Transductive and Inductive Adaptative Inference for Regression and Density Estimation

Pierre Alquier

Université Paris 6
UFR de Mathématiques

Thèse en vue de l'Obtention du Diplôme de
Docteur de l'Université Paris 6
Spécialité: Mathématiques

Transductive and Inductive Adaptative Inference for Regression and Density Estimation

Pierre Alquier

Sous la Direction de **Olivier Catoni**

Thèse soutenue le 08/12/2006
devant le Jury composé de:

Yannick Baraud, Université de Nice Sophia-Antipolis
Olivier Catoni, CNRS et Université Paris 6, Directeur de Thèse
Yuri Goloubev, CNRS et Université Aix-Marseille 1, Rapporteur
Marc Hoffmann, Université de Marne-la-Vallée
Irina Kourkova, Université Paris 6
Pascal Massart, Université Paris-Sud
Dominique Picard, Université Paris 7, Présidente

au vu des Rapports de:

Yuri Goloubev, CNRS et Université Aix-Marseille 1
Gábor Lugosi, ICREA et Pompeu Fabra University

TRANSDUCTIVE AND INDUCTIVE ADAPTATIVE INFERENCE

PIERRE ALQUIER

Abstract. The aim of this thesis is the study of statistical properties of learning algorithm in the case of regression and density estimation. It is divided into three parts.

In the first part, the idea is to generalize Olivier Catoni's PAC-Bayesian theorems ([10]) about classification to the case of regression estimation with a general loss function.

In the second part, we focus more particularly on the least square regression and propose a new iterative algorithm for feature selection. This method can be applied to the case of orthonormal function basis, leading to optimal rates of convergences, as well as to kernel type functions, leading to some variants of the well-known SVM method.

In the third part, we generalize the method proposed in the second part to the density estimation setting with quadratic loss.

Key-words and phrases: statistical learning theory, model selection, least square regression estimation, confidence regions, concentration inequalities, pac-bayesian bounds, non-parametric estimation, adaptative estimation, empirical complexity measure, compression schemes, support vector machines, oracle inequalities, randomized estimator, Gibbs distribution, density estimation, wavelets, bound on the risk.

Résumé en Français: Cette thèse a pour objet l'étude des propriétés statistiques d'algorithmes d'apprentissage dans le cas de l'estimation de la régression et de la densité. Elle est divisée en trois parties.

La première partie consiste en une généralisation des théorèmes PAC-Bayésiens, sur la classification, d'Olivier Catoni ([10]), au cas de la régression avec une fonction de perte générale.

Dans la seconde partie, on étudie plus particulièrement le cas de la régression aux moindres carrés et on propose un nouvel algorithme de sélection de variables. Cette méthode peut être appliquée notamment au cas d'une base de fonctions orthonormales, et conduit alors à des vitesses de convergence optimales, mais aussi au cas de fonctions de type noyau, elle conduit alors à une variante des méthodes dites "machines à vecteurs supports" (SVM).

La troisième partie étend les résultats de la seconde au cas de l'estimation de densité avec perte quadratique.

Remerciements

Je tiens à exprimer mes remerciements les plus sincères à mon directeur de thèse Olivier Catoni, pour m'avoir proposé un sujet de recherche qui m'a passionné, pour sa grande gentillesse et disponibilité, et pour les Mathématiques que j'ai pu apprendre auprès de lui au cours des trois années passées.

Je remercie Gábor Lugosi et Yuri Goloubev pour m'avoir fait l'honneur d'accepter de rapporter ma thèse, ainsi que Yannick Baraud, Marc Hoffmann, Irina Kourkova, Pascal Massart et Dominique Picard celui d'accepter de faire partie de mon jury.

Le Laboratoire de Statistiques du CREST a financé ma thèse et je tiens à en remercier tous les membres pour leur accueil pendant les trois dernières années: Patrice Bertail, Paul Doukhan, Emmanuelle Gautherat, Ghislaine Gayraud, Christian Robert, Judith Rousseau, Jean-Michel Zakoian, Mathieu Cornec, Sophie Dabo-Niang, Stéphanie Dupoiron, Romuald Elie, Eric Gautier, Hugo Harari-Kermadec, Cyrille Joutard, Frédéric Lavancier, Mathieu Rosenbaum, Jessica Tressou...

Je tiens à remercier également tous les membres du Laboratoire de Probabilités et Modèles Aléatoires de Paris 6 au sein duquel j'ai effectué ma thèse, et en particulier ceux dont j'ai pu suivre les cours lors de mon dea: Alexandre Tsybakov, Gérard Kerkyacharian, Lucien Birgé, Stéphane Boucheron, Nicolas Vayatis, Jean Jacod, Zhan Shi... Christophe Chesneau qui organisait le groupe de travail des thésards, avec Stéphane Gaiffas, Thomas Willer, Olivier Wintenberger, Tran Viet Chi, Philippe Rigollet, Guillaume Lécué, Frédéric Guilloux...

J'ai tenté de suivre avec assiduité les séances du groupe de travail "apprentissage" donc j'ai beaucoup appris, je tiens à en remercier les organisateurs Gilles Stoltz et Patricia Reynaud-Bouret, ainsi que tous les participants, en particulier Jean-Yves Audibert.

Je souhaite également remercier un certain nombre de professeurs dont j'ai pu, au cours de mes études, suivre les enseignements qui ont été déterminants dans mon choix de faire des Mathématiques, en particulier Jacques Sauloy, Monique Ramis et l'albigeois légendaire Jean-Paul Lavaux.

Finalement, je remercie mes amis et ma famille, sans pouvoir citer tout le monde, un grand merci encore à Gilles, Josiane, Aline, Marie, Vincent & Marion (merci pour le sauvetage informatique de dernière minute Vincent!), et Jeoffrey.

Contents

Introduction

Most statistical problems can be formulated in the following way: given a set of observations $(Z_1, ..., Z_N) \in \mathcal{Z}^N$ drawn from an unknown probability distribution P, estimate this distribution (or, some function linked with this distribution, like in the regression estimation problem). Actually, it is a well-known fact that the problem cannot be solved without additional assumptions. There is no estimation method that allows to obtain convergence to P with a given rate.

In order to solve this problem, two points of view emerged in statistical inference. The first one consists in making particular assumptions about P, namely that P belongs to $\mathcal{P}$ where $\mathcal{P}$ is a sufficiently small subset of the set of all probability distributions on $\mathcal{Z}$, $\mathcal{P} \subsetneq \mathcal{M}_+^1(\mathcal{Z})$; when $\mathcal{Z}$ is the set of real numbers, $\mathcal{Z} = \mathbb{R}$, $\mathcal{P}$ can be the set of all Gaussian distributions with mean and standard deviation $(m, \sigma) \in \mathbb{R} \times \mathbb{R}_+$, or the set of all probability distributions P that are absolutely continuous with respect to a given measure μ and such that the density function, $\frac{dP}{d\mu}$, has some given regularity. The second point of view is to make no such assumptions on P but to restrict our estimator to belong to a given set $\mathcal{Q}$. In this case, the objective of our estimator is no longer to converge to P at a given rate but to the probability in $\mathcal{Q}$ that is the best approximation of P, say $\arg\min_{Q \in \mathcal{Q}} d(P, Q)$ for some discrepancy measure d. This is the point of view of statistical learning theory, initiated by Vapnik and Cervonenkis in [13].

Statistical learning theory knew a very important development in the last twenty years (see the books of Vapnik [41, 42], Devroye, Györfi and Lugosi [17], Hastie, Tibshirani and Friedman [22] for example, or more recently the paper by Boucheron, Bousquet and Lugosi [8]), and the Structural Risk Minimization strategy proposed by Vapnik and Cervonenkis was successfully implemented through algorithms like Support Vector Machines (SVM, that were introduced by Boser, Guyon and Vapnik [7]). Support Vector Machines are able to perform classification (or regression) of very high-dimensional data, having thus successfull applications in image processing (pattern recognition), speech recognition, bioinformatics...

0.1. **General overview.** In this thesis, we focus on regression estimation algorithm motivated by the learning theory point of view, in both the inductive and the transductive setting (the terminology is due to Vapnik). Let us first shortly define what we mean by transuctive and inductive regression estimation.

In the inductive setting, that is actually the most standard setting, we observe a sample $(X_i, Y_i)_{i=1...N} = (Z_i)_{i=1...N}$ drawn from an unknown probability distribution $P_N = P^{\otimes N}$ on the space:

$$\left((\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})^{\otimes N}\right) = \left(\mathcal{Z}^N, \mathcal{B}_\mathcal{Z}^{\otimes N}\right),$$

where P is an (unknown) probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})$, and $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is some measurable space, as $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$. The aim is to build a function f able to predict a value Y_{N+1} by $f(X_{N+1})$ where (X_{N+1}, Y_{N+1}) is drawn from P, with a small error in expectation.

In the transductive case, we assume that $k \in \mathbb{N}^*$ and that $P_{(k+1)N} = P^{\otimes (k+1)N}$ is some probability measure on the space:

$$\left((\mathcal{X} \times \mathcal{Y})^{(k+1)N}, (\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})^{\otimes (k+1)N}\right) = \left(\mathcal{Z}^{(k+1)N}, \mathcal{B}_\mathcal{Z}^{\otimes (k+1)N}\right)$$

We observe $(X_i, Y_i)_{i=1...N}$ (the learning sample) and $(X_i)_{i=N+1...(k+1)N}$ (the test sample), and we only focus on the estimation of the values $(Y_i)_{i=N+1...(k+1)N}$. The notion of transductive inference was actually introduced by Vapnik, mainly as a tool to study the inductive case. The idea is to introduce the test sample (called in this context shadow sample) in the inductive case, leading to a context where

Hoeffding or Bernstein type inequalities can be applied, and then to get rid of the shadow sample by an integration with respect it (for Hoeffding's inequality, see [23], for Bernstein's inequality, a complete presentation can be found in [30]).

