Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2008, Vol. 44, No. 1, 47–88
DOI: 10.1214/07-AIHP106

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES

www.imstat.org/aihp

Iterative feature selection in least square regression estimation

Pierre Alquier

Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 and Laboratoire de Statistique, Crest, 3, Avenue Pierre Larousse, 92240 Malakoff, France. E-mail: alquier@ensae.fr

Received 7 January 2005; revised 12 December 2005; accepted 23 January 2006

Abstract. This paper presents a new algorithm to perform regression estimation, in both the inductive and transductive setting. The estimator is defined as a linear combination of functions in a given dictionary. Coefficients of the combinations are computed sequentially using projection on some simple sets. These sets are defined as confidence regions provided by a deviation (PAC) inequality on an estimator in one-dimensional models. We prove that every projection the algorithm actually improves the performance of the estimator. We give all the estimators and results at first in the inductive case, where the algorithm requires the knowledge of the distribution of the design, and then in the transductive case, which seems a more natural application for this algorithm as we do not need particular information on the distribution of the design in this case. We finally show a connection with oracle inequalities, making us able to prove that the estimator reaches minimax rates of convergence in Sobolev and Besov spaces.

Résumé. Cette article présente un nouvel algorithme d'estimation de régression, dans les contextes inductifs et transductifs. L'estimateur est défini par une combinaison linéaire de fonctions choisies dans un dictionnaire donné. Les coefficients de cette combinaison sont calculés par des projections successives sur des ensembles simples. Ces ensembles sont définis comme des régions de confiance données par une inégalité de déviation (ou inégalité PAC). On démontre en particulier que chaque projection au cours de l'algorithme améliore effectivement l'estimateur obtenu. On donne tout d'abord les résultats dans le contexte inductif, où l'algorithme nécessite la connaissance de la distribution du design, puis dans le contexte transductif, plus naturel ici puisque l'algorithme s'applique sans la connaissance de cette distribution. On établit finalement un lien avec les inégalités d'oracle, permettant de montrer que notre estimateur atteint les vitesses optimales dans les espaces de Sobolev et de Besov.

MSC: Primary 62G08; secondary 62G15; 68T05

Keywords: Regression estimation; Statistical learning; Confidence regions; Thresholding methods; Support vector machines

1. The setting of the problem

We give here notations and introduce the inductive and transductive settings.

1.1. Transductive and inductive settings

Let $(\mathcal{X}, \mathcal{B})$ be a measure space and let $\mathcal{B}_{\mathbb{R}}$ denote the Borel σ-algebra on $\mathbb{R}$.

1.1.1. The inductive setting

In the inductive setting, we assume that P is a distribution on pairs $Z = (X, Y)$ taking values in $(\mathcal{X} \times \mathbb{R}, \mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})$, that P is such that:

$$P|Y| < \infty,$$

and that we observe N independent pairs $Z_i = (X_i, Y_i)$ for $i \in \{1, \ldots, N\}$. Our objective is then to estimate the regression function on the basis of the observations.

Definition 1.1 (The regression function). *We denote*:

$$f : \mathcal{X} \to \mathbb{R},$$
$$x \mapsto P(Y|X = x).$$

1.1.2. The transductive setting

In the transductive case, we will assume that, for a given integer $k > 0$, $P_{(k+1)N}$ is some exchangeable probability measure on the space $((\mathcal{X} \times \mathbb{R})^{(k+1)N}, (\mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})^{\otimes 2N})$. We will write $(X_i, Y_i)_{i=1,\dots,(k+1)N} = (Z_i)_{i=1,\dots,(k+1)N}$ a random vector distributed according to $P_{(k+1)N}$.

Definition 1.2 (Exchangeable probability distribution). *For any integer j, let $\mathfrak{S}_j$ denote the set of all permutations of $\{1, \dots, j\}$. We say that $P_{(k+1)N}$ is exchangeable if for any $\sigma \in \mathfrak{S}_{(k+1)N}$ we have*: $(X_{\sigma(i)}, Y_{\sigma(i)})_{i=1,\dots,(k+1)N}$ *has the same distribution under $P_{(k+1)N}$ that $(X_i, Y_i)_{i=1,\dots,(k+1)N}$.*

We assume that we observe $(X_i, Y_i)_{i=1,\dots,N}$ and $(X_i)_{i=N+1,\dots,(k+1)N}$; and the observation $(X_i, Y_i)_{i=1,\dots,(k+1)N}$ is usually called the training sample, while the other part of the vector, $(X_i, Y_i)_{i=N+1,\dots,(k+1)N}$ is called the test sample. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1,\dots,(k+1)N}$. This is why Vapnik [22] called this kind of inference "transductive inference" when he introduced it.

Note that in this setting, the pairs (X_i, Y_i) are not necessarily independent, but are identically distributed. We will let P denote their marginal distribution, and we can here again define the regression function f.

Actually, most statistical problems being usually formulated in the inductive setting, the reader may wonder about the pertinence of the study of the transductive setting. Let us think of the following examples: in quality control, or in a sample survey, we try to infer informations about a whole population from observations on a small sample. In this cases, transductive inference seems actually more adapted than inductive inference, with N the size of the sample and $(k + 1)N$ the size of the population. One can see that the use of inductive results in this context is only motivated by the large values of k (the inductive case is the limit case of the transductive case where $k \to +\infty$). In the problems connected with regression estimation or classification, we can imagine a case where a lot of images are collected for example on the internet. The time to label every picture according to the fact that it represents, or not, a given object being too long, one can think of labeling only 1 over $k + 1$ images, and to use then a transductive algorithm to label automatically the other data. We hope that these examples can convince the reader that the use of the transductive setting is not unrealistic. However, the reader that is not convinced should remember that the transductive inference was first introduced by Vapnik mainly as a tool to study the inductive case: there are techniques to get rid of the second part of the sample by taking an expectation with respect to it and obtain results valid in the inductive setting (see for example a result by Panchenko used in this paper, [17]).

1.2. The model

In both settings, we are going to use the same model to estimate the regression function: Θ. The only thing we assume about Θ is that it is a vector space of functions.

Note in particular that we do not assume that f belongs to Θ.

1.3. Overview of the results

In both settings, we give a PAC inequality on the risk of estimators in one-dimensional models of the form:

$$\{\alpha\theta(\cdot), \alpha \in \mathbb{R}\}$$

for a given $\theta \in \Theta$.

This result motivates an algorithm that performs iterative feature selection in order to perform regression estimation. We will then remark that the selection procedure gives the guarantee that every selected feature actually improves the current estimator.

In the inductive setting (Section 2), it means that we estimate $f(\cdot)$ by a function $\hat{\theta}(\cdot) \in \Theta$, but the selection procedure can only be performed if the statistician knows the marginal distribution $P_{(X)}$ of X under P.

In the transductive case (Section 3), the estimation of $Y_{N+1}, \ldots, Y_{(k+1)N}$ can be performed by the procedure without any prior knowledge about the marginal distribution of X under P. We first focus on the case $k = 1$, and then on the general case $k \in \mathbb{N}^*$.

Finally, in Section 4, we use the main result of the paper (the fact that every selected feature improves the performance of the estimator) as an oracle inequality, to compute the rate of convergence of the estimator in Sobolev and Besov spaces.

The last section (Section 5) is dedicated to the proofs.

The literature on iterative methods for regression estimation is very important, let us mention one of the first algorithm, AdaLine, by Widrow and Hoff [23], or more recent versions like boosting, see [18] and the references within. The technique developed here has some similarity with the so-called greedy algorithms, see [3] (and the references within) for a survey and some recent results. However, note that in this techniques, the iterative update of the estimator is motivated by algorithmic issues, and is not motivated statistically. In particular, AdaLine has no guarantee against overfitting if the number of variables m is large (say $m = N$). For greedy algorithms, on has to specify a particular penalization if one wants to get a guarantee against overfitting. The same remark can be done about boosting algorithm. Here, the algorithm is motivated by a statistical result, and as a consequence has theoretical guarantees against overlearning. It stays however computationally feasible, some pseudo-code is given in the paper.

Closer to our technique are the methods of aggregation of statistical estimators, see [16] and [21] and more recently the mirror descent algorithm studied in [13] or [14]. In this papers, oracle inequalities are given ensuring that the estimator performs as well as the best (linear or convex) aggregation of functions in a given family, up to an optimal term. Note that these inequalities are given in expectation, here almost all results are given in a deviation bound (or PAC bound, a bound that is true with high probability, from which we derive a bound in expectation in Section 4). Similar bounds where given for the PAC-Bayesian model aggregation developed by Catoni [7], Yang [24] and Audibert [2]. In some way, the algorithm proposed in this paper can be seen as a practical way to implement these results.

Note that nearly all the methods in the papers mentioned previously where designed especially for the inductive setting. Very few algorithms were created specifically for the transductive regression problem. The algorithm described in this paper seems more adapted to the transductive setting (remember that the procedure can be performed in the inductive setting only if the statistician knows the marginal distribution of X under P, while there is no such assumption in the transductive context).

Let us however start with a presentation of our method in the inductive context.

2. Main theorem in the inductive case, and application to estimation

2.1. Additional definition

Definition 2.1. *We put*:

$$R(\theta) = P\big[\big(Y - \theta(X)\big)^2\big],$$

$$r(\theta) = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \theta(X_i)\big)^2,$$

and in this setting, our objective is $\overline{\theta}$ *given by*:

$$\overline{\theta} \in \arg\min_{\theta\in\Theta} R(\theta).$$

2.2. Main theorem

We suppose that we have an integer $m \in \mathbb{N}$ and that we are given a finite family of functions:

$$\Theta_0 = \{\theta_1, \ldots, \theta_m\} \subset \Theta.$$

Definition 2.2. *Let us put, for any* $k \in \{1, \dots, m\}$:

$$\overline{\alpha}_k = \arg\min_{\alpha \in \mathbb{R}} R(\alpha\theta_k) = \frac{P[\theta_k(X)Y]}{P[\theta_k(X)^2]},$$

$$\hat{\alpha}_k = \arg\min_{\alpha \in \mathbb{R}} r(\alpha\theta_k) = \frac{(1/N)\sum_{i=1}^N \theta_k(X_i)Y_i}{(1/N)\sum_{i=1}^N \theta_k(X_i)^2},$$

$$\mathcal{C}_k = \frac{(1/N)\sum_{i=1}^N \theta_k(X_i)^2}{P[\theta_k(X)^2]}.$$

Theorem 2.1. *Moreover, let us assume that* P *is such that* $|f|$ *is bounded by a constant* B, *and such that*:

$$P\left\{\left[Y - f(X)\right]^2\right\} \leq \sigma^2 < +\infty.$$

We have, for any $\varepsilon > 0$, *with* $P^{\otimes N}$*-probability at least* $1 - \varepsilon$, *for any* $k \in \{1, \dots, m\}$:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \leq \frac{4[1 + \log(2m/\varepsilon)]}{N}\left[\frac{(1/N)\sum_{i=1}^N \theta_k(X_i)^2 Y_i^2}{P[\theta_k(X)^2]} + B^2 + \sigma^2\right]. \tag{2.1}$$

The proof of this theorem is given in Section 5.6.

2.3. Application to regression estimation

2.3.1. Interpretation of Theorem 2.1 *in terms of confidence intervals*

Definition 2.3. *Let us put, for any* $(\theta, \theta') \in \Theta^2$:

$$d_P(\theta, \theta') = \sqrt{P_{(X)}\left[\left(\theta(X) - \theta'(X)\right)^2\right]}.$$

Let also $\|\cdot\|_P$ *denote the norm associated with this distance,* $\|\theta\|_P = d_P(\theta, 0)$, *and* $\langle\cdot,\cdot\rangle_P$ *the associated scalar product*:

$$\left\langle\theta, \theta'\right\rangle_P = P\left[\theta(X)\theta'(X)\right].$$

Because $\overline{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} R(\alpha\theta_k)$ we have:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) = d_P^2(\mathcal{C}_k\hat{\alpha}_k\theta_k, \overline{\alpha}_k\theta_k).$$

So the theorem can be written:

$$P^{\otimes N}\left\{\forall k \in \{1, \dots, m\}, d_P^2(\mathcal{C}_k\hat{\alpha}_k\theta_k, \overline{\alpha}_k\theta_k) \leq \beta(\varepsilon, k)\right\} \geq 1 - \varepsilon,$$

where $\beta(\varepsilon, k)$ is the right-hand side of inequality (2.1).

Now, note that $\overline{\alpha}_k\theta_k$ is the orthogonal projection of:

$$\overline{\theta} = \arg\min_{\theta\in\Theta} R(\theta)$$

onto the space $\{\alpha\theta_k, \alpha \in \mathbb{R}\}$, with respect to the inner product $\langle\cdot,\cdot\rangle_P$:

$$\overline{\alpha}_k = \arg\min_{\alpha\in\mathbb{R}} d_P(\alpha\theta_k, \overline{\theta}).$$

Definition 2.4. *We define, for any k and ε:*

$$\mathcal{CR}(k,\varepsilon) = \left\{\theta \in \Theta\colon \left|\left\langle \theta - \mathcal{C}_k\hat{\alpha}_k\theta_k, \frac{\theta_k}{\|\theta_k\|_P}\right\rangle_P\right| \leq \sqrt{\beta(\varepsilon,k)}\right\}.$$

Then the theorem is equivalent to the following corollary.

Corollary 2.2. *We have*:

$$P^{\otimes N}\left[\forall k \in \{1,\dots,m\}, \overline{\theta} \in \mathcal{CR}(k,\varepsilon)\right] \geq 1-\varepsilon.$$

In other words: $\bigcap_{k\in\{1,\dots,m\}} \mathcal{CR}(k,\varepsilon)$ is a confidence region at level ε for $\overline{\theta}$.

Definition 2.5. *We write $\Pi_P^{k,\varepsilon}$ the orthogonal projection into $\mathcal{CR}(k,\varepsilon)$ with respect to the distance d_P.*

Note that this orthogonal projection is not a projection on a linear subspace of Θ, and so it is not a linear mapping.

2.3.2. The algorithm

The previous corollaries of Theorem 2.1 motivate the following iterative algorithm:

- choose $\theta^{(0)} \in \Theta$, for example, $\theta^{(0)} = 0$;
- at step $n \in \mathbb{N}^*$, we have: $\theta^{(0)},\dots,\theta^{(n-1)}$. Choose $k(n) \in \{1,\dots,m\}$ (this choice can of course be data dependent), and take:

$$\theta^{(n)} = \Pi_P^{k(n),\varepsilon}\theta^{(n-1)};$$

- we can use the following stopping rule: $\|\theta^{(n-1)} - \theta^{(n)}\|_P^2 \leq \kappa$, where $0 < \kappa < \frac{1}{N}$.

Definition 2.6. *Let n_0 denote the stopping step, and*:

$$\hat{\theta}(\cdot) = \theta^{(n_0)}(\cdot)$$

the corresponding function.

2.3.3. Results and comments on the algorithm

Theorem 2.3. *We have*:

$$P^{\otimes N}\left[\forall n \in \{1,\dots,n_0\}, R\left(\theta^{(n)}\right) \leq R\left(\theta^{(n-1)}\right) - d_P^2\left(\theta^{(n)},\theta^{(n-1)}\right)\right] \geq 1-\varepsilon.$$

Proof. This is just a consequence of the preceding corollary. Let us assume that:

$$\forall k \in \{1,\dots,m\}, \quad R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \leq \beta(\varepsilon,k).$$

Let us choose $n \in \{1,\dots,n_0\}$. We have, for a $k \in \{1,\dots,m\}$:

$$\theta^{(n)} = \Pi_P^{k,\varepsilon}\theta^{(n-1)},$$

where $\Pi_P^{k,\varepsilon}$ is the projection into a convex set that contains $\overline{\theta}$. This implies that:

$$\left\langle \theta^{(n)} - \theta^{(n-1)}, \overline{\theta} - \theta^{(n)}\right\rangle_P \geq 0,$$

or:

$$d_P^2\left(\theta^{(n-1)},\overline{\theta}\right) \geq d_P^2\left(\theta^{(n)},\overline{\theta}\right) + d_P^2\left(\theta^{(n-1)},\theta^{(n)}\right),$$

which can be written:

$$R\big[\theta^{(n-1)}\big] - R(\overline{\theta}) \geq R\big[\theta^{(n)}\big] - R(\overline{\theta}) + d_P^2\big(\theta^{(n-1)}, \theta^{(n)}\big). \qquad \square$$

Actually, the main point in the motivation of the algorithm is that, with probability at least $1-\varepsilon$, whatever the current value $\theta^{(n)} \in \Theta$, whatever the feature $k \in \{1, \ldots, m\}$ (even chosen on the basis of the data), $\Pi_P^{k,\varepsilon}\theta^{(n)}$ is a better estimator than $\theta^{(n)}$.