However, we insist on the fact that the transductive inference setting does not seem unrealistic and have an interest on its own. We propose some examples. In a sample survey, one wants to get informations (say Y_i) about a whole population (represented by $X_i, i \in \{1, ..., (k+1)N\}$) but the size of the population prevents the statistician to observe more than a certain fraction $(1/(k+1))$ of the population, $i \in \{1, ..., N\}$. In this case, transductive inference seems adapted to the problem. Another example is the following. One is given a large set of pictures (represented by $X_i, i \in \{1, ..., (k+1)N\}$). One wants to label every picture according to the fact that they represent $(Y_i = 1)$ or not $(Y_i = 0)$ a given object of interest. If k is too large for the whole set to be labeled by a single person, this person can label the N first images at hand, and then use transductive inference to label the other pictures with a small risk of error.

Note that we gave the definitions of both cases in the i. i. d. setting for the sake of simplicity, however in the core of the thesis we will deal with more general cases.

As mentioned previously, in the statistical learning theory approach, we assume that we are given a set of regression functions:

$$\mathcal{R} = \{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\},$$

where Θ is a set of parameters. This does not mean that we are considering a parametric model, we do not assume that $\Theta \subset \mathbb{R}^d$ for any $d \in \mathbb{N}$, we will be typically interested in situations where Θ is a disjoint union of subsets of various dimensions, such as in the case where:

$$\Theta = \bigsqcup_{m \in \mathcal{M}} \Theta_m$$

for a finite or countable index set $\mathcal{M}$, and where $\Theta_m \subset \mathbb{R}^{d_m}$, the dimension d_m depending on the model.

We put, for any measurable nonnegative function $\psi : \mathcal{Y}^2 \to \mathbb{R}$ for any $\theta \in \Theta$, $\psi_i(\theta) = \psi\left(f_\theta(X_i), Y_i\right)$ and:

$$r_1(\psi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \psi_i(\theta),$$

$$r_2(\psi, \theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \psi_i(\theta),$$

$$R(\psi, \theta) = P\left\{\psi\left[f_\theta(X), Y\right]\right\}.$$

So in the inductive setting, our objective is:

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

while in the transductive case it is:

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

However, these two quantities are not observable, and so our objective is to build an estimator (a function of the observations) $\hat{\theta}$ such that:

$$R(\psi, \hat{\theta}) - R(\psi, \overline{\theta})$$

or:

$$r_2(\psi, \hat{\theta}) - r_2(\psi, \theta_2)$$

can be upper bounded with high probability or in expectation. The choice of the empirical risk minimizer $\arg\min_\theta r_1(\psi,\theta)$ does not lead to a good estimator in general (for example in the typical situation we described previously, where some dimensions d_m are large).

This thesis is divided into three parts. In the first one, we propose a method valid in a very general setting, with no particular assumptions on ψ and $\theta \mapsto f_\theta$. The idea is to extend the PAC-Bayesian approach presented by Catoni in [10] for classification to the regression problem. This approach gives new estimation algorithms as well as theoretical guarantees on the performances of these algorithms. We examine several examples, and among them the case where $\mathcal{Y}$ is the set of real numbers $\mathbb{R}$, ψ is the quadratic loss: $\psi(y,y') = (y-y')^2$, and $\theta \mapsto f_\theta$ is linear, namely the least square regression, and we will focus more particularly on the problems due to the large dimension of the parameter space. In a second part, we focus more particularly on the least square regression problem. Some particular properties of $R(\psi,.)$ (or $r_2(\psi,.)$) allow us to propose a new estimation method that selects a few dimensions in Θ that are relevant for regression estimation. The last part is devoted to the extension of this method to the problem of density estimation with quadratic loss. For the sake of simplicity, the main results are presented, in this introduction, in the inductive setting, but in the core of the thesis a particular attention is also given to the transductive setting (that requires generally less hypothesis than the inductive setting).

0.2. **PAC-Bayesian regression estimation.** Note that in learning theory, one needs a structure over the parameter space Θ. A classical structure is a family of disjoints submodels $(\Theta_m, m \in \mathcal{M})$ with:

$$\Theta = \bigsqcup_{m\in\mathcal{M}} \Theta_m$$

such that the capacity, or complexity, of every submodel can controlled. A classical example of complexity control is the VC-dimension, introduced by Vapnik and Cervonenkis [13].

In the PAC-Bayesian approach, the role of structure over Θ is played by a probability distribution on Θ. We consider a σ-algebra $\mathcal{T}$ on Θ, and a distribution $\pi \in \mathcal{M}_+^1(\Theta)$.

The PAC-Bayesian approach is a point of view in statistical learning theory that was initiated by McAllester [28], and its name is due to the fact that in its first form its objective was to combine the major advantages of the learning theory point of view and of the Bayesian statistics[1]. However, note that this approach is not Bayesian, and in particular π does not reflect any prior belief on the localization of the "true" value of the parameter nor a stochastic modelization of $\theta \in \Theta$, π just plays the role of defining a structure over Θ; π is also called the prior distribution in the PAC-Bayesian point of view but does not have any Bayesian interpretation.

The technique used in this thesis are closer to the one developed more recently by Catoni [9, 10, 11]. The idea is to control:

$$\int_\Theta R(\psi,\theta)\rho(d\theta) = \rho\left[R(\psi,.)\right]$$

for any $\rho \in \mathcal{M}_+^1(\Theta)$, with high probability or in expectation (with respect to the sample $(X_1,Y_1),...,(X_N,Y_N)$).

In learning theory, the control the risk of an estimator $\hat\theta$ in a submodel $\Theta_j \subset \Theta$, here $R(\psi,\hat\theta)$, is based on the empirical risk $R(\psi,\hat\theta)$ and on the complexity of Θ_j. So

[1]PAC means "Probably Approximately Correct", the expression was introduced by Valiant [40] in reference to the deviation bounds used in learning theory, that are true with high probability.

the PAC-Bayesian point of view consists in controlling $\rho[R(\psi,.)]$ by its empirical counterpart $\rho[r(\psi,.)]$, and the complexity term is replaced by a measure of the distance between ρ and the π. Here, this is done with the Kullback divergence:

$$\mathcal{K}(\rho,\pi) = \rho\left[\log\left(\frac{d\rho}{d\pi}\right)\right]$$

if $\rho \ll \pi$ and $\mathcal{K}(\rho,\pi) = +\infty$ otherwise.

A natural question is the practical interest of the control of $\rho[R(\psi,.)]$. Several interpretations are discussed in the thesis. First of all, $\rho[R(\psi,.)]$ is exactly the mean risk of the procedure that consists, for any given X_{N+1}, in drawning a parameter θ from the distribution ρ and then predicting Y_{N+1} by $f_\theta(X_{N+1})$.

Let us see a first example of the kind of results presented in the first part of the thesis.

Theorem 0.1 (Theorem 2.3 in the thesis). *Let us assume that ψ takes values in $[0,C]$ for a constant $C > 1$. For any $\varepsilon > 0$, for any $\lambda \in (0, N/C)$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\rho[R(\psi,.)] \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \psi, .\right]\right\} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\},$$

where for any $\alpha \in \mathbb{R}_+^*$:

$$\Phi_\alpha : \left(-\infty, \frac{1}{\alpha}\right) \to \mathbb{R}$$
$$x \mapsto \frac{-\log(1-\alpha x)}{\alpha}.$$

Note that a precise study of $\Phi_{\lambda/N}$ shows that, for any $p \in [0,C]$ and $\lambda \leq (N/C)$ we have:

$$p \leq \Phi_{\frac{\lambda}{N}}(p) \leq p + \frac{\lambda}{2N}p^2,$$

and so one can get the simpler bound, for any $\lambda < N/(2C)$:

$$\Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \psi, .\right]\right\} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$\leq \rho[r(\psi,.)] + \frac{\lambda C^2}{2N} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}$$

(of course, in practical situations one should keep the left-hand side that is more accurate as an upper bound for $\rho[R(\psi,.)]$, this inequality is given here only for the sake of comprehension).

The choice of the parameter λ is discussed in details in the thesis (note that the minimization of the right-hand side is not possible, because the optimal value for λ would depend on ρ, and ρ is allowed to be data-dependent as λ is not). A practical idea, once λ is chosen, is the following. First, choose:

$$\hat{\rho} = \arg\min_{\rho\in\mathcal{M}_+^1(\Theta)} \Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \psi, .\right]\right\} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$= \arg\min_{\rho\in\mathcal{M}_+^1(\Theta)} \left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \psi, .\right]\right\} + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\}$$

and then, for any given X_{N+1}, draw a parameter θ from the distribution $\hat{\rho}$ and predict Y_{N+1} by $f_\theta(X_{N+1})$. We will see in the thesis that $\hat{\rho}$ can be given in explicit form.

Also note that if one wants for some reason to control the risk of a given estimator $\hat{\theta}$, it is sensible to take $\hat{\rho}$ restricted to a small neighborhood of $\hat{\theta}$, because we expect $\rho\left[R\left(\psi,.\right)\right] \simeq R\left(\psi,\hat{\theta}\right)$. This approach is detailed in the thesis.

Another case of interest for the comprehension of Theorem 2.3 is the case where Θ is finite. Let us take, in this case:

$$\pi = \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\delta_\theta$$

and restrict ρ to the set: $\{\delta_\theta, \theta\in\Theta\}$. The theorem becomes, for any $\varepsilon > 0$, for any $\lambda \in (0, N/(2C))$, with P_N-probability at least $1-\varepsilon$, for any $\theta\in\Theta$:

$$\begin{aligned} R\left(\psi,\theta\right) &\leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{ r\left[\Phi_{\frac{\lambda}{N}}\circ\psi,\theta\right] + \frac{\log|\Theta| + \log\frac{1}{\varepsilon}}{\lambda}\right\} \\ &\leq r\left(\psi,\theta\right) + \frac{\lambda C^2}{2N} + \frac{\log|\Theta| + \log\frac{1}{\varepsilon}}{\lambda}. \end{aligned}$$

Note that here, the choice of λ is easy as the right-hand side is minimal for the value:

$$\lambda_0 = \sqrt{\frac{2N\log\frac{|\Theta|}{\varepsilon}}{C^2}}$$

that is not data-dependent. So we take:

$$\lambda = \lambda_0 \wedge \frac{N}{2C},$$

and we obtain, at least N is large enough:

$$R\left(\psi,\theta\right) \leq r\left(\psi,\theta\right) + C\sqrt{\frac{2\log\frac{|\Theta|}{\varepsilon}}{N}}$$

First, this is an incitation to choose the estimator:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} r\left[\psi,\theta\right]$$

the empirical risk minimizer. Moreover, note that we obtain the complexity term $\log|\Theta|$. It is a known fact in learning theory (see Vapnik [41]) that the complexity of a finite set Θ should be measured by $\log|\Theta|$.