So we can choose $k(n)$ as we want in the algorithm. For example, Theorem 2.3 motivates the choice:

$$k(n) = \arg\max_k d_P^2\big(\theta^{(n-1)}, \mathcal{CR}(k,\varepsilon)\big).$$

This version of the algorithm is detailed in Fig. 1. If looking for the exact maximum of

$$d_P\big(\theta^{(n-1)}, \mathcal{CR}(k,\varepsilon)\big)$$

with respect to k is too computationally intensive we can use any heuristic to choose $k(n)$, or even skip this maximization and take:

$$k(1) = 1, \ldots, \qquad k(m) = m, \qquad k(m+1) = 1, \ldots, \qquad k(2m) = m, \ldots.$$

Example 2.1. *Let us assume that $\mathcal{X} = [0,1]$ and let us put $\Theta = \mathbb{L}_2(P_{(X)})$. Let $(\theta_k)_{k\in\mathbb{N}^*}$ be an orthonormal basis of Θ. The choice of m should not be a problem, the algorithm itself avoiding itself overlearning we can take a large value of m like $m = N$. In this setting, the algorithm is a procedure for (soft) thresholding of coefficients. In the particular case of a wavelets basis, see* [10] *or* [15] *for a presentation of wavelets coefficient thresholding. Here, the threshold is not necessarily the same for every coefficient. We can remark that the sequential projection on every k is sufficient here*:

$$k(1) = 1, \ldots, k(m) = m,$$

after that $\theta^{(m+n)} = \theta^{(m)}$ for every $n \in \mathbb{N}$ (because all the directions of the different projections are orthogonals).

Actually, it is possible to prove that the estimator is able to adapt itself to the regularity of the function to achieve a good mean rate of convergence. More precisely, if we assume that the true regression function has an (unknown) regularity β, then it is possible to choose m and ε in such a way that the rate of convergence is:

$$N^{-2\beta/(2\beta+1)} \log N.$$

We prove this point in Section 4.

Remark 2.1. *Note that in its general form, the algorithm does not require any assumption about the dictionary of functions $\Theta_0 = \{\theta_1, \ldots, \theta_m\}$. This family can be non-orthogonal, it can even be redundant (the dimension of the vector space generated by Θ_0 can be smaller than m).*

Remark 2.2. *It is possible to generalize Theorem* 2.1 *to models of dimension larger than* 1. *The algorithm itself can take advantage of these generalizations. This point is developed in* [1], *where some experiences about the performances of our algorithm can also be found.*

2.4. Additional notations for some refinements of Theorem 2.1

Note that an improvement of the inequality in Theorem 2.1 (inequality (2.1)) would allow to apply the same method, but would lead to smaller confidence regions and so to better performances. The end of this section is dedicated to improvements (and generalizations) of this bound.

We have $\varepsilon > 0$, $\kappa > 0$, N observations $(X_1, Y_1), \ldots, (X_N, Y_N)$, m features $\theta_1(\cdot), \ldots, \theta_m(\cdot)$ and $c = (c_1, \ldots, c_m) = (0, \ldots, 0) \in \mathbb{R}^m$. Compute at first every $\hat{\alpha}_k$ and $\beta(\varepsilon, k)$ for $k \in \{1, \ldots, m\}$. Set $n \leftarrow 0$.

Repeat:

- set $n \leftarrow n + 1$;
- set *best_improvement* $\leftarrow 0$;
- for $k \in \{1, \ldots, m\}$, compute:

$$v_k = P\big[\theta_k(X)^2\big],$$

$$\gamma_k \leftarrow \hat{\alpha}_k - \frac{1}{v_k} \sum_{j=1}^{m} c_j P\big[\theta_j(X)\theta_k(X)\big],$$

$$\delta_k \leftarrow v_k \big(|\gamma_k| - \beta(\varepsilon, k)\big)_+^2,$$

and if $\delta_k >$ *best_improvement*, set:

$$\text{best\_improvement} \leftarrow \delta_k,$$

$$k(n) \leftarrow k;$$

- if *best_improvement* > 0 set:

$$c_{k(n)} \leftarrow c_{k(n)} + \operatorname{sgn}(\gamma_{k(n)})\big(|\gamma_{k(n)}| - \beta\big(\varepsilon, k(n)\big)\big)_+;$$

until *best_improvement* $< \kappa$ (where $\operatorname{sgn}(x) = -1$ if $x \leq 0$ and 1 otherwise).

Note that at each step n, $\theta^{(n)}$ is given by:

$$\theta^{(n)}(\cdot) = \sum_{k=1}^{m} c_k \theta_k(\cdot),$$

so after the last step we can return the estimator:

$$\hat{\theta}(\cdot) = \sum_{k=1}^{m} c_k \theta_k(\cdot).$$

Fig. 1. Detailed version of the feature selection algorithm.

Hypothesis. *Until the end of Section* 2, *we assume that* Θ *and* P *are such that*:

$$\forall \theta \in \Theta, \quad P \exp\big[\theta(X)Y\big] < +\infty.$$

Definition 2.7. *For any random variable* T *we put*:

$$V(T) = P\big[(T - PT)^2\big],$$

$$M^3(T) = P\big[(T - PT)^3\big],$$

and we define, for any $\gamma \geq 0$, $P_{\gamma T}$ *by*:

$$\frac{\mathrm{d}P_{\gamma T}}{\mathrm{d}P} = \frac{\exp(\gamma T)}{P[\exp(\gamma T)]}.$$

For any random variables T, T' *and any* $\gamma \geq 0$ *we put*:

$$V_{\gamma T}(T') = P_{\gamma T}\big[(T' - P_{\gamma T}T')^2\big],$$
$$M^3_{\gamma T}(T') = P_{\gamma T}\big[(T' - P_{\gamma T}T')^3\big].$$

Section 2.5 gives an improvement of Theorem 2.1 while Section 2.6 extends it to the case of a data-dependant family Θ_0.

2.5. *Refinements of Theorem* 2.1

Theorem 2.4. *Let us put*:

$$W_\theta = \theta(X)Y - P\big(\theta(X)Y\big).$$

Then we have, for any $\varepsilon > 0$, *with* $P^{\otimes N}$*-probability at least* $1-\varepsilon$, *for any* $k \in \{1, \ldots, m\}$:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \leq \frac{2\log(2m/\varepsilon)}{N}\frac{V(W_{\theta_k})}{P[\theta_k(X)^2]} + \frac{\log^3(2m/\varepsilon)}{N^{3/2}}C_N(P, m, \varepsilon, \theta_k),$$

where we have:

$$C_N(P, m, \varepsilon, \theta_k) = I_{\theta_k}\left(\sqrt{\frac{2\log(2m/\varepsilon)}{NV(W_{\theta_k})}}\right)^2 \frac{\sqrt{2}}{V(W_{\theta_k})^{5/2}P[\theta_k(X)^2]}$$
$$+ I_{\theta_k}\left(\sqrt{\frac{2\log(2m/\varepsilon)}{NV(W_{\theta_k})}}\right)^4 \frac{\log^2(2m/\varepsilon)}{\sqrt{N}V(W_{\theta_k})^6P[\theta_k(X)^2]},$$

with:

$$I_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\beta\gamma W_\theta}(W_\theta)\,\mathrm{d}\beta.$$

For the proof, see Section 5.1.

Actually, the method we proposed requires to be able to compute explicitly the upper bound in this theorem. Remark that, with ε and m fixed:

$$C_N(P, m, \varepsilon, \theta_k) \underset{N\to+\infty}{\longrightarrow} \frac{\sqrt{2}[M^3(W_{\theta_k})]^2}{9V(W_{\theta_k})^{5/2}P[\theta_k(X)^2]},$$

and so we can choose to consider only the first-order term. Another possible choice is to make stronger assumptions on P and Θ_0 that allow to upper bound explicitly $C_N(P, m, \varepsilon, \theta_k)$. For example, if we assume that Y is bounded by C_Y and that $\theta_k(\cdot)$ is bounded by C'_k then W_{θ_k} is bounded by $C_k = 2C_YC'_k$ and we have (basically):

$$C_N(P, m, \varepsilon, \theta_k) \leq \frac{64\sqrt{2}C_k^2}{9V(W_{\theta_k})^{5/2}P[\theta_k(X)^2]} + \frac{4096C_k^4\log^3(2m/\varepsilon)}{81\sqrt{N}V(W_{\theta_k})^6P[\theta_k(X)^2]}.$$

The main problem is actually that the first-order term contains the quantity $V(W_{\theta_k})$ that is not observable, and we would like to be able to replace this quantity by its natural estimator:

$$\hat{V}_k = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i\theta_k(X_i) - \frac{1}{N}\sum_{j=1}^{N} Y_j\theta_k(X_j)\right]^2.$$

The following theorem justifies this method.

Theorem 2.5. *If we assume that there is a constant c such that*:

$$\forall k \in \{1,\dots,m\}, \quad P\left[\exp\left(cW_{\theta_k}^2\right)\right] < \infty,$$

we have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $k \in \{1,\dots,m\}$:

$$R(\mathcal{C}_k\hat{\alpha}_k\theta_k) - R(\overline{\alpha}_k\theta_k) \le \frac{2\log(4m/\varepsilon)}{N}\frac{\hat{V}_k}{P[\theta_k(X)^2]} + \frac{\log(4m/\varepsilon)}{N^{3/2}}C'_N(P,m,\varepsilon,\theta_k),$$

where we have:

$$\hat{V}_k = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i\theta_k(X_i) - \frac{1}{N}\sum_{j=1}^{N} Y_j\theta_k(X_j)\right]^2,$$

and

$$\begin{aligned}
C'_N(P,m,\varepsilon,\theta_k) &= C_N\left(P,m,\frac{\varepsilon}{2},\theta_k\right)\log^2\frac{4m}{\varepsilon} \\
&\quad + \frac{2\log^{1/2}(2m/\varepsilon)}{P[\theta_k(X)^2]}\left[\sqrt{2V\left(W_{\theta_k}^2\right)} + \frac{\log(2m/\varepsilon)}{\sqrt{N}V(W_{\theta_k}^2)}J_{\theta_k}\left(\sqrt{\frac{2\log(2m/\varepsilon)}{NV(W_{\theta_k}^2)}}\right)\right] \\
&\quad + \frac{2\log^{1/2}(4m/\varepsilon)}{P[\theta_k(X)^2]}\left[\sqrt{2V(W_{\theta_k})} + \frac{\log^2(2m/\varepsilon)}{\sqrt{N}V(W_{\theta_k})^3}I_{\theta_k}\left(\sqrt{\frac{2\log(4m/\varepsilon)}{NV(W_{\theta_k})}}\right)\right] \\
&\quad \times\left[\left|\frac{2}{N}\sum_{i=1}^{N}Y_i\theta_k(X_i)\right|\sqrt{\frac{2V(W_{\theta_k})\log(4m/\varepsilon)}{N}} + \frac{\log^{5/2}(2m/\varepsilon)}{NV(W_{\theta_k})^3}I_{\theta_k}\left(\sqrt{\frac{2\log(4m/\varepsilon)}{NV(W_{\theta_k})}}\right)\right]
\end{aligned}$$

and

$$J_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_{\theta_k}^2}\left(W_\theta^2\right)\mathrm{d}\beta.$$

The proof is given in Section 5.1.

2.6. An extension to the case of Support Vector Machines

Thanks to a method due to Seeger [20], it is possible to extend this method to the case where the set Θ_0 is data dependent in the following way:

$$\Theta_0(Z_1,\dots,Z_N,N) = \bigcup_{i=1}^{N}\Theta_0(Z_i,N),$$

where for any $z \in \mathcal{X} \times \mathbb{R}$, the cardinality of the set $\Theta_0(z, N)$ depends only on N, not on z. We will write $m'(N)$ this cardinality. So we have:

$$\left|\Theta_0(Z_1, \ldots, Z_N, N)\right| \leq N\left|\Theta_0(Z_i, N)\right| = Nm'(N).$$

We put:

$$\Theta_0(Z_i, N) = \{\theta_{i,1}, \ldots, \theta_{i,m'(N)}\}.$$

In this case, we need some adaptations of our previous notations.

Definition 2.8. *We put, for $i \in \{1, \ldots, N\}$:*

$$r_i(\theta) = \frac{1}{N-1} \sum_{\substack{j \in \{1,\ldots,N\}, \\ j \neq i}} \big(Y_j - \theta(X_j)\big)^2.$$

For any $(i, k) \in \{1, \ldots, N\} \times \{1, \ldots, m'(N)\}$, we write:

$$\hat{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} r_i(\alpha\theta_{i,k}) = \frac{\sum_{j \neq i} \theta_{i,k}(X_j) Y_j}{\sum_{j \neq i} \theta_{i,k}(X_j)^2},$$

$$\overline{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} R(\alpha\theta_{i,k}) = \frac{P[\theta_{i,k}(X)Y]}{P[\theta_{i,k}(X)^2]},$$

$$\mathcal{C}_{i,k} = \frac{1/(N-1)\sum_{j \neq i} \theta_{i,k}(X_j)^2}{P[\theta_{i,k}(X)^2]}.$$

Theorem 2.6. *We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, \ldots, m'(N)\}$ and $i \in \{1, \ldots, N\}$:*

$$R(\mathcal{C}_{i,k}\hat{\alpha}_{i,k}\theta_{i,k}) - R(\overline{\alpha}_{i,k}\theta_{i,k}) \leq \frac{2\log(2Nm'(N)/\varepsilon)}{N-1} \frac{V(W_{\theta_{i,k}})}{P[\theta_{i,k}(X)^2]} + \frac{\log^3(2Nm'(N)/\varepsilon)}{(N-1)^{3/2}} C_{N-1}\big(P, Nm'(N), \varepsilon, \theta_{i,k}\big).$$

The proof is given in Section 5.1.

We can use this theorem to build an estimator using the algorithm described in the previous subsection, with obvious changes in the notations.

Example 2.2. *Let us consider the case where $\mathcal{H}$ is a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$, and:*

$$\Theta = \big\{\theta(\cdot) = \big\langle h, \Psi(\cdot)\big\rangle, h \in \mathcal{H}\big\}$$

where Ψ is an application $\mathcal{X} \to \Theta$. Let us put $\Theta_0[(x, y), N] = \{\langle \Psi(x), \Psi(\cdot)\rangle\}$. In this case we have $m'(N) = 1$ and the estimator is of the from:

$$\hat{\theta}(\cdot) = \sum_{i=1}^{N} \alpha_{i,1}\big\langle \Psi(X_i), \Psi(\cdot)\big\rangle.$$

Let us define,

$$K\big(x, x'\big) = \big\langle \Psi(x), \Psi\big(x'\big)\big\rangle,$$

the function K is called the kernel, and:

$$I = \{1 \leq i \leq N\colon \alpha_{i,1} \neq 0\},$$

that is called the set of support vectors. Then the estimate has the form of a support vector machine (*SVM*):

$$\hat{\theta}(\cdot) = \sum_{i \in I} \alpha_{i,1} K(X_i, \cdot).$$

SVM where first introduced by Boser, Guyon and Vapnik [5] *in the context of classification, and then generalized by Vapnik* [22] *to the context of regression estimation. For a general introduction to SVM, see also* [6] *and* [9].

Example 2.3. *A widely used kernel is the Gaussian kernel*:

$$K_\gamma\bigl(x, x'\bigr) = \exp\biggl(-\gamma \frac{d^2(x, x')}{2}\biggr),$$

where $d(\cdot, \cdot)$ is some distance over the space $\mathcal{X}$ and $\gamma > 0$. But in practice, the choice of the parameter γ is difficult. A way to solve this problem is to introduce multiscale SVM. We simply take Θ as the set of all bounded functions $\mathcal{X} \to \mathbb{R}$. Now, let us put:

$$\Theta_0\bigl[(x, y), N\bigr] = \bigl\{K_2(x, \cdot), K_{2^2}(X, \cdot), \ldots, K_{2^{m'(N)}}(x, \cdot)\bigr\}.$$

In this case, we obtain an estimator of the form:

$$\hat{\theta}(\cdot) = \sum_{k=1}^{m'(N)} \sum_{i \in I_k} \alpha_{i,k} K_{2^k}(X_i, \cdot),$$

that could be called multiscale SVM. Remark that we can use this technique to define SVM using simultaneously different kernels (*not necessarily the same kernel at different scales*).

3. The transductive case

3.1. Notations

Let us recall that we assume that $k \in \mathbb{N}^*$, that $P_{(k+1)N}$ is some exchangeable probability measure (let us recall that exchangeability is defined in Definition 1.2) on the space $((\mathcal{X} \times \mathbb{R})^{(k+1)N}, (\mathcal{B} \times \mathcal{B}_{\mathbb{R}})^{\otimes(k+1)N})$. Let $(X_i, Y_i)_{i=1,\ldots,(k+1)N} = (Z_i)_{i=1,\ldots,(k+1)N}$ denote a random vector distributed according to $P_{(k+1)N}$.

Let us remark that under this condition, the marginal distribution of every Z_i is the same, we will call P this distribution. In the particular case where the observations are i.i.d., we will have $P_{(k+1)N} = P^{\otimes(k+1)N}$, but what follows still holds for general exchangeable distributions $P_{(k+1)N}$.

We assume that we observe $(X_i, Y_i)_{i=1,\ldots,N}$ and $(X_i)_{i=N+1,\ldots,(k+1)N}$. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1,\ldots,(k+1)N}$.

Definition 3.1. *We put, for any $\theta \in \Theta$*:

$$r_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \bigl(Y_i - \theta(X_i)\bigr)^2,$$

$$r_2(\theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \bigl(Y_i - \theta(X_i)\bigr)^2.$$

Our objective is:

$$\overline{\theta_2} = \arg\min_{\theta\in\Theta} r_2(\theta),$$

if the minimum of r_2 is not unique then we take for $\overline{\theta_2}$ any element of Θ reaching the minimum value of r_2.

Let Θ_0 be a finite family of vectors belonging to Θ, so that $|\Theta_0| = m$. Actually, Θ_0 is allowed to be data-dependent:

$$\Theta_0 = \Theta_0(X_1, \ldots, X_{(k+1)N}),$$

but we assume that the function $(x_1, \ldots, x_{(k+1)N}) \mapsto \Theta_0(x_1, \ldots, x_{(k+1)N})$ is exchangeable with respect to its $(k+1)N$ arguments, and is such that $m = m(N)$ depends only on N, not on $(X_1, \ldots, X_{(k+1)N})$.

The problem of the indexation of the elements of Θ_0 is not straightforward and we must be very careful about it. Let $<_\Theta$ be a complete order on Θ, and write:

$$\Theta_0 = \{\theta_1, \ldots, \theta_m\},$$

where

$$\theta_1 <_\Theta \cdots <_\Theta \theta_m.$$

Remark that, in this case, every θ_h is an exchangeable function of $(X_1, \ldots, X_{(k+1)N})$.

Definition 3.2. *Now, let us write, for any $h \in \{1, \ldots, m\}$*:

$$\alpha_1^h = \arg\min_{\alpha\in\mathbb{R}} r_1(\alpha\theta_h) = \frac{\sum_{i=1}^N \theta_h(X_i)Y_i}{\sum_{i=1}^N \theta_h(X_i)^2},$$

$$\alpha_2^h = \arg\min_{\alpha\in\mathbb{R}} r_2(\alpha\theta_h) = \frac{\sum_{i=N+1}^{(k+1)N} \theta_h(X_i)Y_i}{\sum_{i=N+1}^{(k+1)N} \theta_h(X_i)^2},$$

$$C^h = \frac{(1/N)\sum_{i=1}^N \theta_h(X_i)^2}{(1/(kN))\sum_{i=N+1}^{(k+1)N} \theta_h(X_i)^2}.$$

3.2. Basic results for $k = 1$

In a first time we focus on the case where $k = 1$ as a method due to Catoni [6] brings a substantial simplification of the bound in this case.

Theorem 3.1. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1 - \varepsilon$, for any $h \in \{1, \ldots, m\}$*:

$$r_2\big[\big(C^h\alpha_1^h\big)\cdot\theta_h\big] - r_2\big(\alpha_2^h\cdot\theta_h\big) \le 4\left[\frac{(1/N)\sum_{i=1}^{2N}\theta_h(X_i)^2Y_i^2}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2}\right]\frac{\log(2m/\varepsilon)}{N}.$$

Remark 3.1. *Here again, it is possible to make some hypothesis in order to make the right-hand side of the theorem observable. In particular, if we assume that*:

$$\exists B \in \mathbb{R}_+, \quad P\big(|Y| \le B\big) = 1,$$

then we can get a looser observable upper bound:

$$P_{2N}\left\{\forall k \in \{1, \ldots, m\}, r_2\big[\big(C^h\alpha_1^h\big)\cdot\theta_h\big] - r_2\big(\alpha_2^h\cdot\theta_h\big) \le 4\left[B^2 + \frac{(1/N)\sum_{i=1}^{N}\theta_h(X_i)^2Y_i^2}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2}\right]\frac{\log(2m/\varepsilon)}{N}\right\} \ge 1 - \varepsilon.$$

If we do not want to make this assumption, we can use the following variant, that gives a first-order approximation for the bound.

Theorem 3.2. *For any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $h \in \{1,\dots,m\}$:*

$$r_2\big[\big(C^h\alpha_1^h\big)\cdot\theta_h\big] - r_2\big(\alpha_2^h\cdot\theta_h\big) \le \frac{8\log(4m/\varepsilon)}{N}\left[\frac{(1/N)\sum_{i=1}^N \theta_h(X_i)^2 Y_i^2}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2} + \sqrt{\frac{(1/N)\sum_{i=1}^{2N}\theta_h(X_i)^4 Y_i^4 \log(2m/\varepsilon)}{2N}}\right].$$

Remark 3.2. *Let us assume that Y is such that we know two constants b_Y and B_Y such that:*

$$P\exp\big(b_Y|Y|\big) \le B_Y < \infty.$$

Then we have, with probability at least $1-\varepsilon$:

$$\sup_{i\in\{1,\dots,2N\}} |Y_i| \le \frac{1}{b_Y}\log\frac{2NB_Y}{\varepsilon}.$$

Combining both inequalities leads by a union bound argument leads to:

$$r_2\big[\big(C^h\alpha_1^h\big)\cdot\theta_h\big] - r_2\big(\alpha_2^h\cdot\theta_h\big) \le \frac{8\log(8m/\varepsilon)}{N}\left[\frac{(1/N)\sum_{i=1}^N \theta_h(X_i)^2 Y_i^2}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2} + \sqrt{\frac{(1/N)\sum_{i=1}^{2N}\theta_h(X_i)^4 \log(4m/\varepsilon)\log^4(4NB_Y/\varepsilon)}{2Nb_Y^4}}\right].$$

The proofs of both theorems are given in the proofs section, more precisely in Section 5.2.

Let us compare the first-order term of this theorem to the analogous term in the inductive case (Theorems 2.4 and 2.5). The factor of the variance term is 8 instead of 2 in the inductive case. A factor 2 is to be lost because we have here the variance of a sample of size $2N$ instead of N in the inductive case. But another factor 2 is lost here. Moreover, in the inductive case, we obtained the real variance of $Y\theta_h(X)$ instead of the moment of order 2 here.

In the next subsection, we give several improvements of these bounds, that allows to recover a real variance, and to recover the factor 2. We also give a version that allows to deal with a test sample of different size, this being a generalization of Theorem 3.1 more than of its improved variants.

We then give the analog of the algorithm proposed in the inductive case in this transductive setting.

3.3. Improvements of the bound and general values for k

The proof of all the theorems of this subsection is given in the next section.

3.3.1. Variance term (in the case $k = 1$)

We introduce some new notations.

Definition 3.3. *We write:*

$$\forall\theta\in\Theta,\ r_{1,2}(\theta) = r_1(\theta) + r_2(\theta)$$

and, in the case of a model $k \in \{1,\dots,m\}$:

$$\alpha_{1,2}^h = \arg\min_{\alpha\in\mathbb{R}} r_{1,2}(\alpha\theta_h).$$

The we have the following theorem.

Theorem 3.3. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $h \in \{1,\dots,m\}$:*

$$r_2\big(\mathcal{C}^h \alpha_1^h \theta_h\big) - r_2\big(\alpha_2^h \theta_h\big) \leq 4\left[\frac{(1/N)\sum_{i=1}^{2N}[\theta_h(X_i)Y_i - \alpha_{1,2}^h \theta_h(X_i)^2]^2}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2}\right]\frac{\log(2m/\varepsilon)}{N}.$$

For the proof see Section 5.3.

It is moreover possible to modify the upper bound to make it observable. We obtain that with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,\dots,m\}$:

$$r_2\big[\big(\mathcal{C}^h \alpha_1^h\big)\theta_h\big] - r_2\big(\alpha_2^h \theta_h\big) \leq \frac{16\log(4m/\varepsilon)}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\big(\theta_h(X_i)Y_i - \alpha_1^h \theta_h(X_i)^2\big)^2\right] + \mathcal{O}\left(\left[\frac{\log(m/\varepsilon)}{N}\right]^{3/2}\right).$$

So we can see that this theorem is an improvement on Theorem 3.1 when some features $\theta_h(X)$ are well correlated with Y. But we loose another factor 2 by making the first-order term of the bound observable.

3.3.2. Improvement of the variance term ($k = 1$)

Theorem 3.4. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $h \in \{1,\dots,m\}$:*

$$r_2\big(\mathcal{C}^h \alpha_1^h \theta_h\big) - r_2\big(\alpha_2^h \theta_h\big) \leq \left[\frac{1}{1-2\log(2m/\varepsilon)/N}\right]\frac{2\log(2m/\varepsilon)}{N}\frac{V_1(\theta_h)+V_2(\theta_h)}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2},$$

where:

$$V_1(\theta_h) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i\theta_h(X_i) - \frac{1}{N}\sum_{j=1}^{N}Y_j\theta_h(X_j)\right]^2,$$

$$V_2(\theta_h) = \frac{1}{N}\sum_{i=N+1}^{2N}\left[Y_i\theta_h(X_i) - \frac{1}{N}\sum_{j=N+1}^{2N}Y_j\theta_h(X_j)\right]^2.$$

It is moreover possible to give an observable upper bound: we obtain that with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,\dots,m\}$:

$$r_2\big[\big(\mathcal{C}^h \alpha_1^h\big)\theta_h\big] - r_2\big(\alpha_2^h \theta_h\big) \leq \left[\frac{1}{1-2\log(4m/\varepsilon)/N}\right]\frac{4\log(4m/\varepsilon)}{N}\frac{V_1(\theta_h)}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2}$$
$$+\left[\frac{1}{1-2\log(4m/\varepsilon)/N}\right]2\big(2+\sqrt{2}\big)\left(\frac{\log(6m/\varepsilon)}{N}\right)^{3/2}\frac{\sqrt{(1/N)\sum_{i=1}^{2N}\theta_h(X_i)^4 Y_i^4}}{(1/N)\sum_{i=N+1}^{2N}\theta_h(X_i)^2}.$$

Here again, we can make the bound fully observable under an exponential moment or boundedness assumption about Y. For a complete proof see Section 5.4.

3.3.3. The general case ($k \in \mathbb{N}^*$)

We need some new notations in this case.

Definition 3.4. *Let us put*:

$$\mathbf{P} = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\delta_{Z_i},$$

and, for any $\theta \in \Theta$:

$$\mathbb{V}_\theta = \mathbf{P}\big\{\big[\big(\theta(X)Y\big) - \mathbf{P}\big(\theta(X)Y\big)\big]^2\big\}.$$

Then we have the following theorem.

Theorem 3.5. *Let us assume that we have constants* B_h *and* β_h *such that, for any* $h \in \{1, \dots, m\}$:

$$P\exp\big(\beta_h\big|\theta_h(X_i)Y_i\big|\big) \le B_h.$$

For any $\varepsilon > 0$, *with* $P_{(k+1)N}$ *probability at least* $1 - \varepsilon$ *we have, for any* $h \in \{1, \dots, m\}$:

$$\begin{aligned} &r_2\big(\mathcal{C}^h\alpha_1^h\theta_h\big) - r_2\big(\alpha_2^h\theta_h\big) \\ &\quad\le \frac{(1+1/k)^2}{(1/(kN))\sum_{i=N+1}^{(k+1)N}\theta_h(X_i)^2}\bigg[\frac{2\mathbb{V}_{\theta_h}\log(4m/\varepsilon)}{N} \\ &\qquad + \frac{16(\log(4m/\varepsilon))^{3/2}(\log(4(k+1)mNB_h/\varepsilon))^3}{3\beta_h^3N^{3/2}\mathbb{V}_{\theta_h}^{1/2}} + \frac{64(\log(4m/\varepsilon))^2(\log(4(k+1)mNB_h/\varepsilon))^6}{9\beta_h^6N^2\mathbb{V}_{\theta_h}^2}\bigg]. \end{aligned}$$

Here again, it is possible to replace the variance term by its natural estimator:

$$\hat{\mathbb{V}}_{\theta_h} = \frac{1}{N}\sum_{i=1}^{N}\bigg[\theta_h(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N}\theta_h(X_j)Y_j\bigg]^2.$$

For a complete proof of the theorem see the section dedicated to the proofs (more precisely Section 5.5).

3.4. Application to transductive regression

We give here the interpretation of the preceding theorems in terms of confidence; this motivates an algorithm similar to the one described in the inductive case.

Definition 3.5. *We take, for any* $(\theta, \theta') \in \Theta^2$:

$$d_2\big(\theta, \theta'\big) = \sqrt{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\big[\theta(X_i) - \theta'(X_i)\big]^2}.$$

Let also $\|\theta\|_2 = d_2(\theta, 0)$ *and*:

$$\big\langle\theta, \theta'\big\rangle_2 = \frac{1}{(k+1)N}\sum_{i=N+1}^{(k+1)N}\theta(X_i)\theta'(X_i).$$

We define, for any $h \in \{1, \dots, m\}$ *and* ε:

$$\mathcal{CR}(h, \varepsilon) = \big\{\theta \in \Theta\colon \big|\big\langle\theta - \mathcal{C}^h\alpha_1^h\theta_h, \theta_h\big\rangle_2\big| \le \sqrt{\beta(\varepsilon, h)}\big\},$$

where $\beta(\varepsilon, h)$ *is the upper bound in Theorem* 3.1 (*or in any other theorem given in the transductive section*).

For the same reasons as in the inductive case, these theorems imply the following result.

Corollary 3.6. *We have*:

$$P_{2N}\big[\forall h \in \{1, \dots, m\}, \overline{\theta}_2 \in \mathcal{CR}(h, \varepsilon)\big] \ge 1 - \varepsilon.$$

Definition 3.6. *We call $\Pi_2^{h,\varepsilon}$ the orthogonal projection into $\mathcal{CR}(h,\varepsilon)$ with respect to the distance d_2.*

We propose the following algorithm:

- choose $\theta^{(0)} \in \Theta$ (for example 0);
- at step $n \in \mathbb{N}^*$, we have: $\theta^{(0)}, \ldots, \theta^{(n-1)}$. Choose $h(n)$, for example:

$$h(n) = \underset{h\in\{1,\ldots,m\}}{\arg\max}\ d_2\big(\theta^{(n-1)}, \mathcal{CR}(h,\varepsilon)\big),$$

and take:

$$\theta^{(n)} = \Pi_2^{h(n),\varepsilon}\theta^{(n-1)};$$

- we can use the following stopping rule: $\|\theta^{(n-1)} - \theta^{(n)}\|_2^2 \leq \kappa$ where $0 < \kappa < \frac{1}{N}$.

Definition 3.7. *We write n_0 the stopping step, and:*

$$\theta(\cdot) = \theta^{(n_0)}(\cdot)$$

the corresponding function.

Here again we give a detailed version of the algorithm, see Fig. 2. Remark that as in the inductive case, we are allowed to use whatever heuristic to choose $k(n)$ if we want to avoid the maximization.

Theorem 3.7. *We have:*

$$P_{2N}\big[\forall n \in \{1,\ldots,n_0\}, r_2\big(\theta^{(n)}\big) \leq r_2\big(\theta^{(n-1)}\big) - d_2^2\big(\theta^{(n)}, \theta^{(n-1)}\big)\big] \geq 1-\varepsilon.$$

The proof of this theorem is exactly the same as the proof of Theorem 2.3.

Example 3.1 (Estimation of wavelet coefficients). *Let us consider the case where Θ_0 does not depend on the observations. We can, for example, choose a basis of Θ, or a basis of a subspace of Θ. We obtain an estimator of the form:*

$$\theta(x) = \sum_{h=1}^{m} \alpha^h \theta_h(x).$$

In the case when $(\theta_k)_k$ is a wavelet basis, then we obtain here again a procedure for thresholding wavelets coefficients.

Example 3.2 (SVM and multiscale SVM). *Let us choose Θ as the set of all functions $\mathcal{X} \to \mathbb{R}$, a family of kernels $K_1, \ldots, K_{m'(N)}$ for a $m'(N) \geq 1$ and:*

$$\Theta_0 = \big\{K_h(X_i,\cdot), h \in \big\{1,\ldots,m'(N)\big\}, i \in \big\{1,\ldots,(k+1)N\big\}\big\}.$$

In this case we have $m = (k+1)Nm'(N)$. We obtain an estimator of the form:

$$\theta(x) = \sum_{h=1}^{m'(N)} \sum_{j=1}^{2N} \alpha^{j,h} K_h(X_j, x).$$

Let us put:

$$I_h = \big\{j \in \{1,\ldots,2N\}, \alpha^{j,h} \neq 0\big\}.$$

We have $\varepsilon > 0$, $\kappa > 0$, N observations $(X_1, Y_1), \ldots, (X_N, Y_N)$ and also $X_{N+1}, \ldots, X_{(k+1)N}$, m features $\theta_1(\cdot), \ldots, \theta_m(\cdot)$ and $c = (c_1, \ldots, c_m) = (0, \ldots, 0) \in \mathbb{R}^m$. First, compute every α_1^h and $\beta(\varepsilon, h)$ for $h \in \{1, \ldots, m\}$. Set $n \leftarrow 0$.