Note that, however, in the general case, it is known in learning theory that if one wants to have estimators that achieve optimal rates of convergence in expectation, theorems like 2.3 are not sufficient. It is generally necessary to study "relative bounds", namely bounds on:

$$R(\psi,\hat{\theta}) - R(\psi,\theta_0)$$

for some given value $\theta_0\in\Theta$. A detailed explanation of this phenomenon is given for example in the introduction of the PhD thesis of Audibert [2].

In the PAC-Bayesian setting, we have for example the following result (given here, for the sake of simplicity, in the case where ψ is bounded, however more general versions are given with their proof in the core of the thesis).

Theorem 0.2 (Theorem 2.11 in the thesis). *Let $(\pi,\pi')\in(\mathcal{M}^1_+(\Theta))^2$. Let us assume that ψ takes values in $[0,C]$ for some $C>1$. For any $\varepsilon>0$, for any $\lambda\in(0,N/C)$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $(\rho,\rho')\in(\mathcal{M}^1_+(\Theta))^2$:*

$$(0.1)\quad \Phi_{\frac{\lambda}{N}}\left[\rho R(\psi,.) - \rho' R(\psi,.)\right]$$

$$\leq \rho_{(d\theta)} \otimes \rho'_{(d\theta')} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[\psi_i(\theta) - \psi_i(\theta') \right] \right\} + \frac{\mathcal{K}(\rho, \pi) + \mathcal{K}(\rho', \pi') + \log \frac{1}{\varepsilon}}{\lambda}.$$

Let us choose $\rho' = \pi' = \delta_{\overline{\theta}}$ (and so $\mathcal{K}(\rho', \pi') = 0$), the bound becomes:

$$\rho R(\psi, .) - R\left(\psi, \overline{\theta}\right) \leq \Phi^{-1}_{\frac{\lambda}{N}} \left\{ \rho \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[\psi_i(.) - \psi_i\left(\overline{\theta}\right) \right] \right\} + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} \right\}. \quad (0.2)$$

We prove in the thesis that in the linear case, with quadratic loss $\psi(y, y') = (y - y')^2$, in the case where:

$$\Theta = \mathcal{X} = \left\{ \theta \in \mathbb{R}^d, \|t\| \leq c \right\}$$

for some constant c, under a particular choice for the prior π, we can build an observable posterior $\hat{\rho}$ that is close to minimize the right-hand side of inequality 0.2, such that:

$$P\left[\hat{\rho} R(\psi, .)\right] \leq R\left(\psi, \overline{\theta}\right) + \frac{C' d}{N}$$

where C' is some constant related only to the level of the noise and to c (and d is the dimension of Θ in this particular case).

This approach is satisfying where d is small. However, in a lot of practical applications, the dimension d of Θ is large and can even be greater than N. In this case, the idea is to try to select some "relevant dimensions" in Θ, namely a submodel $\Theta' \subset \Theta$ with dimension $d' < d$ such that $R(\psi, \overline{\theta}')$ is close to $R(\psi, \overline{\theta})$, where:

$$\overline{\theta}' = \arg \min_{\theta \in \Theta'} R(\psi, \theta).$$

The main problem is to define a data-based procedure for the choice of Θ'.

We propose the following approach: let us assume that we have some ($\mathcal{T}$-mesurables) submodels of Θ: $\Theta_i \subset \Theta$ for $i \in \mathcal{I}$. Then, taking ρ and π in $\mathcal{M}^1_+(\Theta_i)$ and ρ' and π' in $\mathcal{M}^1_+(\Theta_j)$ in inequality 0.1 gives a procedure for model selection between Θ_i and Θ_j.

This motivates an iterative algorithm given in the first part of the thesis. Note that the setting used in the thesis is general enough to include model selection in least square regression estimation, selection of support vectors in SVM or selection of a compression set in the context of compression schemes.

0.3. **Iterative feature selection for the least square case.** The second part of the thesis is devoted to relative bounds in the least square regression setting, namely when $\mathcal{Y} = \mathbb{R}$, $\psi(x, x') = (x - x')^2$, Θ is a vector space, and $\theta \mapsto f_\theta$ is linear. Such bounds can be seen as a particular case of the ones obtained in the first part of the thesis (they are obtained by a different approach but we could of course obtain similar ones using the PAC-Bayesian point of view), the main point is that we can give a slightly different interpretation of relative bounds in the linear case, that leads to a new iterative estimation procedure. If one is to compare this procedure to the general model selection method given in the first part, the first thing to note is that the quadratic loss, $\psi(x, x') = (x - x')^2$ is crucial in the second part. From this point of view, the general PAC-Bayesian method of the first part is less restrictive. Moreover, in the inductive setting, the method proposed in the second part requires the knowledge of the distribution of X under P. This hypothesis is made quite often in non-parametric statistics, but is unrealistic in many applications. Note that this major inconvenient does not affect the iterative method of the second part in the transductive setting, where the knowledge of X_i for $i \in \{N+1, ..., (k+1)N\}$ is sufficient. Now, the method given in part 2 has the

following advantage that, at least in its simpler form, it requires only bounds in submodels of Θ of dimension 1. Moreover, the algorithm proposed is very simple to implement, and to interpret, as we will see, as a generalization of thresholding procedures to the case of non necessarily orthogonal features.

Here again, we give the main results in this introduction in the inductive case, although the method proposed here requires clearly more hypothesis than in the transductive setting. So, let us assume that we know the distribution of the design, that means that we know the distribution P_X of X under P.

The idea is to define the following scalar product on Θ:

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

and the associated norm:

$$\|\theta\|_P = \sqrt{\langle \theta, \theta \rangle_P}.$$

Then note that, for any $\theta \in \Theta$ we have:

$$R(\psi, \theta) - R\left(\psi, \overline{\theta}\right) = \|\theta - \overline{\theta}\|_P^2$$

by Pythagore's theorem, and, actually, for any closed subspace $\Theta' \subset \Theta$ with $\overline{\theta}' = \arg\min_{\theta \in \Theta} R(\psi, \theta)$ we have:

$$\forall \theta \in \Theta', \quad R(\psi, \theta) - R\left(\psi, \overline{\theta}'\right) = \|\theta - \overline{\theta}'\|_P^2.$$

Finally let us remark that $\overline{\theta}'$ is the orthogonal projection of $\overline{\theta}$ on Θ':

$$\overline{\theta}' = \arg\min_{\theta \in \Theta'} \|\theta - \overline{\theta}\|_P.$$

We will use the notation $\overline{\theta}' = \Pi_{\Theta'} \overline{\theta}$.

For a particular estimator $\hat{\theta}'$ restricted to Θ' we obtain a bound on the relative risk:

$$R\left(\psi, \hat{\theta}'\right) - R\left(\psi, \overline{\theta}'\right).$$

What follows was motivated by the fact that this bound appears to be numerically very good when $\dim(\Theta') = 1$. We choose a family $(\theta_1, ..., \theta_m) \in \Theta^m$ and define the submodels:

$$\mathcal{M}_i = \{\alpha \theta_i, \alpha \in \mathbb{R}\}$$

and:

$$\overline{\alpha}_i = \arg\min_{\alpha \in \mathbb{R}} R(\psi, \alpha \theta_i).$$

We have for this quantity our estimator that takes here the particular form $\hat{\alpha}_i \theta_i$, where:

$$\hat{\alpha}_i = \frac{\frac{1}{N} \sum_{j=1}^N f_{\theta_i}(X_j) Y_j}{P\left[f_{\theta_i}(X)^2 \right]}.$$

Then our upper bound is given by the following theorem (we give here a weak version with additional hypothesis for the sake of simplicity).

Theorem 0.3 (Theorem 13.1 in the thesis)**.** *Let us assume that the true regression function, that is a measurable version of $x \mapsto P(Y|X = x)$, is such that $\|f\|_\infty \leq C$ for some known constant C. Let us assume that for any $x \in \mathcal{X}$,*

$$P\left\{ \left[Y - f(X) \right]^2 \middle| X = x \right\} \leq \sigma^2$$

for some known constant σ. We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, ..., m\}$:

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

Let us put, for short:

$$\beta(\varepsilon,k)=\frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[\frac{\frac{1}{N}\sum_{i=1}^{N}f_{\theta_k}(X_i)^2Y_i^2+B^2+\sigma^2}{P\left[f_{\theta_k}(X)^2\right]}\right]$$

(note that this bound isn't tight and that a great part of the work in the second part of the thesis is to improve this bound). Actually, these bounds on models of dimension 1 give us some information of the localization of $\overline{\theta}$, as we know that with high-probability:

$$\forall k\in\{1,...,m\},\quad \overline{\theta}\in\mathcal{CR}_{k,\varepsilon}$$

where:

$$\mathcal{CR}_{k,\varepsilon}=\left\{\theta\in\Theta,\quad \|\Pi_{\mathcal{M}_k}\theta-\hat{\alpha}_k\theta_k\|_P\leq\beta(\varepsilon,k)\right\}.$$

It appears that $\mathcal{CR}_{k,\varepsilon}$ is closed, is a convex set, and contains $\overline{\theta}$ with high probability and so we obtain the following corollary.