Repeat:

- set $n \leftarrow n + 1$;
- set *best_improvement* $\leftarrow 0$;
- for $h \in \{1, \ldots, m\}$, compute:

$$v_h = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \theta_h(X_i)^2,$$

$$\gamma_h \leftarrow \alpha_1^h - \frac{1}{v_h} \sum_{j=1}^{m} c_j \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \theta_j(X_i)\theta_h(X_i),$$

$$\delta_h \leftarrow v_h \big(|\gamma_h| - \beta(\varepsilon, h)\big)_+^2,$$

and if $\delta_h >$ *best_improvement*, set:

$$\textit{best\_improvement} \leftarrow \delta_h,$$
$$h(n) \leftarrow h;$$

- if *best_improvement* > 0 set:

$$c_{h(n)} \leftarrow c_{h(n)} + \operatorname{sgn}(\gamma_{h(n)})\big(|\gamma_{h(n)}| - \beta\big(\varepsilon, h(n)\big)\big)_+;$$

until *best_improvement* $< \kappa$.

Return the estimation:

$$[\tilde{Y}_{N+1}, \ldots, \tilde{Y}_{(k+1)N}] = \big[\hat{\theta}(X_{N+1}), \ldots, \hat{\theta}(X_{(k+1)N})\big],$$

where:

$$\hat{\theta}(\cdot) = \sum_{h=1}^{m} c_h \theta_h(\cdot).$$

Fig. 2. Detailed version of the feature selection algorithm in the transductive case.

We have:

$$\theta(x) = \sum_{h=1}^{m'(N)} \sum_{j \in I_h} \alpha^{j,h} K_h(X_i, x),$$

that is a Support Vector Machine with different kernel estimate; *like in Example* 2.3, *the kernels* K_h *can be the same kernel taken at different scales.*

Example 3.3 (Kernel PCA Kernel Projection Machine). *Take the same Θ and consider the kernel*:

$$K\left(x,x'\right)=\left\langle \Psi(x),\Psi\left(x'\right)\right\rangle .$$

Let us consider a principal component analysis (*PCA*) *of the family*:

$$\left\{K(X_i,\cdot),\ldots,K(X_{(k+1)N},\cdot)\right\}$$

by performing a diagonalization of the matrix:

$$\left(K(X_i,X_j)\right)_{1\leq i,j\leq(k+1)N}.$$

This method is known as Kernel PCA, *see for example* [19]. *We obtain eigenvalues*:

$$\lambda^1\geq\cdots\geq\lambda^{(k+1)N}$$

and associated eigenvectors $e^1,\ldots,e^{(k+1)N}$, *associated to elements of* Θ:

$$k_1(\cdot)=\sum_{i=1}^{(k+1)N}e_i^1K(X_i,\cdot),\ldots,k_{(k+1)N}(\cdot)=\sum_{i=1}^{(k+1)N}e_i^{(k+1)N}K(X_i,\cdot)$$

that are exchangeable functions of the observations. Using the family:

$$\Theta_0=\{k_1,\ldots,k_{(k+1)N}\},$$

we obtain an algorithm that selects which eigenvectors are going to be used in the regression estimation. This is very close to the Kernel Projection Machine (*KPM*) *described by Blanchard*, *Massart*, *Vert and Zwald* [4] *in the context of classification.*

4. Rates of convergence in Sobolev and Besov spaces

We conclude this paper by coming back to the inductive case. We use Theorem 2.3 as an oracle inequality to show that the obtained estimator is adaptative, which means that if we assume that the true regression function f has an unknown regularity β, then the estimator is able to reach the optimal speed of convergence $N^{-2\beta/(2\beta+1)}$ up to a $\log N$ factor.

4.1. Presentation of the context

Here we assume that $\mathcal{X}$ is a compact interval of $\mathbb{R}$, that $\Theta=\mathbb{L}_2(P_{(X)})$ and that P is such that $Y=f(X)+\eta$ with η independent of X, $P\eta=0$ and $P(\eta^2)\leq\sigma^2<+\infty$.

We assume that $(\theta_k)_{k\in\mathbb{N}^*}$ is an orthonormal basis of Θ. We still have to choose $m\in\mathbb{N}$ and we will take $\Theta_0=\{\theta_1,\ldots,\theta_m\}$.

Remark that the orthogonality means here that $P[\theta_k(X)^2]=1$ for any $k\in\mathbb{N}^*$, and that:

$$P\left[\theta_k(X)\theta_{k'}(X)\right]=0$$

for any $k'\neq k$.

4.2. Rate of convergence of the estimator: the Sobolev space case

Now, let us put:

$$\overline{\theta}^m=\underset{\theta\in\mathrm{Span}(\Theta_0)}{\arg\min}\ R(\theta)$$

(that depends effectively on m by $\Theta_0 = \{\theta_1, \dots, \theta_m\}$), and let us assume that f satisfies the two following conditions: it is regular, namely there is an unknown $\beta \geq 1$ and a $C \geq 0$ such that:

$$\left\|\overline{\theta}^m - f\right\|_P^2 \leq Cm^{-2\beta},$$

and that we have a constant $B < \infty$ such that:

$$\sup_{x \in \mathcal{X}} f(x) \leq B$$

with B known to the statistician. It follows that:

$$\|f\|_P^2 \leq B^2.$$

If follows that every set, for $k \in \{1, \dots, m\}$:

$$\mathcal{F}_k = \left\{ \sum_{j=1}^{\infty} \alpha_j \theta_j \colon \alpha_k^2 \leq B^2 \right\} \cap \Theta$$

is a convex set that contains f and such that the orthogonal projection: $\Pi_P^{\mathcal{F},m} = \Pi_P^{\mathcal{F}_m} \cdots \Pi_P^{\mathcal{F}_1}$ (where $\Pi_P^{\mathcal{F}_k}$ denotes the orthogonal projection on $\mathcal{F}_k$) can only improve an estimator:

$$\forall \theta, \quad \left\|\Pi_P^{\mathcal{F},m}\theta - f\right\|_P^2 \leq \|\theta - f\|_P^2.$$

Actually, note that this projection just consists in thresholding very large coefficients to a limited value. This modification is necessary in what follows, but this is just a technical remark: most of the time, our estimator won't be modified by $\Pi_P^{\mathcal{F},m}$ for any m.

Remember also that in this context, the estimator given in Definition 2.6 is just:

$$\hat{\theta} = \Pi_P^{m,\varepsilon} \cdots \Pi_P^{1,\varepsilon} 0.$$

Theorem 4.1. *Let us assume that $\Theta = \mathbb{L}_2(P_{(X)})$, $\mathcal{X} = [0,1]$ and $(\theta_k)_{k \in \mathbb{N}^*}$ is an orthonormal basis of Θ. Let us assume that we are in the idealized regression model*:

$$Y = f(X) + \eta,$$

where $P\eta = 0$, $P(\eta^2) \leq \sigma^2 < \infty$ and η and X are independent, and σ is known. Let us assume that $f \in \Theta$ is such that there is an unknown $\beta \geq 1$ and an unknown $C \geq 0$ such that:

$$\|\overline{\theta}_m - f\|_P^2 \leq Cm^{-2\beta},$$

and that we have a constant $B < \infty$ such that:

$$\sup_{x \in \mathcal{X}} f(x) \leq B$$

with B known to the statistician. Then our estimator $\hat{\theta}$ (given in Definition 2.6 *with $n_0 = m$ here, build using the bound $\beta(\varepsilon, k)$ given in Theorem* 2.1*), with $\varepsilon = N^{-2}$ and $m = N$, is such that, for any $N \geq 2$,*

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},N}\hat{\theta} - f\right\|_P^2\right] \leq C'(C, B, \sigma)\left(\frac{\log N}{N}\right)^{2\beta/(2\beta+1)}.$$

Here again, the proof is given at the end of the paper (Section 5.7). Let us just remark that, in the case where $\mathcal{X} = [0, 1]$, P is the Lebesgue measure, and $(\theta_k)_{k\in\mathbb{N}^*}$ is the trigonometric basis, the condition:

$$\left\|\overline{\theta}^m - f\right\|_P^2 \leq Cm^{-2\beta}$$

is satisfied for $C = C(\beta, L)$ as soon as $f \in W(\beta, L)$ where $W(\beta, L)$ is the Sobolev class:

$$\left\{ f \in \mathcal{L}^2 \colon\ f^{(\beta-1)} \text{ is absolutely continuous and } \int_0^1 f^{(\beta)}(x)^2 \lambda(\mathrm{d}x) \leq L^2 \right\}.$$

The minimax rate of convergence in $W(\beta, L)$ is $N^{-2\beta/(2\beta+1)}$, so we can see that our estimator reaches the best rate of convergence up to a $\log N$ factor with an unknown β.

4.3. Rate of convergence in Besov spaces

We here extend the previous result to the case of a Besov space $B_{s,p,q}$ in the case of a wavelet basis (see [11] or [12]).

Theorem 4.2. *Let us assume that $\mathcal{X} = [-A, A]$, that $P_{(X)}$ is uniform on $\mathcal{X}$ and that $(\psi_{j,k})_{j=0,\ldots,+\infty,k\in\{1,\ldots,2^j\}}$ is a wavelet basis, together with a function ϕ, satisfying the conditions given in* [11], *with ϕ and $\psi_{0,1}$ supported by $[-A, A]$. Let us assume that $f \in B_{s,p,q}$ with $s > \frac{1}{p}$, $1 \leq p, q \leq \infty$, with*:

$$B_{s,p,q} = \left\{ g \colon [-A, A] \to \mathbb{R},\ g(\cdot) = \alpha\phi(\cdot) + \sum_{j=0}^{\infty}\sum_{k=1}^{2^j} \beta_{j,k}\psi_{j,k}(\cdot), \right.$$
$$\left. \sum_{j=0}^{\infty} 2^{jq(s-1/2-1/p)} \left[\sum_{k=1}^{2^j} |\beta_{j,k}|^p \right]^{q/p} = \|g\|_{s,p,q}^q < +\infty \right\}$$

(*with obvious changes for $p = +\infty$ or $q = +\infty$*) *with unknown constants s, p and q and that for any x, $|f(x)| \leq B$ for a known constant B. Let us choose*:

$$\{\theta_1, \ldots, \theta_m\} = \{\phi\} \cup \left\{ \psi_{j,k},\ j = 1, \ldots, 2^{\lfloor \log N/\log 2 \rfloor},\ k = 1, \ldots, 2^j \right\}$$

(*so $\frac{N}{2} \leq m \leq N$*) *and $\varepsilon = N^{-2}$ in the definition of $\hat{\theta}$. Then we have*:

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},N}\hat{\theta} - f\right\|_P^2\right] = \mathcal{O}\left(\left(\frac{\log N}{N}\right)^{2s/(2s+1)} (\log N)^{(1-2/((1+2s)q))_+}\right).$$

Let us remark that we obtain nearly the same rate of convergence than in [11], namely the minimax rate of convergence up to a $\log N$ factor.

For the proof, see Section 5.7.

5. Proofs

The order of the proofs is exactly the order of apparition of the results in the paper, except for the first theorem (Theorem 2.1): its proof using lemmas proved in the transductive setting, it is given after the proof of the transductive theorems.

5.1. Proof of Theorems 2.4–2.6

First, we prove a lemma that is the basis of proofs of Theorems 2.4–2.6.

Lemma 5.1. *We have, for any* $\theta \in \Theta$, $\gamma > 0$ *and* $\eta \geq 0$:

$$P\exp(\gamma W_\theta - \eta) = \exp\left\{\frac{\gamma^2}{2}V(W_\theta) + \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta\right\},$$

and

$$P\exp(-\gamma W_\theta - \eta) = \exp\left\{\frac{\gamma^2}{2}V(W_\theta) - \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta\right\}.$$

Proof. For the first equality, we write:

$$\begin{aligned}
\log P\exp(\gamma W_\theta - \eta) &= \log P\exp(\gamma W_\theta) - \eta \\
&= \int_0^\gamma P_{\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta = \int_0^\gamma (\gamma-\beta) V_{\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta \\
&= \frac{\gamma^2}{2}V(W_\theta) + \int_0^\gamma \frac{(\gamma-\beta)^2}{2} M^3_{\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta \\
&= \frac{\gamma^2}{2}V(W_\theta) + \frac{\gamma^3}{2}\int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta}(W_\theta)\,\mathrm{d}\beta - \eta.
\end{aligned}$$

For the reverse equality, the proof is exactly the same, replacing γ by $-\gamma$. □

We can now give the proof of both theorems.

Proof of Theorem 2.4. Let us choose $k \in \{1,\dots,m\}$, for any $\lambda_k > 0$ and $\eta_k \geq 0$ we have:

$$\begin{aligned}
&P^{\otimes N}\exp\left\{\frac{\lambda_k}{N}\sum_{i=1}^N \big[Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\big] - \eta_k\right\} \\
&= \left\{P\exp\left[\frac{\lambda_k}{N}W_{\theta_k} - \frac{\eta_k}{N}\right]\right\}^N \\
&= \exp\left[\frac{\lambda_k^2}{2N}V(W_{\theta_k}) + \frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta - \eta_k\right]
\end{aligned}$$

by the first equality of Lemma 5.1. By the same way, using the reverse inequality we obtain:

$$\begin{aligned}
&P^{\otimes N}\exp\left\{\frac{\lambda_k}{N}\sum_{i=1}^N \big[P\big(Y\theta_k(X)\big) - Y_i\theta_k(X_i)\big] - \eta_k\right\} \\
&= \exp\left[\frac{\lambda_k^2}{2N}V(W_{\theta_k}) - \frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta - \eta_k\right].
\end{aligned}$$

So we obtain, for any $k \in \{1,\dots,m\}$, for any $\lambda_k > 0$ and $\eta_k \geq 0$:

$$\begin{aligned}
&P^{\otimes N}\exp\left\{\lambda_k\left|\frac{1}{N}\sum_{i=1}^N Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\right| - \eta_k\right\} \\
&\leq 2\exp\left[\frac{\lambda_k^2}{2N}V(W_{\theta_k}) - \eta_k\right]\cosh\left[\frac{\lambda_k^3}{2N^2}\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right] \\
&\leq 2\exp\left[\frac{\lambda_k^2}{2N}V(W_{\theta_k}) - \eta_k + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right)^2\right],
\end{aligned}$$

since, for any $x \in \mathbb{R}$, we have:

$$\cosh(x) \leq \exp\left(\frac{x^2}{2}\right).$$

Now, let us choose $\varepsilon > 0$ and put:

$$\eta_k = \frac{\lambda_k^2}{2N} V(W_{\theta_k}) + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right)^2 - \log\frac{\varepsilon}{2m}.$$

We obtain:

$$P^{\otimes N}\sum_{k=1}^m \exp\left\{\lambda_k\left|\frac{1}{N}\sum_{i=1}^N Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\right|\right.$$
$$\left. - \frac{\lambda_k^2}{2N} V(W_{\theta_k}) + \frac{\lambda_k^6}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right)^2 + \log\frac{\varepsilon}{2m}\right\} \leq \varepsilon$$

and so:

$$P^{\otimes N}\left[\forall k \in \{1,\ldots,m\}, \left|\frac{1}{N}\sum_{i=1}^N Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\right|\right.$$
$$\left. \leq \frac{\lambda_k}{2N} V(W_{\theta_k}) + \frac{\lambda_k^5}{8N^4}\left(\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right)^2 + \frac{\log(2m/\varepsilon)}{\lambda_k}\right] \geq 1-\varepsilon.$$

Now, we put:

$$\lambda_k = \sqrt{\frac{2N\log(2m/\varepsilon)}{V(W_{\theta_k})}}.$$

We obtain, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $k \in \{1,\ldots,m\}$:

$$\left|\frac{1}{N}\sum_{i=1}^N Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\right| \leq \sqrt{\frac{2V(W_{\theta_k})\log(2m/\varepsilon)}{N}}$$
$$+ \frac{\log^{5/2}(2m/\varepsilon)}{NV(W_{\theta_k})^3}\left(\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_k/N)W_{\theta_k}}(W_{\theta_k})\,\mathrm{d}\beta\right)^2.$$

For short, we take the notation of the theorem:

$$I_{\theta_k}(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\beta\gamma W_{\theta_k}}(W_{\theta_k}).$$

Now, dividing both sides by:

$$P\big[\theta_k(X)^2\big]$$

we obtain:

$$|\hat{\alpha}_k\mathcal{C}_k - \overline{\alpha}_k| \leq \frac{1}{P[\theta_k(X)^2]}\left[\sqrt{\frac{2V(W_{\theta_k})\log(2m/\varepsilon)}{N}} + \frac{I^2_{\theta_k}(\lambda_k/N)\log^{5/2}(2m/\varepsilon)}{NV(W_{\theta_k})^3}\right].$$

In order to conclude, just remark that:

$$R(\hat{\alpha}_k \mathcal{C}_k \theta_k) - R(\overline{\alpha}_k \theta_k) = |\hat{\alpha}_k \mathcal{C}_k - \overline{\alpha}_k|^2 P\big[\theta_k(X)^2\big]. \qquad \square$$

Proof of Theorem 2.5. Remark that, for any $\theta \in \Theta$:

$$V(W_\theta) = P\big(W_\theta^2\big) - P(W_\theta)^2,$$

we will deal with each term separately. For the first term, let us remark that we obtain the following result that is obtained exactly as Lemma 5.1. For any $\theta \in \Theta$:

$$P \exp\big\{\gamma\big[P\big(W_\theta^2\big) - W_\theta^2\big] - \eta\big\} = \exp\left\{\frac{\gamma^2}{2} V\big(W_\theta^2\big) + \frac{\gamma^3}{2} \int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_\theta^2}\big(W_\theta^2\big)\,\mathrm{d}\beta - \eta\right\}.$$

Let us apply this result to every θ_k for $k \in \{1, \ldots, m\}$:

$$P^{\otimes N} \exp\left\{\lambda_k \left[P\big(W_{\theta_k}^2\big) - \frac{1}{N}\sum_{i=1}^N Y_i^2 \theta_k(X_i)^2\right] - \eta_k\right\} = \exp\left\{\frac{\lambda_k^2}{2N} V\big(W_{\theta_k}^2\big) + \frac{\lambda_k^3}{2N} J_k\left(\frac{\lambda_k}{N}\right) - \eta_k\right\},$$

where:

$$J_\theta(\gamma) = \int_0^1 (1-\beta)^2 M^3_{\gamma\beta W_{\theta_k}^2}\big(W_\theta^2\big)\,\mathrm{d}\beta.$$

Taking

$$\eta_k = \frac{\lambda_k^2}{2N} V\big(W_{\theta_k}^2\big) + \frac{\lambda_k^3}{2N^2} J_{\theta_k}\left(\frac{\lambda_k}{N}\right) + \log\frac{2m}{\varepsilon}$$

and

$$\lambda_k = \sqrt{\frac{2N\log(2m/\varepsilon)}{V(W_{\theta_k}^2)}}$$

we obtain that the following inequality is satisfied with $P^{\otimes N}$-probability at least $1 - \frac{\varepsilon}{2}$, for any k:

$$\begin{aligned} P\big(W_{\theta_k}^2\big) &\le \frac{1}{N}\sum_{i=1}^N Y_i^2 \theta_k(X_i)^2 + \sqrt{\frac{2V(W_{\theta_k}^2)\log(2m/\varepsilon)}{N}} + \frac{\log(2m/\varepsilon)}{NV(W_{\theta_k}^2)} J_{\theta_k}\left(\sqrt{\frac{2\log(2m/\varepsilon)}{NV(W_{\theta_k}^2)}}\right) \\ &= \frac{1}{N}\sum_{i=1}^N Y_i^2 \theta_k(X_i)^2 + \mathcal{A}_k \end{aligned} \tag{5.1}$$

for short. Now, we try to upper bound the second term, $-P(W_\theta)^2$. Remark that, for any θ:

$$\begin{aligned} \left(\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i)\right)^2 - P(W_\theta)^2 &= \left(\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i) - P(W_\theta)\right)\left(\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i) + P(W_\theta)\right) \\ &\le \left|\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i) - P(W_\theta)\right| \\ &\quad \times \left\{2\left|\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i)\right| + \left|\frac{1}{N}\sum_{i=1}^N Y_i \theta(X_i) - P(W_\theta)\right|\right\}. \end{aligned}$$

Remember that in the proof of Theorem 2.4 we got the upper bound, with probability at least $1-\frac{\varepsilon}{2}$, for any k:

$$\left|\frac{1}{N}\sum_{i=1}^{N} Y_i\theta_k(X_i) - P\big(Y\theta_k(X)\big)\right| \leq \sqrt{\frac{2V(W_{\theta_k})\log(4m/\varepsilon)}{N}} + \frac{\log^{5/2}(4m/\varepsilon)}{NV(W_{\theta_k})^3} I_{\theta_k}\left(\sqrt{\frac{2\log(4m/\varepsilon)}{NV(W_{\theta_k})}}\right)^2,$$

that gives:

$$\begin{aligned}
-P(W_{\theta_k})^2 &\leq -\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\theta_k(X_i)\right)^2 + \left\{\sqrt{\frac{2V(W_{\theta_k})\log(4m/\varepsilon)}{N}} + \frac{\log^{5/2}(4m/\varepsilon)}{NV(W_{\theta_k})^3} I_{\theta_k}\left(\sqrt{\frac{2\log(4m/\varepsilon)}{NV(W_{\theta_k})}}\right)^2\right\} \\
&\quad \times \left\{2\left|\frac{1}{N}\sum_{i=1}^{N} Y_i\theta_k(X_i)\right| + \sqrt{\frac{2V(W_{\theta_k})\log(4m/\varepsilon)}{N}} + \frac{\log^{5/2}(4m/\varepsilon)}{NV(W_{\theta_k})^3} I_{\theta_k}\left(\sqrt{\frac{2\log(4m/\varepsilon)}{NV(W_{\theta_k})}}\right)^2\right\} \\
&= -\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\theta(X_i)\right)^2 + \mathcal{B}_k \qquad (5.2)
\end{aligned}$$

for short. Let us combine inequalities (5.1) and (5.2). We obtain that, with probability at least $1-\varepsilon$, for every k we have:

$$V(W_{\theta_k}) = P\big(W_{\theta_k}^2\big) - P(W_{\theta_k})^2 \leq \frac{1}{N}\sum_{i=1}^{N} Y_i^2\theta_k(X_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\theta_k(X_i)\right)^2 + \mathcal{A}_k + \mathcal{B}_k = \hat{V}_k + \mathcal{A}_k + \mathcal{B}_k. \qquad \square$$

Proof of Theorem 2.6. This proof is a variant of the proof of Theorem 2.4, the method it uses is due to Seeger [20]. Let us define, for any $i \in \{1,\dots,N\}$:

$$P_i(\cdot) = P^{\otimes N}(\cdot|Z_i).$$

Let us choose $(i,k) \in \{1,\dots,N\}\times\{1,\dots,m'(N)\}$, for any $\lambda_{i,k} = \lambda_{i,k}(Z_i) > 0$ and $\eta_{i,k} = \eta_{i,k}(Z_i) \geq 0$ we have:

$$\begin{aligned}
&P_i \exp\left\{\frac{\lambda_{i,k}}{N-1}\sum_{j\neq i}\big[Y_j\theta_{i,k}(X_j) - P\big(Y\theta_{i,k}(X)\big)\big] - \eta_{i,k}\right\} \\
&\quad \leq \exp\left[\frac{\lambda_{i,k}}{2(N-1)} V(W_{\theta_{i,k}}) + \frac{\lambda_{i,k}^3}{2(N-1)^2}\int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_{i,k}/N-1)W_{\theta_{i,k}}}(W_{\theta_{i,k}})\,\mathrm{d}\beta - \eta_{i,k}\right]
\end{aligned}$$

by the first equality of Lemma 5.1. In the same way, we obtain the reverse inequality and, combining both results, for any $(i,k) \in \{1,\dots,N\}\times\{1,\dots,m'(N)\}$, for any $\lambda_{i,k} > 0$ and $\eta_{i,k} \geq 0$:

$$\begin{aligned}
&P_i \exp\left\{\lambda_{i,k}\left|\frac{1}{N-1}\sum_{j\neq i} Y_j\theta_{i,k}(X_j) - P\big(Y\theta_{i,k}(X)\big)\right| - \eta_{i,k}\right\} \\
&\quad \leq 2\exp\left[\frac{\lambda_{i,k}^2}{2(N-1)} V(W_{\theta_{i,k}}) - \eta_{i,k}\right]\cosh\left[\frac{\lambda_{i,k}^3}{2(N-1)^2} I_{i,k}\right] \\
&\quad \leq 2\exp\left[\frac{\lambda_{i,k}^2}{2(N-1)} V(W_{\theta_{i,k}}) - \eta_{i,k} + \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2\right],
\end{aligned}$$

where:

$$I_{i,k} = \int_0^1 (1-\beta)^2 M^3_{(\beta\lambda_{i,k}/N)W_{\theta_{i,k}}}(W_{\theta_{i,k}})\,\mathrm{d}\beta$$

for short. Now, let us choose $\varepsilon > 0$ and put:

$$\eta_{i,k} = \frac{\lambda_{i,k}^2}{2(N-1)} V(W_{\theta_{i,k}}) + \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 - \log \frac{\varepsilon}{2Nm'(N)}.$$

We obtain:

$$\begin{aligned} P^{\otimes N} &\sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} \exp\left\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j \theta_{i,k}(X_j) - P\big(Y\theta_{i,k}(X)\big) \right| \right. \\ &\left. - \frac{\lambda_{i,k}^2}{2(N-1)} V(W_{\theta_{i,k}}) - \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 + \log \frac{\varepsilon}{2Nm'(N)} \right\} \\ = P^{\otimes N} &\sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} P_i \exp\left\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j \theta_{i,k}(X_j) - P\big(Y\theta_{i,k}(X)\big) \right| \right. \\ &\left. - \frac{\lambda_{i,k}^2}{2(N-1)} V(W_{\theta_{i,k}}) - \frac{\lambda_{i,k}^6}{8(N-1)^4} I_{i,k}^2 + \log \frac{\varepsilon}{2Nm'(N)} \right\} \leq \varepsilon. \end{aligned}$$

Now, we put:

$$\lambda_{i,k} = \sqrt{\frac{2N \log(2Nm'(N)/\varepsilon)}{V(W_{\theta_{i,k}})}},$$

and achieve the proof exactly as for Theorem 2.4. □

5.2. Proof of Theorems 3.1 *and* 3.2

Here again, the first thing to do is to prove a general deviation inequality. This one is a variant of the one given by Catoni [6]. We go back to the notations of Theorem 3.1 and 3.2, with test sample of size N.

Definition 5.1. *Let $\mathcal{G}$ denote the set of all functions*:

$$\begin{aligned} &g : (\mathcal{X} \times \mathbb{R})^{2N} \times \mathbb{R}^2 \to \mathbb{R}, \\ &\big(Z_1, \ldots, Z_{2N}, u, u'\big) \mapsto g\big(Z_1, \ldots, Z_{2N}, u, u'\big) = g\big(u, u'\big) \end{aligned}$$

for the sake of simplicity, such that g is exchangeable with respect to its $2N$ first arguments.

Lemma 5.2. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, \ldots, Z_{2N})$, for any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$ and any $g \in \mathcal{G}$:*

$$\mathcal{P} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \big\{ g\big[\theta(X_{i+N}), Y_{i+N}\big] - g\big[\theta(X_i), Y_i\big] \big\} - \frac{\lambda^2}{c_g N^2} \sum_{i=1}^{2N} g\big[\theta(X_i), Y_i\big]^2 - \eta \right) \leq \mathcal{P} \exp(-\eta)$$

and the reverse inequality:

$$\mathcal{P} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \big\{ g\big[\theta(X_i), Y_i\big] - g\big[\theta(X_{i+N}), Y_{i+N}\big] \big\} - \frac{\lambda^2}{c_g N^2} \sum_{i=1}^{2N} g\big[\theta(X_i), Y_i\big]^2 - \eta \right) \leq \mathcal{P} \exp(-\eta),$$

where we write:

$$\eta = \eta\big((X_1, Y_1), \ldots, (X_{2N}, Y_{2N})\big),$$
$$\lambda = \lambda\big((X_1, Y_1), \ldots, (X_{2N}, Y_{2N})\big)$$

for short, *and*:

$$c_g = \begin{cases} 2 & \text{if } g \text{ is nonnegative,} \\ 1 & \text{otherwise.} \end{cases}$$

Proof. In order to prove the first inequality, we write:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\big\{g\big[\theta(X_{i+N}), Y_{i+N}\big] - g\big[\theta(X_i), Y_i\big]\big\} - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N} g\big[\theta(X_i), Y_i\big]^2 - \eta\right)$$
$$= \mathcal{P}\exp\left(\sum_{i=1}^{N}\log\cosh\left\{\frac{\lambda}{N}g\big[\theta(X_{i+N}), Y_{i+N}\big] - \frac{\lambda}{N}g\big[\theta(X_i), Y_i\big]\right\} - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N} g\big[\theta(X_i), Y_i\big]^2 - \eta\right).$$

This last step is true because $\mathcal{P}$ is exchangeable. We conclude by using the inequality:

$$\forall x \in \mathbb{R}, \quad \log\cosh x \leq \frac{x^2}{2}.$$

We obtain:

$$\log\cosh\left\{\frac{\lambda}{N}g\big[\theta(X_{i+N}), Y_{i+N}\big] - \frac{\lambda}{N}g\big[\theta(X_i), Y_i\big]\right\} \leq \frac{\lambda^2}{2N^2}\big\{g\big[\theta(X_{i+N}), Y_{i+N}\big] - g\big[\theta(X_i), Y_i\big]\big\}^2$$
$$\leq \frac{\lambda^2}{c_g N^2} g\big[\theta(X_i), Y_i\big]^2.$$

The proof for the reverse inequality is exactly the same. □

We can now give the proof of the theorems.

Proof of Theorem 3.1. From now on we assume that the hypothesis of Theorem 3.1 are satisfied. Let us choose $\varepsilon' > 0$ and apply Lemma 5.2 with $\eta = -\log\varepsilon'$, and g such that $g(u, u') = uu'$. We obtain: for any exchangeable distribution $\mathcal{P}$, for any measurable function λ: $(\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\big[\theta(X_{i+N})Y_{i+N} - \theta(X_i)Y_i\big] - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N}\theta(X_i)^2 Y_i^2 + \log\varepsilon'\right) \leq \varepsilon'$$

and the reverse inequality:

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\big[\theta(X_i)Y_i - \theta(X_{i+N})Y_{i+N}\big] - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N}\theta(X_i)^2 Y_i^2 + \log\varepsilon'\right) \leq \varepsilon'.$$

Let us denote:

$$f\big(\theta, \varepsilon', \lambda\big) = \lambda\left|\frac{1}{N}\sum_{i=1}^{N}\big[\theta(X_{i+N})Y_{i+N} - \theta(X_i)Y_i\big] - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N}\theta(X_i)^2 Y_i^2\right| + \log\varepsilon'.$$

The previous inequalities imply that: for any exchangeable $\mathcal{P}$, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$:

$$\mathcal{P} \exp f\big((Z_1, \ldots, Z_{2N}), \theta, \varepsilon', \lambda\big) \leq 2\varepsilon'. \tag{5.3}$$

Now, let us introduce a new conditional probability measure:

$$\overline{P} = \frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} \delta_{(X_{\sigma_i}, Y_{\sigma_i})_{i \in \{1,\ldots,2N\}}}.$$

Remark that P_{2N} being exchangeable, we have, for any bounded function $h : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$,

$$P_{2N} h = P_{2N}(\overline{P} h).$$

The measure $\overline{P}$ is exchangeable, so we can apply Eq. (5.3). For any values of $Z_1, \ldots, Z_{2N}$ we have:

$$\forall \theta \in \Theta, \quad \overline{P} \exp f\big((Z_1, \ldots, Z_{2N}), \theta, \varepsilon', \lambda\big) \leq 2\varepsilon'.$$

In particular, we can choose $\theta = \theta(Z_1, \ldots, Z_{2N})$ as an exchangeable function of $(Z_1, \ldots, Z_{2N})$, because we will have:

$$\begin{aligned}
&\frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} \exp f\big((Z_{\sigma(1)}, \ldots, Z_{\sigma(2N)}), \theta(Z_{\sigma(1)}, \ldots, Z_{\sigma(2N)}), \varepsilon', \lambda\big) \\
&\quad = \frac{1}{(2N)!} \sum_{\sigma \in \mathfrak{S}_{2N}} \exp f\big((Z_{\sigma(1)}, \ldots, Z_{\sigma(2N)}), \theta(Z_1, \ldots, Z_{2N}), \varepsilon', \lambda\big) \leq \varepsilon'.
\end{aligned}$$

Here, we choose as functions θ the members of Θ_0: $\theta_1, \ldots, \theta_m$ (remember that we choose this indexation in such a way that for any k, θ_k is an exchangeable function of $(Z_1, \ldots, Z_{2N})$). We have, for any $\lambda_1, \ldots, \lambda_m$ that are m exchangeable functions of $(Z_1, \ldots, Z_{2N})$:

$$\begin{aligned}
&P_{2N}\big[\exists k \in \{1, \ldots, m\}, f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0\big] \\
&\quad = P_{2N}\left[\bigcup_{k=1}^{m} \big\{ f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0 \big\}\right] \\
&\quad \leq P_{2N}\left[\sum_{k=1}^{m} 1\big(f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0\big)\right] \\
&\quad = P_{2N}\overline{P}\left[\sum_{k=1}^{m} 1\big(f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0\big)\right] \\
&\quad = P_{2N} \sum_{k=1}^{m} \overline{P}\big[1\big(f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0\big)\big] \\
&\quad \leq P_{2N} \sum_{k=1}^{m} \overline{P} \exp f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big).
\end{aligned}$$

Now let us apply inequality (5.3), we obtain:

$$P_{2N}\big[\exists k \in \{1, \ldots, m\}, f\big((Z_1, \ldots, Z_{2N}), \theta_k, \varepsilon', \lambda_k\big) > 0\big] \leq P_{2N} \sum_{k=1}^{m} 2\varepsilon' = 2\varepsilon' m = \varepsilon$$

if we choose:

$$\varepsilon' = \frac{\varepsilon}{2m}.$$

From now, we assume that the event:

$$\left\{\forall k \in \{1,\dots,m\},\ f\left((Z_1,\dots,Z_{2N}),\theta_k,\frac{\varepsilon}{2m},\lambda_k\right) \le 0\right\}$$

is satisfied. It can be written, for any $k \in \{1,\dots,m\}$:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\left[\theta_k(X_{i+N})Y_{i+N} - \theta_k(X_i)Y_i\right]\right| \le \frac{\lambda_k}{N^2}\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2 + \frac{\log(2m/\varepsilon)}{\lambda_k}.$$

Let us divide both inequalities by:

$$\frac{1}{N}\sum_{i=N+1}^{2N}\theta_k(X_i)^2.$$

We obtain, for any $k \in \{1,\dots,m\}$:

$$\left|\alpha_2^k - \mathcal{C}^k\alpha_1^k\right| \le \frac{(\lambda_k/N^2)\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2 + (\log(2m/\varepsilon))/\lambda_k}{(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2}.$$

It is now time to choose the functions λ_k. We try to optimize the right-hand side with respect to λ_k, and obtain a minimal value for:

$$\lambda_k = \sqrt{\frac{N\log(2m/\varepsilon)}{(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2}}.$$

This choice is admissible because it is exchangeable with respect to $(Z_1,\dots,Z_{2N})$.