Corollary 0.4. *Under the same hypothesis, for any $\varepsilon>0$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $\theta\in\Theta$, for any $k\in\{1,...,m\}$:*

$$\|\Pi_{\mathcal{CR}_{k,\varepsilon}}\theta-\overline{\theta}\|_P\leq\|\theta-\overline{\theta}\|_P,$$

or more explicitly:

$$R\left(\psi,\Pi_{\mathcal{CR}_{k,\varepsilon}}\theta\right)-R\left(\psi,\overline{\theta}\right)\leq R\left(\psi,\theta\right)-R\left(\psi,\overline{\theta}\right).$$

We can interpret it in the following way: when we have an estimator θ, for any k, $\Pi_{\mathcal{CR}_{k,\varepsilon}}\theta$ is a better estimator than θ. So we propose the following algorithm (in general form):

- start with $\theta(0)=0$;
- at step n, we have $\theta(n)$, choose a direction $k(n)$ (the choice of $k(n)$ is discussed in the thesis, several examples are proposed) and take:
$$\theta(n+1)=\Pi_{\mathcal{CR}_{k(n),\varepsilon}}\theta(n);$$
- stop when a given criterion is satisfied (here again, several cases are discussed in the core of the thesis).

We study two particular cases, that we introduced previously. The first case is SVM. In this particular case we can show that the search for the exact projection of 0 on the confidence region:

$$\mathcal{CR}_{\varepsilon}=\bigcap_{k=1}^{m}\mathcal{CR}_{k,\varepsilon}$$

leads to a minimization problem very close to the one that is usually taken to compute SVM estimators. However, successive projections on the $\mathcal{CR}_{k,\varepsilon}$ are less computationally intensive. Moreover, the algorithm being given in a very general form, we can note that it gives a theoretical background to almost any reasonable heuristic for the choice of support vectors in SVM.

The second case is the case where $(\theta_1,...,\theta_m)$ is the beginning of $(\theta_i)_{i=1}^{+\infty}$, an orthogonal basis of $\mathcal{L}^2$, with the trivial indexation $f_\theta(x)=\theta(x)$. In both examples, we obtain very good results on simulations with our algorithm. Moreover, in the $\mathcal{L}^2$ case, we remark that the order of projection does not have an influence of the obtained estimator, so we can take:

$$\hat{\theta}=\Pi_{\mathcal{CR}_{m,\varepsilon}}...\Pi_{\mathcal{CR}_{1,\varepsilon}}0.$$

The previous corollary implies the following result that can be interpreted as an oracle inequality.

Corollary 0.5. *Under the same hypothesis, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $m' \in \{1, ..., m\}$:*

$$R\left(\psi, \hat{\theta}\right) = R\left(\psi, \Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0\right) \leq R\left(\psi, \Pi_{\mathcal{CR}_{m',\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0\right).$$

Namely, if m is too large, the algorithm adapts itself to the problem and does not overfit the data. Actually, it is well known that if the true regression function f has a regularity β, then the optimal choice in estimation by projections is:

$$m = N^{\frac{1}{2\beta+1}}$$

leading to estimation at the minimax rate of convergence:

$$\frac{1}{N^{\frac{2\beta}{2\beta+1}}}.$$

Here, if we assume that we do not know β, we can take for example $m = N$ and the previous corollary leads to the following theorem.

Theorem 0.6 (Theorem 13.2 in the thesis)**.** *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 $B \geq 0$ such that:

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

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

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

with C known to the statistician. Then we can build a slight modification of $\hat{\theta}$, $\tilde{\theta} = g(\hat{\theta})$, such that (with $\varepsilon = N^{-2}$ and $m = N$), for any $N \geq 2$,

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

So the estimator achieves the minimax rate of convergence up to a log factor.

Similar results when f is assumed to belong to a Besov space, and when the θ_i are wavelets. Note that in this case, the estimator $f_{\overline{\theta}_m}$ is exactly a soft-thresholded wavelet estimator. So, the iterative feature selection method proposed in part 2 and briefly described previously can be seen as a generalization of thresholding procedures to cases where the features are not orthogonal, for example in the case of SVM.

0.4. **The density estimation case.** The end (part 3) of the thesis is devoted to the generalization of this method to the density estimation case (with quadratic loss). The construction is the same, but we have to adapt the scalar product. If we assume that $Z_1, ..., Z_N$ are generated by a probability distribution P that is assumed to have a density

$$f(.) = \frac{dP}{d\mu}(.)$$

with respect to a measure μ, given a model $\{f_\theta, \theta \in \Theta\}$ where $\theta \mapsto f_\theta$ is linear, we take the following scalar product:

$$\langle \theta, \theta' \rangle = \int_{\mathcal{X}} f_\theta(x) f_{\theta'}(x) \mu(dx)$$

and the associated norm:

$$\|\theta\| = \sqrt{\langle \theta, \theta \rangle}.$$

The method has the same theoretical guarantees as in the regression case, we build an iterative algorithm such that at each step, we build an estimator $\theta(n+1)$ from an estimator $\theta(n)$, such that:

$$\|\theta(n+1) - \overline{\theta}\| \leq \|\theta(n) - \overline{\theta}\|$$

with high probability, where here:

$$\overline{\theta} \in \arg\min_{\theta \in \Theta} \int_{\mathcal{X}} \left[f_\theta(x) - f(x) \right]^2 \mu(dx).$$

In order to obtain this, we have to give an upper bound on the risk of estimators in unidimensional models. Actually, the density estimation case provides a slightly different context that allows us to use a change of variable technique in the deviation inequality (a technique developed by Catoni in recent papers [11]), leading to tighter bounds (at least for small values of N).

Note that it would be possible to use others loss functions to perform density estimation. For example, logarithmic loss function could be used. The algorithm proposed in the third part of the thesis is dedicated to the quadratic loss, however, the PAC-Bayesian techniques given in the first part allow to deal with the density estimation case with more general loss functions.

Part 1. PAC-Bayesian Regression Estimation

The aim of this part is to generalize the PAC-Bayesian theorems proved by Catoni [10, 11] in the classification setting to the regression estimation case. We focus on two cases: the "usual" inductive setting, and the transductive setting where we try to estimate the regression function only on a finite given set of points. First, we give control of the deviations of the risk of randomized estimators. This allows to bound the risk of very general estimation procedures, and gives a criterion to make a choice between different procedures. We then focus on some particular cases: compression schemes, and least square linear regression estimation. One of the consequence of the results obtained is to justify the use of a wide range of algorithms for the selection of the set of support vectors in SVM methods.

1. Introduction: the setting of the problem

In this part we propose methods to perform regression estimation in large dimension. Namely, inputs X_i (represented by a large set of features) and outputs Y_i are given, and one wants to be able to predict Y when given a new input X. The learning theory point of view introduced by Vapnik and Cervonenkis ([13], see Vapnik [41] for a presentation of the main results in English) gives a setting that proved to be adapted to deal with estimation problems in large dimension, allowing to select dimensions of X that may be relevant to predict Y (model selection). This point of view received an important interest over the past few years, see for example Boucheron, Bousquet and Lugosi that present some recent advances [8]. All these works have in common the use of a "structure" on the parameter space.

The PAC-Bayesian point of view on learning theory, introduced by McAllester [28, 29] in the context of classification (the case where $Y \in \{-1, +1\}$), uses as a structure a probability distribution over the parameter space. It can deal with very general problems and gives results about model selection and aggregation. McAllester's bound were improved by Catoni [9, 10, 12, 11], and Audibert [2].

The aim of this part is to extend Catoni's results (in [11]) to the more general context of regression estimation (Y not restricted to $\{-1, +1\}$). In [9] some results were given on the particular case of linear least square regression estimation. In [1] Audibert gives a method is proposed for aggregation of estimators using the PAC-Bayesian point of view, still with the least square loss. Here, we extend the PAC-Bayesian setting a more general context (not necessarily linear, and generic loss). Sections 2 and 3 are devoted to the general results for the inductive and the transductive inference (definitions are given later in the introduction). In sections 4 and 5 we apply these results in two particular cases, an extension of Littlestone and Warmuth's compression schemes [27] and least square linear regression estimation. In both cases we insist on the model selection aspect, in order to obtain a "short" representation of the regression function, in the spirit of Rissanen's minimum description length [45]. Both cases include variants of support vector machines (SVM), allowing to use very general heuristics for the choice of support vectors, as well as for the choice of the kernel.

The end of this introduction is devoted to introduce notations used in the whole part, to emphasize the difference between inductive and transductive inference and to present the particularity of the PAC-Bayesian point of view.

1.1. Transductive and inductive inference.

The transductive inference (as opposed to the inductive inference, which is the "usual" point of view in statistics) was introduced in statistical learning theory by Vapnik [41]. We can think of it in the following way: in the inductive setting, we try to estimate a whole function (in

this thesis, the regression function) while in the transducive case, we try to estimate only the values of this function at some given points of interest.

More precisely, in the inductive setting, we will assume that we observe a sample $(X_i, Y_i)_{i=1\dots N} = (Z_i)_{i=1\dots N}$ drawn from a probability distribution:

$$P_N = \bigotimes_{i=1}^{N} p_i$$

on the space:

$$\left((\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})^{\otimes N}\right) = \left(\mathcal{Z}^N, \mathcal{B}_\mathcal{Z}^{\otimes N}\right),$$

where every p_i is a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})$, and $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is some measurable space, as $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$.

Note that it is usual to consider that for every i, $p_i = P$ where P is a probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})$, and so to consider that we are in the i. i. d. case, where $P_N = P^{\otimes N}$. The aim is to be able to predict a value Y_{N+1} from the knowledge of the value X_{N+1} where (X_{N+1}, Y_{N+1}) is drawn from P, with a small error in expectation. However, in practice it can be the case that some experiments cannot be repeated under the same conditions, so X_1 and X_2 do not have the same distribution. We can think of the deterministic design regression, where the values $(X_i)_{i=1,\dots,N}$ form a deterministic grid of $\mathcal{X}$ and only the Y_i are random. This is why, unless we mention it explicitly, we will not assume that we are in the particular case $P_N = P^{\otimes N}$.

In the transductive case, we assume that $k \in \mathbb{N}^*$ and that $P_{(k+1)N}$ is some partially exchangeable probability measure on the space:

$$\left((\mathcal{X} \times \mathcal{Y})^{(k+1)N}, (\mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})^{\otimes (k+1)N}\right) = \left(\mathcal{Z}^{(k+1)N}, \mathcal{B}_\mathcal{Z}^{\otimes (k+1)N}\right)$$

(the definition of a partially exchangeable probability measure will be given in the section devoted to the transductive setting, just remark that $P_{(k+1)N} = P^{\otimes (k+1)N}$ satisfies this condition).

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

Actually, most statistical problems are usually formulated in the inductive setting, and one may wonder about the pertinence 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 $k = +\infty$). 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 the transductive inference to label 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: one can 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. This fact is discussed in more details in what follows.

1.2. **The model.** We assume that we are given a set of regression functions:

$$\mathcal{R} = \{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\},$$

where $(\Theta, \mathcal{F})$ is a measurable set of parameters. The function $(x, \theta) \mapsto f_\theta(x)$ is assumed to be measurable:

$$(\mathcal{X} \times \Theta, \mathcal{B}_\mathcal{X} \otimes \mathcal{F}) \to (\mathcal{Y}, \mathcal{B}_\mathcal{Y}).$$

Note that this does not mean that we are considering a single parametric model, we do not necessarily assume that $\Theta \subset \mathbb{R}^d$ for any $d \in \mathbb{N}$. We will actually be more interested in situations where Θ is a disjoint union of subsets of various dimensions:

$$\Theta = \bigsqcup_{m \in \mathcal{M}} \Theta_m$$

for a finite (or countable) set $\mathcal{M}$, and where $\Theta_m \subset \mathbb{R}^{d_m}$, the dimension d_m depending on the model.

1.3. **Risk and loss functions.**

Definition 1.1. We put, for any measurable nonnegative function $\psi : \mathcal{Y}^2 \to \mathbb{R}$ for any $\theta \in \Theta$, $\psi_i(\theta) = \psi\left(f_\theta(X_i), Y_i\right)$ and:

$$r_1(\psi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \psi_i(\theta),$$

$$r_2(\psi, \theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \psi_i(\theta),$$

$$R(\psi, \theta) = \frac{1}{N} \sum_{i=1}^{N} p_i \left\{ \psi\Big[f_\theta(X), Y\Big] \right\}.$$

We will moreover use the notation $r(\psi, \theta) = r_1(\psi, \theta)$ in the inductive setting.

If ψ is a loss function, namely $\psi_i(\theta)$ measures a distance between Y_i and $f_\theta(X_i)$, then the risk $r_2(\psi, \theta)$ is the quantity that we want to minimize (with respect to θ) in the transductive setting. In the inductive setting, $R(\psi, \theta)$ is to be minimized. In both cases however, these quantities are not observable to the statistician (only r_1 is, we will use the notation r_1 in the transductive setting and r in the inductive setting.). Note that in the inductive setting, i. i. d. case, where $P_N = P^{\otimes N}$ we have simply:

$$R(\psi, \theta) = P\left\{ \psi\Big[f_\theta(X), Y\Big] \right\}.$$

Example 1.1. In the classification setting, $\mathcal{Y} = \{y_1, ..., y_p\}$ with $p \geq 2$ and we use the loss function:

$$\psi(y, y') = \delta_y(y').$$

In the particular case where $p = 2$ we choose $\mathcal{Y} = \{-1, +1\}$ and we have:

$$\psi(y, y') = \mathbb{1}_{\mathbb{R}_+^*}(yy').$$

However, in many practital situations, algorithmic considerations lead to use a convex upper bound of this loss, like:

$$\psi(y, y') = (1 - yy')_+ = \max(1 - yy', 0) \quad \text{the "hinge loss",}$$

$$\psi(y, y') = \exp(-yy') \quad \text{the exponential loss,}$$

$$\psi(y, y') = (1 - yy')^2 \quad \text{the least square loss.}$$

For example, Cortes and Vapnik [14] generalized the SVM technique to non-separable data using the hinge loss, while Schapire, Freund, Bartlett and Lee [34] gave a statistical interpretation of boosting algorithm thanks to the exponential loss. See Zhang [46] for a complete study of the performance of classification methods using these loss functions.

Example 1.2. In the case where $\mathcal{Y} = \mathbb{R}$, it is usual to take: $\psi(y, y')$ as a distance between y and y'. For example, a widely used loss function is:

$$\psi(y, y') = |y - y'|^p$$

where $p \in [1, +\infty)$. For $p = 2$ we obtain the standard least square regression setting.

Definition 1.2. We put, in the case where $\mathcal{Y} = \mathbb{R}$ and where $\mathcal{B}_{\mathcal{Y}}$ is the Borel σ-algebra on $\mathbb{R}$, for any $p \in [1, +\infty)$:

$$l^p : \mathbb{R}^2 \to \mathbb{R}_+$$
$$(y, y') \mapsto |y - y'|^p.$$

Definition 1.3. For any ψ, let us put:

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

(the dependence with respect to ψ is not made explicit by the notation but there will we no ambiguity in the thesis, as ψ will we fixed once and for all).

1.4. **PAC-Bayesian approach and the Legendre transform of the Kullback divergence function.** Note that in statistical learning, the bounds on the risk $R(\psi, \hat{\theta})$ of an estimator $\hat{\theta}$ often depends on the empirical risk of $\hat{\theta}$, $r(\psi, \hat{\theta})$, and on a measure of the complexity of the submodel of Θ used to build $\hat{\theta}$.

In the PAC-Bayesian approach, initiated by McAllester [28, 29] (note however that the techniques used in this part are closer to the ones initiated by Catoni [11]), we do not consider any longer complexity measures of subspaces of Θ. This structure is replaced by the use of a "prior" probability measure over the parameter set Θ: $\pi \in \mathcal{M}_+^1(\Theta)$. Note that this measure is called prior in reference to Bayesian analysis, however its interpretation is different here. We do not think of it as a stochastic modelization of $\theta \in \Theta$. Its only role is to replace the structure of submodels of Θ. The aim of the PAC-Bayesian approach is to obtain PAC bounds on the integrated risk:

$$\int_\Theta r_2(\psi, \theta)\rho(d\theta) = \rho[r_2(\psi, .)] \quad \text{or} \quad \int_\Theta R(\psi, \theta)\rho(d\theta) = \rho[R(\psi, .)],$$

(according to the fact that we are in the transductive, or inductive setting) where $\rho \in \mathcal{M}_+^1(\Theta)$ is whatever posterior distribution, depending on π and on the observed data. The bounds here will depend on the empirical counterpart of $\rho[R(\psi, .)]$: $\rho[r(\psi, .)]$, and on a measure of the distance between ρ and π, that replaces the complexity term in the approach using submodels.

This measure of the distance between ρ and π will be made by the use of the Kullback divergence.

Definition 1.4. We let $\mathcal{K}(\rho, \pi)$ denote the Kullback divergence between ρ and π, given by:

$$\mathcal{K}(\rho, \pi) = \begin{cases} \rho \log\left[\frac{d\rho}{d\pi}(.)\right] & \text{if } \rho \ll \pi, \\ \\ \infty \text{ otherwise .} \end{cases}$$

Note that the fact to upper bound $\rho[R(\psi,.)]$ rather than $R(\psi,\hat{\theta})$ for some estimator seems to be unnatural. However, note the following facts:

- the quantity $\rho[R(\psi,.)]$ is the mean risk, under P, of the procedure that consists, for every X, to drawn $\hat{\theta}\in\Theta$ according to ρ and to predict Y by $f_{\hat{\theta}}(X)$;
- if Θ is countable, π is under the form $\pi=\sum_{\theta\in\Theta}p_\theta\delta_\theta$ with every $p_\theta\geq 0$ and $\sum_{\theta\in\Theta}p_\theta=1$, and the choice of $\rho=\delta_{\hat{\theta}}$ for some estimator $\hat{\theta}$ leads to $\rho[R(\psi,.)]=R(\psi,\hat{\theta})$, and $\mathcal{K}(\rho,\pi)=\log p_\theta^{-1}$;
- if Θ is not countable, and if we want to upper bound the risk of a given estimator $\hat{\theta}$ using a PAC-Bayesian technique, we can take ρ as the uniform measure over a small neighborhood of θ, this approach is detailed later;
- finally, in some cases of interest (for example, linear least square regression, as we will see later in this part), $\theta\mapsto R(\psi,\theta)$ is convex, and so by Jensen's inequality:

$$\rho[R(\psi,.)]\geq R\left[\psi,\int_\Theta\theta\rho(d\theta)\right],$$

which means that we are able to upper bound $R(\psi,\hat{\theta})$ for every estimator of this form.

The following definitions and lemma will be used in this whole part.