So we have, for any $k \in \{1,\dots,m\}$:

$$\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right| \le 2\frac{\sqrt{(1/N^2)\sum_{i=1}^{2N}[\theta_k(X_i)^2Y_i^2]\log(2m/\varepsilon)}}{(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2}.$$

Finally, remark that:

$$\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right| = \sqrt{\frac{r_2[(\mathcal{C}^k\alpha_1^k)\theta_k] - r_2(\alpha_2^k\theta_k)}{(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2}},$$

which leads to the conclusion that for any $k \in \{1,\dots,m\}$:

$$r_2\left[\left(\mathcal{C}^k\alpha_1^k\right)\theta_k\right] - r_2\left(\alpha_2^k\theta_k\right) \le 2^2\frac{(1/N^2)\sum_{i=1}^{2N}[\theta_k(X_i)^2Y_i^2]\log(2m/\varepsilon)}{(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2}.$$

This ends the proof. □

Proof of Theorem 3.2. We write:

$$\frac{1}{N}\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2 = \frac{1}{N}\sum_{i=1}^{N}\theta_k(X_i)^2Y_i^2 + \frac{1}{N}\sum_{i=N+1}^{2N}\theta_k(X_i)^2Y_i^2$$

and try to upper bound the second term. We apply Lemma 5.2, but this time with g such that $g(u) = (uu')^2$ that is nonnegative, and obtain, for any ε, for any (exchangeables) θ and λ:

$$\frac{1}{N}\sum_{i=N+1}^{2N}\theta_k(X_i)^2Y_i^2 \le \frac{1}{N}\sum_{i=1}^{N}\theta_k(X_i)^2Y_i^2 + \frac{\lambda}{2N}\left(\frac{1}{N}\right)\sum_{i=1}^{2N}\theta_k(X_i)^4Y_i^4 + \frac{\log\varepsilon}{\lambda}.$$

We choose:

$$\lambda = \sqrt{\frac{2N\log\varepsilon}{(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^4Y_i^4}},$$

we apply this result to every $\theta \in \Theta_0$, and combine it with Theorem 3.1 by a union bound argument to obtain the result. □

5.3. Proof of Theorem 3.3

First of all, we give the following obvious variant of Lemma 5.2:

Lemma 5.3. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, \dots, Z_{2N})$, for any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\theta \in \Theta$:*

$$\mathcal{P}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\left\{\left[\theta(X_{i+N})Y_{i+N} - \alpha(\theta)\theta(X_{i+N})^2\right] - \left[\theta(X_i)Y_i - \alpha(\theta)\theta(X_i)^2\right]\right\} - \frac{\lambda^2}{N^2}\sum_{i=1}^{2N}\left[\theta(X_i)Y_i - \alpha(\theta)\theta(X_i)^2\right]^2 - \eta\right) \le \mathcal{P}\exp(-\eta)$$

and the reverse inequality, where:

$$\alpha(\theta) = \arg\min_{\alpha\in\mathbb{R}} r_{1,2}(\alpha\theta).$$

Proof. This is actually just an application of Lemma 5.2, we just need to remark that $\alpha(\theta)$ is an exchangeable function of $(Z_1, \dots, Z_{2N})$, and so we can take in Lemma 5.2:

$$g\big(u, u'\big) = uu' - u^2\alpha(\theta),$$

that means that:

$$g\big[\theta(X_i), Y_i\big] = \theta(X_i)Y_i - \alpha(\theta)\theta(X_i)^2.$$ □

Proof of Theorem 3.3. Proceeding exactly in the same way as in the proof of Theorem 3.1, we obtain the following inequality with probability at least $1 - \varepsilon$:

$$r_2\big(\mathcal{C}^k\alpha_1^k\theta_k\big) - r_2\big(\alpha_2^k\theta_k\big) \le 4\left[\frac{(1/N)\sum_{i=1}^{2N}[\theta_k(X_i)Y_i - \alpha_{1,2}^k\theta_k(X_i)^2]^2}{(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2}\right]\frac{\log(2m/\varepsilon)}{N}. \tag{5.4}$$

This proves the theorem. □

Before giving the proof of the next theorem, let us see how we can make the first-order term observable in this theorem. For example, we can write:

$$\left[\theta_k(X_i)Y_i - \alpha_{1,2}^k\theta_k(X_i)^2\right]^2 = \left[\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right]^2 + \left[\alpha_1^k - \alpha_{1,2}^k\right]^2\theta_k(X_i)^4 + 2\left[\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right]\left[\alpha_1^k - \alpha_{1,2}^k\right]\theta_k(X_i)^2.$$

Remark that it is obvious that:

$$\left|\alpha_1^k - \alpha_{1,2}^k\right| \leq \left|\alpha_1^k - \alpha_2^k\right|,$$

and so:

$$\left[\theta_k(X_i)Y_i - \alpha_{1,2}^k\theta_k(X_i)^2\right]^2 \leq \left[\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right]^2 + \left[\alpha_1^k - \alpha_2^k\right]^2\theta_k(X_i)^4 + 2\left|\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right|\left|\alpha_1^k - \alpha_2^k\right|\theta_k(X_i)^2.$$

Now, just write:

$$\alpha_1^k - \alpha_2^k = \left(1 - \mathcal{C}^k\right)\alpha_1^k - \left(\mathcal{C}^k\alpha_1^k - \alpha_2^k\right)$$

and so we get:

$$\begin{aligned}\left[\theta_k(X_i)Y_i - \alpha_{1,2}^k\theta_k(X_i)^2\right]^2 \leq{}& \left[\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right]^2 + \left[\mathcal{C}^k\alpha_1^k - \alpha_2^k\right]^2\theta_k(X_i)^4 \\ &+ 2\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right|\left|\left(1 - \mathcal{C}^k\right)\alpha_1^k\right|\theta_k(X_i)^4 + \left(1 - \mathcal{C}^k\right)^2\left(\alpha_1^k\right)^2\theta_k(X_i)^4 \\ &+ 2\left|\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right|\left|\mathcal{C}^k\alpha_1^k - \alpha_2^k\right|\theta_k(X_i)^2 \\ &+ 2\left|\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right|\left|\left(\mathcal{C}^k - 1\right)\alpha_1^k\right|\theta_k(X_i)^2.\end{aligned}$$

So finally, Eq. (5.4) left us with a second degree inequality with respect to $|\mathcal{C}^k\alpha_1^k - \alpha_2^k|$ or $r_2(\mathcal{C}^k\alpha_1^k\theta_k) - r_2(\alpha_2^k\theta_k)$ that we can solve to obtain the following result: with probability at least $1 - \varepsilon$, as soon as we have:

$$\left[\frac{1}{N}\sum_{i=N+1}^{2N}\theta_k(X_i)^2\right]^2 > \left[\frac{1}{N}\sum_{i=1}^{2N}\theta_k(X_i)^4\right]\frac{4\log(2m/\varepsilon)}{N},$$

which is always true for large enough N, the quantity $|\mathcal{C}^k\alpha_1^k - \alpha_2^k|$ belongs to the interval:

$$\left[\frac{2\log(2m/\varepsilon)}{N}\frac{b \pm \sqrt{b^2 + a((N/\log(2m/\varepsilon))[(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2]^2 - (4/N)\sum_{i=1}^{2N}\theta_k(X_i)^4)}}{[(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2]^2 - (4\log(2m/\varepsilon)/N)[(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^4]}\right]$$

with the following notations:

$$a = \frac{1}{N}\sum_{i=1}^{2N}\left[\left|\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right| + \left|\alpha_k^1\left(1 - \mathcal{C}^k\right)\right|\theta_k(X_i)^2\right]^2,$$

$$b = \frac{1}{N}\sum_{i=1}^{2N}2\theta_k(X_i)^2\left[\left|\alpha_k^1\left(1 - \mathcal{C}^k\right)\right|\theta_k(X_i)^2 + \left|\theta_k(X_i)Y_i - \alpha_1^k\theta_k(X_i)^2\right|\right].$$

Remark that only one of the bounds of the interval is positive. So we obtain the following result: with P_{2N}-probability at least $1 - \varepsilon$, as soon as:

$$\left[\frac{1}{N}\sum_{i=N+1}^{2N}\theta_k(X_i)^2\right]^2 > \left[\frac{1}{N}\sum_{i=1}^{2N}\theta_k(X_i)^4\right]\frac{4\log(2m/\varepsilon)}{N}$$

we have:

$$\begin{aligned}
&\forall k \in \{1,\dots,m\},\\
&r_2\big[\big(\mathcal{C}^k\alpha_1^k\big)\theta_k\big] - r_2\big(\alpha_2^k\theta_k\big)\\
&\quad\le \frac{4\log^2(2m/\varepsilon)}{N^2}\left[\frac{1}{N}\sum_{i=1}^{2N}\theta_k(X_i)^2\right]\\
&\qquad\times\left[\frac{b+\sqrt{b^2+a((N/\log(2m/\varepsilon))[(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2]^2-(4/N)\sum_{i=1}^{2N}\theta_k(X_i)^4)}}{[(1/N)\sum_{i=N+1}^{2N}\theta_k(X_i)^2]^2-((4\log(2m/\varepsilon))/N)[(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^4]}\right]^2.
\end{aligned}$$

We can notice that this bound may be written:

$$\begin{aligned}
r_2\big[\big(\mathcal{C}^k\alpha_1^k\big)\theta_k\big] - r_2\big(\alpha_2^k\theta_k\big) &\le \frac{8a\log(2m/\varepsilon)}{N} + \mathcal{O}\left(\left[\frac{\log(m/\varepsilon)}{N}\right]^{3/2}\right)\\
&= \frac{8\log(2m/\varepsilon)}{N}\left[\frac{1}{N}\sum_{i=1}^{2N}\big(\theta_k(X_i)Y_i-\alpha_1^k\theta_k(X_i)^2\big)^2\right] + \mathcal{O}\left(\left[\frac{\log(m/\varepsilon)}{N}\right]^{3/2}\right).
\end{aligned}$$

The next step would be now to replace the bound by an observable quantity, by getting a bound like:

$$\frac{1}{N}\sum_{i=1}^{2N}\big(\theta_k(X_i)Y_i-\alpha_1^k\theta_k(X_i)^2\big)^2 \le \frac{2}{N}\sum_{i=1}^{N}\big(\theta_k(X_i)Y_i-\alpha_1^k\theta_k(X_i)^2\big)^2 + \mathcal{O}\left(\frac{\log(m/\varepsilon)}{N}\right)$$

with high probability. This can be done very simply, using Lemma 5.2 with this time:

$$g\big(u,u'\big) = \big(uu'-u^2\alpha(\theta)\big)^2.$$

We obtain the bound:

$$r_2\big[\big(\mathcal{C}^k\alpha_1^k\big)\theta_k\big] - r_2\big(\alpha_2^k\theta_k\big) \le \frac{16\log(4m/\varepsilon)}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\big(\theta_k(X_i)Y_i-\alpha_1^k\theta_k(X_i)^2\big)^2\right] + \mathcal{O}\left(\left[\frac{\log(m/\varepsilon)}{N}\right]^{3/2}\right).$$

5.4. Proof of Theorem 3.4

The proof is exactly similar, we just use a new variant of lemma 5.2, that is based on an idea introduced by Catoni [8] in the context of classification.

Definition 5.2. *Let us write*:

$$T_\theta(Z_i) = \theta(X_i)Y_i$$

for short. We also introduce a conditional probability measure:

$$\mathcal{P}^{(2)} = \frac{1}{N!}\sum_{\sigma\in\mathfrak{S}_N}\delta_{(Z_1,\dots,Z_N,Z_{N+\sigma(1)},\dots,Z_{N+\sigma(N)})}.$$

Remark that, because $\mathcal{P}$ is exchangeable, we have, for any function h:

$$\mathcal{P}h = \mathcal{P}\big[\mathcal{P}^{(2)}h\big].$$

Lemma 5.4. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, \ldots, Z_{2N})$, for any measurable function $\eta : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any measurable function $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+^*$ which is such that, for any $i \in \{1, \ldots, 2N\}$:*

$$\lambda(Z_1, \ldots, Z_{2N}) = \lambda(Z_1, \ldots, Z_{i-1}, Z_{i+N}, Z_{i+1}, \ldots, Z_{i+N-1}, Z_i, Z_{i+N+1}, \ldots, Z_{2N}),$$

for any $\theta \in \Theta$:

$$\mathcal{P} \exp\left\{ \frac{\mathcal{P}^{(2)}\lambda}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] - \mathcal{P}^{(2)}\left[\frac{\lambda^2}{2N^2} \frac{1}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big]^2 \right] - \eta \right\} \leq \mathcal{P} \exp(-\eta)$$

and the reverse inequality.