Definition 1.5. For any measurable function $h:\Theta\to\mathbb{R}$, for any measure $\rho\in\mathcal{M}_+^1(\Theta)$ we put:

$$\rho(h)=\sup_{B\in\mathbb{R}}\int_{\theta\in\Theta}[h(\theta)\wedge B]\,\rho(d\theta).$$

Lemma 1.1 (Legendre transform of the Kullback divergence function)**.** *For any measurable function $h:\Theta\to\mathbb{R}$ such that $\pi\exp[h(\theta)]<+\infty$ we have:*

$$\log\pi\exp(h)=\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\Big(\rho(h)-\mathcal{K}(\rho,\pi)\Big),\tag{1.1}$$

with convention $\infty-\infty=-\infty$. Moreover, as soon as h is upper-bounded on the support of π, the supremum with respect to ρ in the right-hand side is reached for the Gibbs distribution, $\pi_{\exp(h)}$ given by:

$$\forall\theta\in\Theta,\quad\frac{d\pi_{\exp(h)}}{d\pi}(\theta)=\frac{\exp[h(\theta)]}{\pi[\exp(h)]}.$$

Proof. We give the proof given in Catoni [10]. Let us assume that h is upper-bounded on the support of π. Let us remark that ρ is absolutely continuous with respect to π if and only if it is absolutely continuous with respect to $\pi_{\exp(h)}$. Let us assume that this is the case, then we have:

$$\begin{aligned}\mathcal{K}(\rho,\pi_{\exp(h)})=\rho\left\{\log\left(\frac{d\rho}{d\pi}\right)-h\right\}+\log\pi\exp(h)\\ =\mathcal{K}(\rho,\pi)-\rho(h)+\log\pi\exp(h).\end{aligned}$$

The left-hand side of this equation is nonnegative and cancels only for $\rho=\pi_{\exp(h)}$. Note that this equation is still valid if ρ is not absolutely continuous with respect to π (it just says that $+\infty=+\infty$ in this case). So we obtain:

$$0=\inf_{\rho\in\mathcal{M}_+^1(\Theta)}[\mathcal{K}(\rho,\pi)-\rho(h)]+\log\pi\exp(h).$$

This proves the second part of lemma 1.1. For the first part, we do not assume any longer that h is upper bounded on the support of π. Then we have:

$$\log \pi \exp(h) = \sup_{B \in \mathbb{R}} \log \pi \exp(h \wedge B) = \sup_{B \in \mathbb{R}} \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho\left(h \wedge B\right) - \mathcal{K}(\rho, \pi)\right]$$

$$= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \sup_{B \in \mathbb{R}} \left[\rho\left(h \wedge B\right) - \mathcal{K}(\rho, \pi)\right]$$

$$= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \sup_{B \in \mathbb{R}} \left[\rho\left(h \wedge B\right)\right] - \mathcal{K}(\rho, \pi) \right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho(h) - \mathcal{K}(\rho, \pi)\right].$$

□

2. PAC-Bayesian regression in the inductive setting

We use in this whole section the notations devoted to the inductive setting defined in the introduction. We also assume that ψ is chosen in such a way that for any $\theta \in \Theta$, $R(\psi, \theta)$ exists and belongs to $\mathbb{R}$. Finally, we fix a prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$.

2.1. Main lemma.

We need the following definition, that is used in what follows.

Definition 2.1. We put, for any $\alpha \in \mathbb{R}_+^*$:

$$\Phi_\alpha : \left(-\infty, \frac{1}{\alpha}\right) \to \mathbb{R}$$

$$t \mapsto -\frac{\log\left(1 - \alpha t\right)}{\alpha}.$$

Note that Φ_α is invertible, that for any $u \in \mathbb{R}$:

$$\Phi_\alpha^{-1}(u) = \frac{1 - \exp\left(-\alpha u\right)}{\alpha},$$

and that $\frac{2(\Phi_\alpha(x) - x)}{\alpha x^2} \underset{x \to 0}{\longrightarrow} 1$. Also note that for $\alpha > 0$, Φ_α is convex and that $\Phi_\alpha(x) \geq x$. For $\alpha < 0$, Φ_α is concave and $\Phi_\alpha(x) \leq x$.

We can now state the lemma.

Lemma 2.1. *We have, for any $\lambda \in \mathbb{R}_+^*$, for any $\theta \in \Theta$:*

$$P_N \exp\left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, \theta\right) \right] - \frac{\lambda}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} \left(\psi_i(\theta) \wedge \frac{N}{\lambda} \right) \right\} = 1.$$

The proof being almost trivial, let us give it now and make the remarks about the lemma after the proof.

Proof. For any $\lambda \in \mathbb{R}_+^*$, for any $\theta \in \Theta$:

$$P_N \exp\left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, \theta\right) \right] - \frac{\lambda}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} \left(\psi_i(\theta) \wedge \frac{N}{\lambda} \right) \right\}$$

$$P_N \exp\left\{ \sum_{i=1}^N \log\left[1 - \frac{\lambda}{N}\left(\psi_i(\theta) \wedge \frac{N}{\lambda}\right)\right] - N \log\left[1 - \frac{\lambda}{N} R\left(\psi \wedge \frac{N}{\lambda}, \theta\right)\right] \right\}$$

$$= \prod_{i=1}^N P_N \left[\frac{1 - \frac{\lambda}{N}\left(\psi_i(\theta) \wedge \frac{N}{\lambda}\right)}{1 - \frac{\lambda}{N} R\left(\psi \wedge \frac{N}{\lambda}, \theta\right)} \right].$$

□

Let us now make the following remarks about this result. First of all, we can deduce from it the following deviation inequality, by upper bounding the function $\mathbb{1}_{\mathbb{R}_+^*}(.)$ by $\exp(.)$: for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, for any $\theta \in \Theta$:

$$P_N\left\{R\left(\psi \wedge \frac{N}{\lambda}, \theta\right) \leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right), \theta\right] + \frac{\log \frac{1}{\varepsilon}}{\lambda}\right\}\right\} \geq 1 - \varepsilon.$$

So we have a control of the quantity $R\left(\psi \wedge \frac{N}{\lambda}, \theta\right)$ based only on empirical quantities. Note the interest of this point: in most deviation inequalities, we can have a control of $R(\psi, \theta)$ that depends on empirical quantities and on a theoretical variance of $\psi[f_\theta(X), Y]$ under every p_i. This distribution being unknown to the statistician, however, the use of these inequalities does not allow a direct control of $R(\psi, \theta)$ by empirical quantities only. Here, we control $R\left(\psi \wedge \frac{N}{\lambda}, \theta\right)$ by empirical quantities only.

The problem is that we control a thresholded version of the risk: $R\left(\psi \wedge \frac{N}{\lambda}, \theta\right)$ and not $R(\psi, \theta)$. Actually, let us remark that this point is natural. If we observe N values $\psi[f_\theta(X_i), Y_i]$ quite small (of order 1, say), it is possible that $\psi[f_\theta(X), Y]$ takes a large value (say of order N/λ) with probability smaller than $1/N$: this event is too rare to be observed with our sample. However, it leads to a change of order $1/\lambda$ of the risk: that is precisely the order of the bound. In order to see this, let us present the following toy example.

Example 2.1. We assume that $P_N = P^{\otimes N}$ and that $\mathcal{X} = \{x\}$ and Y takes two values: $Y = 0$, with probability $1 - 1/N$, or $Y = cN/\lambda$, with small probability $1/N$. Finally, we consider a θ such that $f_\theta(x) = 0$, and $\psi = l^1$. Note that:

$$R(\psi, \theta) = \left(1 - \frac{1}{N}\right) 0 + \frac{1}{N} \frac{cN}{\lambda} = \frac{c}{\lambda}.$$

Now, with probability at least $(1 - 1/N)^N$ we have every Y_i equal to 0 and so:

$$r\left(\Phi_{\lambda/N} \circ \psi, \theta\right) = 0.$$

So we cannot hope to upper bound $R(\psi, \theta)$ by a bound in $r(\psi, \theta) + \log(\varepsilon^{-1})/\lambda$, because we can choose c as large as we want. However, note that:

$$R\left(\psi \wedge \frac{N}{\lambda}, \theta\right) = \frac{1 \wedge c}{\lambda}.$$

So, we cannot guess the probability to have $\psi[f_\theta(X), Y]$ greater than N/λ without any assumptions: that would mean that we are able to estimate accurately the probability of events that we have not observed. This can be done, if we formulate some assumptions about P_N or ψ. In order to explain this, let us introduce the following definition.

Definition 2.2. For any $t \in \mathbb{R}$, $\alpha \in \mathbb{R}_+^*$, with the notation $(t)_+ = t \vee 0$, and for any $\psi : \mathcal{Y}^2 \to \mathbb{R}$ and $\theta \in \Theta$:

$$\Delta_\alpha(\psi, \theta) = R\left[\left(\psi - \frac{1}{\alpha}\right)_+, \theta\right].$$

Note that:

$$\Delta_{\frac{\lambda}{N}}(\psi, \theta) = R\left(\psi \wedge \frac{N}{\lambda}, \theta\right) - R(\psi, \theta).$$

So the deviation inequality becomes: for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, for any $\theta \in \Theta$:

$$P_N\left\{R(\psi, \theta) \leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right), \theta\right] + \frac{\log \frac{1}{\varepsilon}}{\lambda}\right\} + \Delta_{\frac{\lambda}{N}}(\psi, \theta)\right\} \geq 1 - \varepsilon.$$

Now, we can see some examples of hypotheses about P or ψ which allow to upper bound the term $\Delta_{\lambda/N}(\psi, \theta)$. Basically, note that, if ψ is upper-bounded by a constant $c > 0$, then we have, as soon as $\lambda \leq N/c$:

$$\Delta_{\frac{\lambda}{N}}(\psi, \theta) = 0.$$

We have also the following result.

Lemma 2.2. *Let us assume that we are in the case where $\mathcal{Y} = \mathbb{R}$ and where $\mathcal{B}_{\mathcal{Y}}$ is the Borel σ-algebra on $\mathbb{R}$, and that $\psi = l^p$ for a $p \in [1, +\infty)$. Let us assume that we are in the i. i. d. case, the distribution p_i of every pair $Z_i = (X_i, Y_i)$ is actually P and so $P_N = P^{\otimes N}$. Let us assume that we know that the true regression function f (defined as a measurable version of $x \mapsto P[Y|X = x]$) is such that $\|f\|_\infty \leq C/2$ for some $C \geq 0$. Then it makes sense to consider only parameters $\theta \in \Theta$ in such a way that $\|f_\theta\|_\infty \leq C/2$. Let us moreover assume that there are two constants $b > 0$ and $B < +\infty$ such that for any $x \in \mathcal{X}$:*

$$P\left\{\exp\Big[b|Y - f(X)|\Big] \Big| X = x\right\} \leq B.$$

Then we have, for any $\theta \in \Theta$ such that $\|f_\theta\|_\infty \leq C/2$:

$$\Delta_{\frac{\lambda}{N}}(l^p, \theta) \leq B \exp\left\{b\left[C - 2^{\frac{1}{p}-1}\left(\frac{N}{\lambda}\right)^{\frac{1}{p}}\right]\right\} \int_0^{+\infty} \exp\left[-b2^{\frac{1}{p}-1}t^{\frac{1}{p}}\right] dt.$$

Proof. We have:

$$\begin{aligned}
R\left[\left(\psi - \frac{N}{\lambda}\right)_+, .\right] &= P\left[\left(|f_\theta(X) - Y|^p - \frac{N}{\lambda}\right)_+\right] \\
&= \int_0^{+\infty} P\left[|f_\theta(X) - Y|^p - \frac{N}{\lambda} > t\right] dt \\
&= \int_0^{+\infty} P\left[|f_\theta(X) - Y| > \left(t + \frac{N}{\lambda}\right)^{\frac{1}{p}}\right] dt \\
&\leq \int_0^{+\infty} P\left\{|f(X) - Y| > 2^{\frac{1}{p}-1}\left[t^{\frac{1}{p}} + \left(\frac{N}{\lambda}\right)^{\frac{1}{p}}\right] - C\right\} dt \\
&\leq \int_0^{+\infty} P \exp\left\{|f(X) - Y| - 2^{\frac{1}{p}-1}\left[t^{\frac{1}{p}} + \left(\frac{N}{\lambda}\right)^{\frac{1}{p}}\right] + C\right\} dt \\
&\leq B \exp\left\{b\left[C - 2^{\frac{1}{p}-1}\left(\frac{N}{\lambda}\right)^{\frac{1}{p}}\right]\right\} \int_0^{+\infty} \exp\left[-b2^{\frac{1}{p}-1}t^{\frac{1}{p}}\right] dt.
\end{aligned}$$

This is exactly the result claimed in the lemma. □

Example 2.2. The conditions of lemma 2.2, with $p = 2$, are satisfied when $Y - f(X)$ is independent of X and is Gaussian with mean 0 and variance σ^2. In this case we wan check that for any $b > 0$ we can take:

$$B = \exp\left(\frac{b^2\sigma^2}{2}\right).$$

However in this particular case we have the specific bound:

$$P\Big[|f(X) - Y| \geq t\Big] \leq \sqrt{\frac{\sigma^2}{2\pi}} \frac{\exp\left(\frac{-t^2}{2\sigma^2}\right)}{t}$$

leading to the following result, valid for any $\theta \in \Theta$ such that $\|f_\theta\|_\infty \leq C/2$ and for any λ such that $\lambda < N/(2C^2)$:

$$\Delta_{\frac{\lambda}{N}}(l^2, \theta) \leq 2\sigma^3 \sqrt{\frac{2}{\pi}} \frac{\exp\left[\frac{1}{2\sigma^2}\left(C^2 - \frac{N}{2\lambda}\right)\right]}{\sqrt{\frac{N}{2\lambda} - C^2}}.$$

2.2. **A basic PAC-Bayesian theorem.**

Theorem 2.3. *For any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\Phi_{\frac{\lambda}{N}}\left\{\rho\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\} \leq \rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right), .\right]\right\} + \frac{\mathcal{K}(\rho, \pi) + \log\frac{1}{\varepsilon}}{\lambda}.$$

Proof. We apply lemma 2.1 and integrate it with respect to π to obtain that:

$$\pi P_N \exp\left\{\lambda \Phi_{\frac{\lambda}{N}}\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right] - \frac{\lambda}{N}\sum_{i=1}^N \Phi_{\frac{\lambda}{N}}\left(\psi_i(.) \wedge \frac{N}{\lambda}\right)\right\} = 1.$$

Using Fubini's theorem and multiplying both members by ε we have:

$$P_N \pi \exp\left\{\lambda \Phi_{\frac{\lambda}{N}}\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right] - \frac{\lambda}{N}\sum_{i=1}^N \Phi_{\frac{\lambda}{N}}\left(\psi_i(.) \wedge \frac{N}{\lambda}\right) - \log\frac{1}{\varepsilon}\right\} = \varepsilon.$$

Now, we apply lemma 1.1 and we obtain:

$$P_N \exp \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{\lambda\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\} - \frac{\lambda}{N}\sum_{i=1}^N \rho\left[\Phi_{\frac{\lambda}{N}}\left(\psi_i(.) \wedge \frac{N}{\lambda}\right)\right] - \mathcal{K}(\rho, \pi) - \log\frac{1}{\varepsilon}\right\} = \varepsilon.$$

This implies that:

$$P_N\left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{\lambda\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\} - \frac{\lambda}{N}\sum_{i=1}^N \rho\left[\Phi_{\frac{\lambda}{N}}\left(\psi_i(.) \wedge \frac{N}{\lambda}\right)\right] - \mathcal{K}(\rho, \pi) - \log\frac{1}{\varepsilon}\right\} \geq 0\right] \leq \varepsilon.$$

As $\Phi_{\frac{\lambda}{N}}$ is convex we have:

$$\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\} \geq \Phi_{\frac{\lambda}{N}}\left\{\rho\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\}.$$

So we obtain:

$$P_N\left\{\forall \rho \in \mathcal{M}_+^1(\Theta), \quad \Phi_{\frac{\lambda}{N}}\left\{\rho\left[R\left(\psi \wedge \frac{N}{\lambda}, .\right)\right]\right\} \leq \rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right), .\right]\right\} + \frac{\mathcal{K}(\rho, \pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\} \geq 1 - \varepsilon,$$

that is the stated result. □

We are now going to see how to apply this theorem to perform regression estimation. Let us assume just for a while that on the support of π, $P[\psi(Y, f_\theta(X)) \leq 1] = 1$. This is for example the case if we take $\psi = l^2$, $\mathcal{Y} = [0, 1]$ and $0 \leq f_\theta(x) \leq 1$

for any $(x, \theta) \in \mathcal{X} \times \Theta$. Then we have, for any $\varepsilon > 0$, for any $\lambda \in [1, N)$, with P_N-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\Phi_{\frac{\lambda}{N}} \left\{ \rho \left[R\left(\psi, .\right) \right] \right\} \leq \rho \left[r\left(\Phi_{\frac{\lambda}{N}} \circ \psi, . \right) \right] + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda},$$

or:

$$\rho \left[R\left(\psi, .\right) \right] \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho \left[r\left(\Phi_{\frac{\lambda}{N}} \circ \psi, . \right) \right] + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} \right\} = \mathcal{B}(\rho, \varepsilon, \lambda).$$

Example 2.3 (Finite or countable parameter set)**.** Let us first consider the case where Θ is finite, and where we take:

$$\pi = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \delta_\theta.$$

The restriction $\rho \in \{\delta_\theta, \theta \in \Theta\}$ leads to the following result: for any $\varepsilon > 0$, for any $\lambda \in [1, N)$, with P_N-probability at least $1 - \varepsilon$, for any $\theta \in \Theta$:

$$R\left(\psi, \theta\right) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left[r\left(\Phi_{\frac{\lambda}{N}} \circ \psi, \theta \right) + \frac{\log |\Theta| + \log \frac{1}{\varepsilon}}{\lambda} \right].$$

Note that the fact that the complexity of a finite model Θ should be measured by $\log |\Theta|$ is a well known fact of learning theory, see Vapnik [41] for example. However, note that here it is possible to give more importance to some values $\theta \in \Theta$, and to deal with the case where Θ is countable, by changing π. Let us choose, for any $\theta \in \Theta$, $p_\theta \geq 0$ such that $\sum_{\theta \in \Theta} p_\theta = 1$. Let us take:

$$\pi = \sum_{\theta \in \Theta} p_\theta \delta_\theta$$

and still $\rho \in \{\delta_\theta, \theta \in \Theta\}$, we obtain this time: for any $\varepsilon > 0$, for any $\lambda \in [1, N)$, with P_N-probability at least $1 - \varepsilon$, for any $\theta \in \Theta$:

$$R\left(\psi, \theta\right) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left[r\left(\Phi_{\frac{\lambda}{N}} \circ \psi, \theta \right) + \frac{\log \frac{1}{p_\theta} + \log \frac{1}{\varepsilon}}{\lambda} \right].$$

Note that this upper bound is valid even if the "prior" distribution $(p_\theta)_{\theta \in \Theta}$ does not represent a belief on the "true" value of the parameter as in the Bayesian case. The interpretation of π is not Bayesian. So, for a given λ, we propose to choose the estimator:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \Phi_{\frac{\lambda}{N}}^{-1} \left[r\left(\Phi_{\frac{\lambda}{N}} \circ \psi, \theta \right) + \frac{\log \frac{1}{p_\theta} + \log \frac{1}{\varepsilon}}{\lambda} \right].$$

We will discuss the choice of λ in a more general setting.

Now, we come back to a general parameter set Θ. By an application of lemma 1.1 we know that, for a given λ, $\mathcal{B}(\rho, \varepsilon, \lambda)$ reaches its minimum for

$$\rho = \pi_{\exp\left[-\lambda r\left(\Phi_{\lambda/N} \circ \psi, .\right)\right]}$$

and that this minimum is equal to:

$$\mathcal{B}\left(\pi_{\exp\left\{-\lambda r\left(\Phi_{\lambda/N} \circ \psi, .\right)\right]}, \varepsilon, \lambda \right) = \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \frac{\log \frac{1}{\varepsilon} - \log \pi \exp\left[-\lambda r\left(\Phi_{\lambda/N} \circ \psi, .\right)\right]}{\lambda} \right\}.$$

The choice of λ is more problematic here because λ is not allowed to be data dependent. However, there is a simple way to solve this problem, given by Catoni [10]. Let us choose $a \in (1, N/2)$ and put:

$$\Lambda = \left\{ a^l, \quad 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{2}}{\log a} \right\rfloor \right\}.$$

By a union bound on Λ we obtain, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$ and $\lambda \in \Lambda$:

$$\rho\left[R\left(\psi, .\right)\right] \leq \mathcal{B}\left(\rho, \frac{\varepsilon}{|\Lambda|}, \lambda\right).$$

Now, let us remark that for any $\lambda \in [1, N/2]$ there is a $\lambda' \in \Lambda$ such that $\lambda \leq \lambda' \leq a\lambda$. We can state the following corollary.