Proof. Let $\mathcal{L}hs$ denote the left-hand side of Lemma 5.4. For short, let us put:

$$s(\theta) = \frac{1}{N} \sum_{i=1}^{N} \big[\theta(X_{i+N})Y_{i+N} - \theta(X_i)Y_i\big]^2 = \frac{1}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big]^2.$$

Then we have:

$$\begin{aligned} \mathcal{L}hs &= P_{2N} \exp P^{(2)} \left(\frac{\lambda}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] - \frac{\lambda^2}{2N} s(\theta) - \eta \right) \\ &\leq P_{2N} P^{(2)} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] - \frac{\lambda^2}{2N} s(\theta) - \eta \right), \end{aligned}$$

by Jensen's conditional inequality. Now, we can conclude as in Lemma 5.2:

$$\begin{aligned} \mathcal{L}hs &= P_{2N} \exp\left(\sum_{i=1}^{N} \log\cosh\left\{ \frac{\lambda}{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] \right\} - \frac{\lambda^2}{2N} s(\theta) - \eta \right) \\ &\leq P_{2N} \exp\left(\frac{\lambda^2}{2N^2} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big]^2 - \frac{\lambda^2}{2N} s(\theta) - \eta \right) \\ &= P_{2N} \exp(-\eta). \end{aligned}$$

□

Proof of Theorem 3.4. We apply both inequalities of Lemma 5.4 to every $\theta_k, k \in \{1, \ldots, m\}$, and we take:

$$\lambda = \sqrt{\frac{2N \log(2m/\varepsilon)}{s(\theta)}}.$$

We obtain, for any $k \in \{1, \ldots, m\}$:

$$\mathcal{P} \exp\left\{ \frac{\mathcal{P}^{(2)}\lambda}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] - \log\frac{2m}{\varepsilon} - \eta \right\} \leq \varepsilon.$$

Or, with probability at least $1 - \varepsilon$, for any k:

$$\frac{1}{N} \sum_{i=1}^{N} \big[T_\theta(Z_i) - T_\theta(Z_{i+N})\big] \leq \sqrt{\frac{2\log(2m/\varepsilon)}{N}} \big[\mathcal{P}^{(2)}\big(s(\theta)^{-1/2}\big)\big]^{-1},$$

so:

$$\left[\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right]^2 \leq \frac{2\log(2m/\varepsilon)}{N}\mathcal{P}^{(2)}s(\theta).$$

We end the first part of the proof by noting that:

$$\mathcal{P}^{(2)}s(\theta) = V_1(\theta) + V_2(\theta) + \left[\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right]^2.$$

Now, let us see how we can obtain the second part of the theorem. Note that:

$$V_2(\theta) = \frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)^2 - \left(\frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right)^2.$$

We upper bound the first term by using Lemma 5.2 with $g(\theta(X_i), Y_i) = \theta(X_i)^2 Y_i^2 = T_\theta(Z_i)^2$, so with probability at least $1-\varepsilon$, for any k:

$$\frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)^2 \leq \frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i)^2 + \sqrt{\frac{2\log(m/\varepsilon)(1/N)\sum_{i=1}^{2N} T_\theta(Z_i)^4}{N}}.$$

For the second-order term, we use both inequalities of Lemma 5.2 with $g(\theta(X_i), Y_i) = \theta(X_i)Y_i = T_\theta(Z_i)$, so with probability at least $1-\varepsilon$, for any k:

$$\begin{aligned}\left(\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i)\right)^2 - \left(\frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right)^2 &\leq \left|\frac{1}{N}\sum_{i=1}^{N} T_\theta(Z_i) - \frac{1}{N}\sum_{i=N+1}^{2N} T_\theta(Z_i)\right|\left|\frac{1}{N}\sum_{i=1}^{2N} T_\theta(Z_i)\right| \\ &\leq 2\sqrt{\frac{(1/N)\sum_{i=1}^{2N} T_\theta(Z_i)^2\log(2m/\varepsilon)}{N}}\,\frac{1}{N}\sum_{i=1}^{2N}\left|T_\theta(Z_i)\right|.\end{aligned}$$

Putting all pieces together (and replacing ε by $\varepsilon/3$) ends the proof. □

5.5. Proof of Theorem 3.5

Proof of Theorem 3.5. We introduce the following conditional probability measures, for any $i \in \{1, \ldots, N\}$:

$$\begin{aligned}\mathrm{P}_i = \frac{1}{(k+1)!} &\\ \times \sum_{\sigma\in\mathfrak{S}_{k+1}} &\delta_{(Z_1,\ldots,Z_{i-1},Z_{N(\sigma(1)-1)+i},Z_{i+1},\ldots,Z_{N+i-1},Z_{N(\sigma(2)-1)+i},Z_{N+i+1},\ldots,\ldots,Z_{kN+i-1},Z_{N(\sigma(k+1)-1)+i},Z_{kN+i+1},\ldots,Z_{(k+1)N})}.\end{aligned}$$

and

$$\mathrm{P} = \bigotimes_{i=1}^{N} \mathrm{P}_i$$

and, finally, remember that:

$$\mathbf{P} = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N} \delta_{Z_i}.$$

Note that, by exchangeability, for any nonnegative function

$$h:(\mathcal{X}\times\mathbb{R})^{(k+1)N}\to\mathbb{R}$$

we have, for any $i\in\{1,\dots,N\}$:

$$P_{(k+1)N}\mathrm{P}_i h(Z_1,\dots,Z_{2N})=P_{(k+1)N}h(Z_1,\dots,Z_{2N}).$$

Lemma 5.5. *Let χ be a function $\mathbb{R}\to\mathbb{R}$. For any exchangeable functions λ, $\eta:(\mathcal{X}\times\mathbb{R})^{(k+1)N}\to\mathbb{R}_+$ and $\theta:(\mathcal{X}\times\mathbb{R})^{(k+1)N}\to\Theta$ we have*:

$$\begin{aligned}
&\mathrm{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\chi\big[\theta(X_i)Y_i\big]-\frac{1}{N}\sum_{i=1}^{N}\chi\big[\theta(X_i)Y_i\big]\right]-\eta\right\}\\
&\quad\le\exp(-\eta)\exp\left\{\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbf{P}\big\{\big[\chi\big(\theta(X)Y\big)-\mathbf{P}\chi\big(\theta(X)Y\big)\big]^2\big\}\right.\\
&\qquad\left.+\frac{\lambda^3(1+k)^3}{6N^2k^3}\Big[\sup_{i\in\{1,\dots,(k+1)N\}}\chi\big(\theta(X_i)Y_i\big)-\inf_{i\in\{1,\dots,(k+1)N\}}\chi\big(\theta(X_i)Y_i\big)\Big]^3\right\},
\end{aligned}$$

where we put $\lambda=\lambda(Z_1,\dots,Z_{(k+1)N})$, $\theta=\theta(Z_1,\dots,Z_{(k+1)N})$ and $\eta=\eta(Z_1,\dots,Z_{(k+1)N})$ for short. We have the reverse inequality as well.

Before giving the proof, let us introduce the following useful notations.

Definition 5.3. *We put, for any $\theta\in\Theta$, for any function χ:*

$$\chi_i^\theta=\chi\big(Y_i\theta(X_i)\big),$$

and

$$\chi^\theta=\chi\big(Y\theta(X)\big)$$

that means that:

$$\mathbf{P}\chi^\theta=\frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\chi_i^\theta.$$

We also put:

$$\mathcal{S}_\chi(\theta)=\sup_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta-\inf_{i\in\{1,\dots,(k+1)N\}}\chi_i^\theta.$$

Proof of the Lemma 5.5. Remark that, for any exchangeable functions λ, $\eta:(\mathcal{X}\times\mathbb{R})^{(k+1)N}\to\mathbb{R}_+$ and $\theta:(\mathcal{X}\times\mathbb{R})^{kN}\to\Theta$ we have:

$$\begin{aligned}
&\mathrm{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}g\big[\theta(X_i)Y_i\big]-\frac{1}{N}\sum_{i=1}^{N}g\big[\theta(X_i)Y_i\big]\right]-\eta\right\}\\
&\quad=\exp(-\eta)\prod_{i=1}^{N}\mathrm{P}_i\exp\left\{\frac{\lambda}{kN}\sum_{j=1}^{k}\chi_{i+jN}^\theta-\frac{\lambda}{N}\chi_i^\theta\right\}\\
&\quad=\exp(-\eta)\prod_{i=1}^{N}\exp\left\{\frac{\lambda}{kN}\sum_{j=0}^{k}\chi_{i+jN}^\theta\right\}\prod_{i=1}^{N}\mathrm{P}_i\exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^\theta\right\},
\end{aligned}$$

where we put $\lambda = \lambda(Z_1, \ldots, Z_{kN})$, $\theta = \theta(Z_1, \ldots, Z_{kN})$ and $\eta = \eta(Z_1, \ldots, Z_{kN})$ for short.

Now, we have:

$$\log \prod_{i=1}^{N} \mathrm{P}_i \exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^{\theta}\right\} = \sum_{i=1}^{N} \log \mathrm{P}_i \exp\left\{-\frac{\lambda(1+k)}{kN}\chi_i^{\theta}\right\},$$

and, for any $i \in \{1, \ldots, N\}$:

$$\begin{aligned}
&\log \mathrm{P}_i \exp\left\{-\frac{\lambda(1+k)}{Nk}\chi_i^{\theta}\right\} \\
&\quad = -\frac{\lambda(1+k)}{Nk}\mathrm{P}_i\chi_i^{\theta} + \frac{\lambda^2(1+k)^2}{2N^2k^2}\mathrm{P}_i\big[\big(\chi_i^{\theta} - \mathrm{P}_i\chi_i^{\theta}\big)^2\big] \\
&\qquad - \int_0^{\lambda(1+k)/(Nk)} \frac{1}{2}\left(\frac{\lambda(1+k)}{Nk} - \beta\right)^2 \frac{1}{\mathrm{P}_i \exp[-\beta\chi_i^{\theta}]}\mathrm{P}_i\left[\left(\chi_i^{\theta} - \frac{\mathrm{P}_i\{\chi_i^{\theta}\exp[-\beta\chi_i^{\theta}]\}}{\mathrm{P}_i \exp[-\beta\chi_i^{\theta}]}\right)^3 \exp\big(-\beta\chi_i^{\theta}\big)\right] \mathrm{d}\beta.
\end{aligned}$$

Note that, for any $\beta \geq 0$:

$$\frac{1}{\mathrm{P}_i \exp[-\beta\chi_i^{\theta}]}\mathrm{P}_i\left[\left(\chi_i^{\theta} - \frac{\mathrm{P}_i\{\chi_i^{\theta}\exp[-\beta\chi_i^{\theta}]\}}{\mathrm{P}_i \exp[-\beta\chi_i^{\theta}]}\right)^3 \exp\big(-\beta\chi_i^{\theta}\big)\right] \leq \Big[\sup_{j\in\{1,\ldots,k\}} \chi_{i+(j-1)N}^{\theta} - \inf_{j\in\{1,\ldots,k\}} \chi_{i+(j-1)N}^{\theta}\Big]^3,$$

and so:

$$\begin{aligned}
\log \prod_{i=1}^{N} \mathrm{P}_i \exp\left\{-\frac{\lambda(1+k)}{Nk}\chi_i^{\theta}\right\} &\leq -\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda(1+k)}{k}\mathrm{P}_i\chi_i^{\theta} + \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda^2(1+k)^2}{2Nk^2}\mathrm{P}_i\big[\big(\chi_i^{\theta} - \mathrm{P}_i\chi_i^{\theta}\big)^2\big] \\
&\quad + \frac{\lambda^3(1+k)^3}{6N^2k^3}\Big[\sup_{i\in\{1,\ldots,(k+1)N\}} \chi_i^{\theta} - \inf_{i\in\{1,\ldots,(k+1)N\}} \chi_i^{\theta}\Big]^3.
\end{aligned}$$

Note that:

$$\mathrm{P}_i\chi_i^{\theta} = \frac{1}{k+1}\sum_{j=0}^{k}\chi_{i+jN}^{\theta}$$

and so:

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{P}_i\chi_i^{\theta} = \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\chi_i^{\theta} = \mathbf{P}\chi^{\theta};$$

remark also that:

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{P}_i\big[\big(\chi_i^{\theta} - \mathbb{P}_i\chi_i^{\theta}\big)^2\big] \leq \frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\left[\chi_i^{\theta} - \left(\frac{1}{(k+1)N}\sum_{j=1}^{(k+1)N}\chi_j^{\theta}\right)\right]^2 = \mathbf{P}\big[\big(\chi^{\theta} - \mathbf{P}\chi^{\theta}\big)^2\big],$$

we obtain:

$$\begin{aligned}
&\mathrm{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\theta(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N}\theta(X_i)Y_i\right] - \eta\right\} \\
&\quad = \exp(-\eta)\exp\left\{\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbf{P}\big[\big(\chi^{\theta} - \mathbf{P}\chi^{\theta}\big)^2\big] + \frac{\lambda^3(1+k)^3}{6N^2k^3}\Big[\sup_{i\in\{1,\ldots,(k+1)N\}} \chi_i^{\theta} - \inf_{i\in\{1,\ldots,(k+1)N\}} \chi_i^{\theta}\Big]^3\right\}.
\end{aligned}$$

The proof of the reverse inequality is exactly the same. □

Let us choose here again χ such that $\chi(u)=u$, namely: $\chi = id$. By the use of a union bound argument on elements of Θ_0 we obtain, for any $\varepsilon > 0$, for any exchangeable function $\lambda : (\mathcal{X}\times\mathbb{R})^{(k+1)N} \to \mathbb{R}_+$, with probability at least $1-\varepsilon$, for any $h \in \{1,\dots,m\}$:

$$\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\theta_h(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N}\theta_h(X_i)Y_i$$
$$\leq \frac{\lambda(1+1/k)^2}{2N}\mathbf{P}\Big[\big(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h}\big)^2\Big] + \frac{\lambda^2(1+1/k)^3}{6N^2}\mathcal{S}_{id}(\theta_h)^3 + \frac{\log(m/\varepsilon)}{\lambda}.$$

Let us choose, for any $h \in \{1,\dots,m\}$:

$$\lambda = \sqrt{\frac{2N\log(m/\varepsilon)}{(1+1/k)^2\mathbf{P}[(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h})^2]}},$$

the bound becomes:

$$\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\theta_h(X_i)Y_i - \frac{1}{N}\sum_{i=1}^{N}\theta_h(X_i)Y_i$$
$$\leq \left(1+\frac{1}{k}\right)\left[2\sqrt{\frac{\mathbf{P}[(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h})^2]\log(m/\varepsilon)}{2N}} + \frac{\mathcal{S}_{id}(\theta_h)^3\log(m/\varepsilon)}{3N\mathbf{P}[(\chi^{\theta_h}-\mathbf{P}\chi^{\theta_h})^2]}\right].$$

We use the reverse inequality exactly in the same way, we then combine both inequality by a union bound argument and obtain the following result. For any $\varepsilon > 0$, with $P_{(k+1)N}$ probability at least $1-\varepsilon$ we have, for any $h \in \{1,\dots,m\}$:

$$r_2\big(\mathcal{C}^h\alpha_1^h\theta_h\big) - r_2\big(\alpha_2^h\theta_h\big) \leq \frac{(1+1/k)^2}{(1/(kN))\sum_{i=N+1}^{(k+1)N}\theta_h(X_i)^2}\left[\frac{2\mathbb{V}_{\theta_h}\log(2m/\varepsilon)}{N}\right.$$
$$\left. + \frac{2(\log(2m/\varepsilon))^{3/2}\mathcal{S}_{id}(\theta_h)^3}{3N^{3/2}\mathbb{V}_{\theta_h}^{1/2}} + \frac{(\log(2m/\varepsilon))^2\mathcal{S}_{id}(\theta_h)^6}{9N^2\mathbb{V}_{\theta_h}^2}\right], \tag{5.5}$$

remember that:

$$\mathbb{V}_\theta = \mathbf{P}\Big\{\big[\big(\theta(X)Y\big) - \mathbf{P}\big(\theta(X)Y\big)\big]^2\Big\}.$$

We now give a new lemma.

Lemma 5.6. *Let us assume that P is such that, for any $h \in \{1,\dots,m\}$:*

$$\exists\beta_h > 0,\ \exists B_h \geq 0, \quad P\exp\big(\beta_h\big|\theta_h(X)Y\big|\big) \leq B_h.$$

This is for example the case if $\theta_h(X_i)Y_i$ is subgaussian, with any $\beta_h > 0$ and

$$B_h = 2\exp\left\{\frac{\beta_h^2}{2}P\Big[\big(\theta_h(X)Y\big)^2\Big]\right\}.$$

Then we have, for any $\varepsilon \geq 0$:

$$P_{(k+1)N}\left\{\sup_{1\leq i\leq (k+1)N}\theta_h(X_i)Y_i \leq \frac{1}{\beta_h}\log\frac{(k+1)NB_h}{\varepsilon}\right\} \geq 1-\varepsilon.$$

Proof. We have:

$$P_{(k+1)N}\Big(\sup_{1\leq i\leq (k+1)N}\theta_h(X_i)Y_i\geq s\Big)=P_{(k+1)N}\big(\exists i\in\{1,\dots,(k+1)N\},\theta_h(X_i)Y_i\geq s\big)$$
$$=\sum_{i=1}^{(k+1)N}P1_{\theta_h(X_i)Y_i\geq s}$$
$$\leq (k+1)NP\exp\big(\beta_h\big|\theta_h(X_i)Y_i-s\big|\big)\leq (k+1)NB_h\exp(-\beta_h s).$$

Now, let use choose:

$$s=\frac{1}{\beta_h}\log\frac{(k+1)NB_h}{\varepsilon},$$

and we obtain the lemma. □

As a consequence, using a union bound argument, we have, for any $\varepsilon\geq 0$, with probability at least $1-\varepsilon$, for any $h\in\{1,\dots,m\}$:

$$\mathcal{S}_{id}(\theta_h)=\sup_{i\in\{1,\dots,(k+1)N\}}\theta_h(X_i)Y_i-\inf_{i\in\{1,\dots,(k+1)N\}}\theta_h(X_i)Y_i\leq\frac{2}{\beta_h}\log\frac{2(k+1)mNB_h}{\varepsilon}.$$

By plugging the lemma into Eq. (5.5) we obtain the theorem. □

5.6. Proof of Theorem 2.1*: integration of the transductive results*

Actually, the proof is quite direct now: instead of using the techniques given in the section devoted to the inductive case, we use a result valid in the transductive case and integrate it with respect to the test sample. This idea is quite classical in learning theory, and was actually one of the reason for the introduction of the transductive setting (see [22] for example). There are several ways to perform this integration (see for example [6]), here we choose to apply a result obtained by Panchenko [17] that gives a particularly simple result here.

Lemma 5.7 ([17], Corollary 1). *Let us assume that we have i.i.d. variables $T_1,\dots,T_N$ (with distribution P and values in $\mathbb{R}$) and an independent copy $T'=(T'_1,\dots,T'_N)$ of $T=(T_1,\dots,T_N)$. Let $\xi_j(T,T')$ for $j\in\{1,2,3\}$ be three measurables functions taking values in $\mathbb{R}$, and $\xi_3\geq 0$. Let us assume that we know two constants $A\geq 1$ and $a>0$ such that, for any $u>0$:*

$$P^{\otimes 2N}\big[\xi_1\big(T,T'\big)\geq\xi_2\big(T,T'\big)+\sqrt{\xi_3\big(T,T'\big)u}\big]\leq A\exp(-au).$$

Then, for any $u>0$:

$$P^{\otimes 2N}\big\{P^{\otimes 2N}\big[\xi_1\big(T,T'\big)|T\big]\geq P^{\otimes 2N}\big[\xi_2\big(T,T'\big)|T\big]+\sqrt{P^{\otimes 2N}\big[\xi_3\big(T,T'\big)|T\big]u}\big\}\leq A\exp(1-au).$$

Proof of Theorem 2.1. A simple application of the first inequality of Lemma 5.2 (given as a tool for the proof of the transductive results) with $\varepsilon>0$, any $k\in\{1,\dots,m\}$, $g=id$, $\eta=1+\log\frac{2m}{\varepsilon}$ and:

$$\lambda_k=\sqrt{\frac{N\eta}{(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2}}$$

leads us to the following bound, for any k:

$$P^{\otimes 2N}\exp\left[\sqrt{N\eta}\frac{(1/N)\sum_{i=1}^{N}[\theta_k(X_i)Y_i-\theta_k(X_{i+N})Y_{i+N}]}{\sqrt{(1/N)\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2}}-2\eta\right]\leq\exp(-\eta),$$

or:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\theta_k(X_i)Y_i-\theta_k(X_{i+N})Y_{i+N}\right]\geq\sqrt{\frac{4\eta}{N^2}\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2}\right]\leq\exp(-\eta)=\frac{\varepsilon}{2k\exp(1)}.$$

We now apply Panchenko's lemma with:

$$T_i=\theta_k(X_i)Y_i,\qquad T_i'=\theta_k(X_{i+N})Y_{i+N},$$
$$\xi_1(T,T')=\frac{1}{N}\sum_{i=1}^{N}T_i,\qquad \xi_2(T,T')=\frac{1}{N}\sum_{i=1}^{N}T_i',$$
$$\xi_3(T,T')=\frac{2}{N^2}\sum_{i=1}^{2N}\theta_k(X_i)^2Y_i^2\geq 0,$$

and $A=a=1$. We obtain:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\theta_k(X_i)Y_i-P\left[\theta_k(X)Y\right]\right]\geq\sqrt{\frac{4\eta}{N^2}\sum_{i=1}^{N}\left[\theta_k(X_i)^2Y_i^2+P\left[\theta_k(X)^2Y^2\right]\right]}\right]\leq\exp(1-\eta)=\frac{\varepsilon}{2k}.$$

Remark finally that:

$$P\left[\theta_k(X)^2Y^2\right]\leq P\left[\theta_k(X)^2\right]\left(B^2+\sigma^2\right).$$

We proceed exactly in the same way with the reverse inequalities for any k and combine the obtained $2m$ inequalities to obtain the result:

$$\begin{aligned}
&P^{\otimes N}\left\{\exists k\in\{1,\dots,m\},\frac{1}{N}\sum_{i=1}^{N}\left|\theta_k(X_i)Y_i-P\left[\theta_k(X)Y\right]\right|\right.\\
&\left.\geq\sqrt{\frac{4+4\log(2m/\varepsilon)}{N^2}\sum_{i=1}^{N}\left\{\theta_k(X_i)^2Y_i^2+P\left[\theta_k(X)^2\right]\left(B^2+\sigma^2\right)\right\}}\right\}\\
&=P^{\otimes 2N}\left\{\exists k\in\{1,\dots,m\},\frac{1}{N}\sum_{i=1}^{N}\left|\theta_k(X_i)Y_i-P\left[\theta_k(X)Y\right]\right|\right.\\
&\left.\geq\sqrt{\frac{4+4\log(2m/\varepsilon)}{N^2}\sum_{i=1}^{N}\left\{\theta_k(X_i)^2Y_i^2+P\left[\theta_k(X)^2\right]\left(B^2+\sigma^2\right)\right\}}\right\}\leq\varepsilon
\end{aligned}$$

that ends the proof. □

5.7. Proof of Theorems 4.1 *and* 4.2*: Theorem* 2.3 *used as an oracle inequality*

Proof of Theorem 4.1. Let us begin the proof with a general m and ε, the reason of the choice $m=N$ and $\varepsilon=N^{-2}$ will become clear. Let us also call $\mathcal{E}(\varepsilon)$ the event satisfied with probability at least $1-\varepsilon$ in Theorem 2.1. We have:

$$P^{\otimes N}\left[\left\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\right\|_P^2\right]=P^{\otimes N}\left[1_{\mathcal{E}(\varepsilon)}\left\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\right\|_P^2\right]+P^{\otimes N}\left[(1-1_{\mathcal{E}(\varepsilon)})\left\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\right\|_P^2\right].$$

First of all, it is obvious that:

$$P^{\otimes N}\big[(1-1_{\mathcal{E}(\varepsilon)})\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big]\le 2P^{\otimes N}\big[(1-1_{\mathcal{E}(\varepsilon)})\big(\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}\big\|_P^2+\|f\|_P^2\big)\big]$$
$$\le 2\varepsilon\big(B^2m+B^2\big)=2\varepsilon(m+1)B^2.$$

For the other term, just remark that, for any $m'\le m$:

$$\big\|\Pi_P^{\mathcal{F},N}\hat{\theta}-f\big\|_P^2=\big\|\Pi_P^{\mathcal{F},m}\Pi_P^{m,\varepsilon}\cdots\Pi_P^{1,\varepsilon}0-f\big\|_P^2\le\big\|\Pi_P^{m,\varepsilon}\cdots\Pi_P^{1,\varepsilon}0-f\big\|_P^2\le\big\|\Pi_P^{m',\varepsilon}\cdots\Pi_P^{1,\varepsilon}0-f\big\|_P^2$$
$$\le\sum_{k=1}^{m'}\frac{4[1+\log(2m/\varepsilon)]}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\theta_k(X_i)^2Y_i^2+B^2+\sigma^2\right]+\|\overline{\theta}_{m'}-f\|_P^2.$$

This is where Theorem 2.3 has been used as an oracle inequality: the estimator that we have, with $m\ge m'$, is better than the one with the "good choice" m'. We also have:

$$P^{\otimes N}\big[1_{\mathcal{E}(\varepsilon)}\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big]\le P^{\otimes N}\left[\sum_{k=1}^{m'}\frac{4[1+\log(2m/\varepsilon)]}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\theta_k(X_i)^2Y_i^2+B^2+\sigma^2\right]\right]+\big(m'\big)^{-2\beta}C$$
$$\le m'\frac{8[1+\log(2m/\varepsilon)]}{N}\big[B^2+\sigma^2\big].$$

So finally, we obtain, for any $m'\le m$:

$$P^{\otimes N}\big[\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big]\le m'\frac{8[1+\log(2m/\varepsilon)]}{N}\big[B^2+\sigma^2\big]+\big(m'\big)^{-2\beta}C+2\varepsilon(m+1)B^2.$$

The choice of:

$$m'=\left(\frac{N}{\log N}\right)^{1/(2\beta+1)}$$

leads to a first term of order $N^{-2\beta/(2\beta+1)}\log\frac{m}{\varepsilon}(\log N)^{2\beta/(2\beta+1)}$ and a second term of order $N^{-2\beta/(2\beta+1)}\times(\log N)^{2\beta/(2\beta+1)}$. The choice of $m=N$ and $\varepsilon=N^{-2}$ gives a first and a second term of the desired order $N^{-2\beta/(2\beta+1)}(\log N)^{2\beta/(2\beta+1)}$ while keeping the third term at order N^{-1}. This proves the theorem. □

Proof of Theorem 4.2. Here again let us write $\mathcal{E}(\varepsilon)$ the event satisfied with probability at least $1-\varepsilon$ in Theorem 2.1. We have:

$$P^{\otimes N}\big[\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big]=P^{\otimes N}\big[1_{\mathcal{E}(\varepsilon)}\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big]+P^{\otimes N}\big[(1-1_{\mathcal{E}(\varepsilon)})\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\big].$$

For the first term we still have:

$$\big\|\Pi_P^{\mathcal{F},m}\hat{\theta}-f\big\|_P^2\le 2(m+1)B^2.$$

For the second term, let us write the expansion of f into our wavelet basis:

$$f=\alpha\phi+\sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k},$$

and

$$\hat{\theta}(x)=\tilde{\alpha}\phi+\sum_{j=0}^{J}\sum_{k=1}^{2^j}\tilde{\beta}_{j,k}\psi_{j,k}$$

the estimator $\hat{\theta}$. Let us put $J = 2^{\lfloor (\log N)/\log 2 \rfloor}$.

$$\begin{aligned}
\left\| \Pi_P^{\mathcal{F},m} \hat{\theta} - f \right\|_P^2 \leq \|\hat{\theta} - f\|_P^2 &= \left\| \Pi_P^{m,\varepsilon} \cdots \Pi_P^{1,\varepsilon} 0 - f \right\|_P^2 \\
&= (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 + \sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2 \\
&\leq (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 1\left(|\beta_{j,k}| \geq \kappa\right) + \sum_{j=0}^{J} \sum_{k=1}^{2^j} \beta_{j,k}^2 1\left(|\beta_{j,k}| < \kappa\right) + \sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2
\end{aligned}$$

for any $\kappa \geq 0$, as soon as $\mathcal{E}(\varepsilon)$ is satisfied (here again we used Theorem 2.3 as an oracle inequality). Now, we follow the technique used in [11] and [12] (see also the end of the third chapter in [7]). As soon as $\mathcal{E}(\varepsilon)$ is satisfied we have:

$$\begin{aligned}
\sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 1\left(|\beta_{j,k}| \geq \kappa\right) &\leq \frac{8(B^2 + \sigma^2) \log(2m/\varepsilon)}{N} \sum_{j=0}^{J} \sum_{k=1}^{2^j} 1\left(|\beta_{j,k}| \geq \kappa\right) \\
&\leq \frac{8(B^2 + \sigma^2) \log(2m/\varepsilon)}{N} \sum_{j=0}^{J} \sum_{k=1}^{2^j} \left(\frac{|\beta_{j,k}|}{\kappa} \right)^{2/(2s+1)} \\
&= \frac{8(B^2 + \sigma^2) \log(2m/\varepsilon)}{N} \kappa^{-2/(2s+1)} \sum_{j=0}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^{2/(2s+1)}.
\end{aligned}$$

In the same way, we have:

$$\sum_{j=0}^{J} \sum_{k=1}^{2^j} \beta_{j,k}^2 1\left(|\beta_{j,k}| < \kappa\right) \leq \kappa^{2-2/(1+2s)} \sum_{j=0}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^{2/(1+2s)}.$$

So we have to give an upper bound on the quantity:

$$\sum_{j=0}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^{2/(2s+1)}.$$

By Hölder's inequality we have, as soon as $p \geq \frac{2}{2s+1}$:

$$\sum_{j=0}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^{2/(2s+1)} \leq \sum_{j=0}^{J} \left[2^{j(1+1/2-1/p)} \sum_{k=1}^{2^j} |\beta_{j,k}|^p \right]^{2/(1+2s)} \leq \|f\|_{s,p,q}^{2/(1+2s)} J^{(1-2/((1+2s)q))_+},$$

let us put $C' = \|f\|_{s,p,q}^{2/(1+2s)}$. Finally, note that we have, for $p \geq 2$:

$$\sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2 \leq \sum_{j=J+1}^{\infty} \left(\sum_{k=1}^{2^j} \beta_{j,k}^p \right)^{2/p} 2^{j(1-2/p)}.$$

As $f \in B_{s,p,q} \subset B_{s,p,\infty}$ we have:

$$\left(\sum_{k=1}^{2^j} \beta_{j,k}^p \right)^{2/p} \leq C' 2^{-2j(s+1/2-1/p)}$$

for some C'' and so:

$$\sum_{j=J+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2 \le C'''2^{-2Js}$$

for some C'''. In the case where $p<2$ we use (see [12], for $s>\frac{1}{p}-\frac{1}{2}$):

$$B_{s,p,q}\subset B_{s-1/p+1/2,2,q}$$

to obtain:

$$\sum_{j=J+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2 \le C''''2^{-2J(s+1/2-1/p)} \le C''''2^{-J}.$$

So we have:

$$P^{\otimes N}d^2(\tilde f, f) \le 2(m+1)\varepsilon\big(B^2+\sigma^2\big)+\frac{8(B^2+\sigma^2)\log(2m/\varepsilon)}{N}\big(1+C'\kappa^{-2/(1+2s)}J^{(1-2/((1+2s)q))_+}\big)$$
$$+C'\kappa^{2-2/(1+2s)}J^{(1-2/((1+2s)q))_+}+C'''\big(2^{-J}\big)^{2s}+C''''2^{-J}.$$

Let us remember that:

$$\frac{N}{2}\le m=2^J\le N$$

and that $\varepsilon=N^{-2}$, and take:

$$\kappa=\sqrt{\frac{\log N}{N}}$$

to obtain the desired rate of convergence. □

Acknowledgments

I would like to thank my PhD advisor, Professor Olivier Catoni, for his constant help, and the anonymous referee for very useful comments and remarks.

References

[1] P. Alquier. Transductive and inductive adaptative inference for regression and density estimation. PhD thesis, University Paris 6, 2006.
[2] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré. Probab. Statist.* **40** (2004) 685–736. MR2096215
[3] A. Barron, A. Cohen, W. Dahmen and R. DeVore. Adaptative approximation and learning by greedy algorithms. Preprint, 2006.
[4] G. Blanchard, P. Massart, R. Vert and L. Zwald. Kernel projection machine: a new tool for pattern recognition. In *Advances in Neural Inf. Proc. Systems* (*NIPS*, *2004*) 1649–1656, Mit Press, 2005.
[5] B. E. Boser, I. M. Guyon and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM, 1992.
[6] O. Catoni. A pac-Bayesian approach to adaptative classification. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2003.
[7] O. Catoni. Statistical learning theory and stochastic optimization. *Saint-Flour Summer School on Probability Theory*. *Lecture Notes in Math*. **1851**. Springer, Berlin, 2004. MR2163920
[8] O. Catoni. Improved Vapnik–Cervonenkis bounds. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2005.
[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge University Press, 2000.
[10] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelets. *Biometrika* **81** (1994) 425–455. MR1311089

[11] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.* **24** (1996) 508–539. MR1394974

[12] W. Härdle, G. Kerkyacharian, D. Picard and A. B. Tsybakov. *Wavelets, Approximations and Statistical Applications* **129**. Springer, New York, 1998. MR1618204

[13] A. Juditsky, A. Nazin, A. Tsybakov and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Probl. Inf. Transm.* **41** (2005) 368–384. MR2198228

[14] A. Juditsky, P. Rigollet and A. Tsybakov. Mirror averaging, aggregation and model selection. In *Meeting on Statistical and Probabilistic Methods of Model Selection*, pp. 2688–2691. Oberwolfach reports, 2005.

[15] G. Kerkyacharian and D. Picard. Regression in random design and warped wavelets. *Bernoulli* **10** (2004) 1053–1105. MR2108043

[16] A. Nemirovski. Topics in non-parametric statistics. *Saint-Flour Summer School on Probability Theory* 85–277. *Lecture Notes in Math.* **1738**. Springer, Berlin, 2000. MR1775640

[17] D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Ann. Probab.* **31** (2003) 2068–2081. MR2016612

[18] R. Schapire, Y. Freund, P. Bartlett and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** (1998) 1651–1686. MR1673273

[19] B. Schölkopf, A. J. Smola and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10** (1998) 1299–1319.

[20] M. Seeger. Pac-Bayesian generalization error bounds for Gaussian process classification. *J. Mach. Learn. Res.* **3** (2002) 233–269. MR1971338

[21] A. Tsybakov. Optimal aggregation of classifiers instatistical learning. *Ann. Statist.* **32** (2004) 135–156. MR2051002

[22] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1998. MR1367965

[23] B. Widrow and M. Hoff. Adaptative switching circuits. In *IRE WESCON Convention Record*, *Part 4*, *Computers: Man–Machine Systems*, 96–104, 2005.

[24] Y. Yang. Aggregating regression procedures to improve performances. *Bernoulli* **10** (2004) 25–47. MR2044592