Corollary 2.4. *Let us assume that ψ takes values in $[0,1]$. For any $\varepsilon > 0$, and $a \in (1, N/2)$, with P_N-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\rho\left[R\left(\psi, .\right)\right] \leq \inf_{\lambda \in [1, N/2]} \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho\left[r\left(\Phi_{\frac{a\lambda}{N}} \circ \psi, .\right)\right] + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1 + \left\lfloor \frac{\log \frac{N}{2}}{\log a} \right\rfloor}{\varepsilon}}{\lambda} \right\}.$$

So we can now propose the following estimation method:

- choose ε and a;
- compute

$$\lambda_0 = \arg \min_{\lambda \in [1, N/2]} \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho\left[r\left(\Phi_{\frac{a\lambda}{N}} \circ \psi, .\right)\right] + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1 + \left\lfloor \frac{\log \frac{N}{2}}{\log a} \right\rfloor}{\varepsilon}}{\lambda} \right\};$$

- choose $\hat{\rho} = \pi_{\exp\left[-\lambda_0 r\left(\Phi_{a\lambda_0/N} \circ \psi, .\right)\right]}$;
- for any new value X predict Y by $\hat{Y}$ obtained by drawing θ from the distribution $\hat{\rho}$ and taking $\hat{Y} = f_\theta(X)$.

We know that the mean risk of this procedure is $\hat{\rho}\left[R\left(\psi, \theta\right)\right]$ and so according to the corollary it cannot exceed:

$$\Phi_{\frac{\lambda_0}{N}}^{-1} \left\{ \frac{\log \frac{1 + \left\lfloor \frac{\log \frac{N}{2}}{\log a} \right\rfloor}{\varepsilon} - \log \pi \exp\left[-\lambda_0 r\left(\Phi_{\lambda_0/N} \circ \psi, .\right)\right]}{\lambda_0} \right\}.$$

Remark 2.1. As we already mentioned in the introduction, note that if $\theta \mapsto R(\psi, \theta)$ is convex (this is the case for example if $\psi = l^p$ and if $\theta \mapsto f_\theta$ is linear) then we have:

$$\rho\left[R\left(\psi, \theta\right)\right] \geq R\left[\psi, \rho(\theta)\right],$$

in this notation θ is the canonical process on $(\Theta, \mathcal{T}, \rho)$, that means that we can write:

$$\rho(\theta) = \int_\Theta t \rho(dt).$$

This avoids us the randomization step: we can choose the function $f_{\rho(\theta)}$ as an estimation of the regression function. The procedure becomes, for any new value X, predict Y by $\hat{Y} = f_{\hat{\rho}(\theta)}(X)$.

Let us now give a look at what can be done in the case of an unbounded function ψ. For the sake of simplicity we state the following corollary of Theorem 2.3 with a single value λ, but the union bound on Λ can still be done.

Corollary 2.5. *For any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\rho\left[R\left(\psi,.\right)\right] \leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right),.\right]\right\} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\} + \rho\left[\Delta_{\frac{\lambda}{N}}(\psi,.)\right].$$

So, in the general case, we are going to use the same method than in the case where ψ is bounded, with an additional hypothesis to obtain an explicit upper bound for the term:

$$\Delta_{\frac{\lambda}{N}}(\psi,\theta) = R\left[\left(\psi - \frac{N}{\lambda}\right)_+,\theta\right].$$

2.3. **Deviation under the posterior.** The aim of this subsection is to prove that it makes sense to draw once and for all a value $\hat{\theta}$ from a well-chosen posterior distribution ρ instead of drawing a new value for every new prevision.

Theorem 2.6. *For any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$, for any $\eta > 0$, with ρ-probability at least $1-\eta$ over θ:*

$$R(\psi,\theta) \leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right),\theta\right] + \frac{1}{\lambda}\left[-\log\pi\exp\left\{-\lambda r\left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda)\right),.\right]\right\} + \log\left(\frac{d\rho}{d\pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}}(\theta)\right) + \log\frac{1}{\eta\varepsilon}\right]\right\} + \Delta_{\frac{\lambda}{N}}(\psi,\theta).$$

In particular, for the particular choice of the optimal Gibbs posterior, with P_N-probability at least $1-\varepsilon$, for any $\eta > 0$, with $\pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}$-probability at least $1-\eta$ over θ:

$$R(\psi,\theta) \leq \Phi_{\frac{\lambda}{N}}^{-1}\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right),\theta\right] + \frac{1}{\lambda}\left[-\log\pi\exp\left\{-\lambda r\left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda)\right),.\right]\right\} + \log\frac{1}{\eta\varepsilon}\right]\right\} + \Delta_{\frac{\lambda}{N}}(\psi,\theta).$$

Proof. For the sake of shorteness, let us put:

$$\hat{\rho} = \pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}.$$

Let follow the proof of Theorem 2.3. At a certain point we have, for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$:

$$\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi\wedge\frac{N}{\lambda},.\right)\right]\right\} - \rho\left\{r\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right),.\right]\right\} - \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right\} \leq 0,$$

and this inequality implies:

$$\log \rho \exp\Bigg\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] - \log \frac{1}{\varepsilon} + \log \pi \exp\left\{ -\lambda r \left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda) \right), . \right] \right\} + \log\left(\frac{d\hat{\rho}}{d\rho} \right) \Bigg\}$$

$$= \log \hat{\rho} \exp\Bigg\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] - \log \frac{1}{\varepsilon} + \log \pi \exp\left\{ -\lambda r \left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda) \right), . \right] \right\} \Bigg\}$$

$$= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \Bigg\{ \lambda \rho \left\{ \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] \right\} - \log \frac{1}{\varepsilon} + \log \pi \exp\left\{ -\lambda r \left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda) \right), . \right] \right\} - \mathcal{K}(\rho, \hat{\rho}) \Bigg\}$$

$$= \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \Bigg\{ \lambda \rho \left\{ \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] \right\} - \lambda \rho \left\{ r \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), . \right] \right\} - \log \frac{1}{\varepsilon} - \mathcal{K}(\rho, \pi) \Bigg\} \leq 0.$$

This implies, for any η:

$$\hat{\rho}\Bigg\{ \theta \in \Theta : \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] \geq - \log \pi \exp\left\{ -\lambda r \left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda) \right), . \right] \right\} + \log\left(\frac{d\rho}{d\hat{\rho}}(\theta) \right) + \log \frac{1}{\eta \varepsilon} \Bigg\} \leq \hat{\rho} \exp\Bigg\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R\left(\psi \wedge \frac{N}{\lambda}, . \right) \right] - \log \frac{1}{\varepsilon \eta} + \log\left(\frac{d\rho}{d\hat{\rho}} \right) + \log \pi \exp\left\{ -\lambda r \left[\Phi_{\lambda/N} \circ \left(\psi \wedge (N/\lambda) \right), . \right] \right\} \Bigg\} \leq \eta.$$

This ends the proof. □

2.4. **Introduction of moment hypothesis.** We have seen that Theorem 2.3 (or corollary 2.5) leads to an observable bound as soon as some assumption (exponential moments, boundedness) is satisfied. Here, we propose another change of variable that leads to an observable bound under the hypothesis that for a given $s > 1$, for any $\theta \in \Theta$:

$$R(|\psi|^s, \theta) < +\infty.$$

Theorem 2.7. *Let us assume that for a given $s > 1$, for any $\theta \in \Theta$, $R(|\psi|^s, \theta) < +\infty$. For any $\alpha \in (0,1)$, for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\Phi_{\frac{\lambda}{N}} \left\{ \alpha \rho \left[R\left(\psi - \frac{1}{s} \left(\frac{(s-1)\lambda}{sN} \right)^{s-1} |\psi|^s, . \right) \right] \right\} \leq \rho \left\{ r \left[\Phi_{\frac{\lambda}{N}} \circ \left(\alpha\psi - \frac{\alpha}{s} \left(\frac{(s-1)\lambda}{sN} \right)^{s-1} |\psi|^s \right), . \right] \right\} + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda}.$$

Proof. This is just an application of Theorem 2.3 where we replace ψ by:

$$\alpha\left[\psi - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s\right],$$

and note that, for any $\lambda \in \mathbb{R}_+^*$:

$$\frac{\lambda\alpha}{N}\left[\psi - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s\right] = \alpha\left[\frac{\lambda\psi}{N} - \frac{1}{s}\left(\frac{s-1}{s}\right)^{s-1}\left|\frac{\lambda\psi}{N}\right|^s\right] \leq \alpha < 1.$$

□

A few lines may help to interpret Theorem 2.7. Note that this implies the following bound for $\rho\left[R\left(\psi\right)\right]$, with probability at least $1-\varepsilon$:

$$\begin{aligned}\rho\left[R\left(\psi\right)\right] &\\ \leq \frac{1}{\alpha}\Phi_{\frac{\lambda}{N}}^{-1}&\left\{\rho r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\alpha\psi - \frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s\right),.\right] + \frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\lambda}\right\}\\ &+\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\rho\left[R\left(|\psi|^s,.\right)\right].\end{aligned}$$

So this theorem is almost the same as Theorem 2.3 page 29, but the thresholding is replaced by the moment term:

$$\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\rho\left[R\left(|\psi|^s,.\right)\right], \tag{2.1}$$

allowing to replace the boundedness assumption by the existence of a moment. Remark that a new problem emerges as the term $R\left(|\psi|^s,.\right)$ in the right-hand side is not observable to the statistician. Additional hypothesis are required to control this term. For example we can use the following lemma.

Lemma 2.8. *Let us assume that we are in the i. i. d. case, $P_N = P^{\otimes N}$. Let us assume that $\mathcal{Y} = \mathbb{R}$, $\psi = l^p$, and that for any $(x,\theta) \in \mathcal{X}\times\Theta$, $|f_\theta(x)| \leq C$. Let us assume that P is such that:*

$$P(Y^{sp}) \leq M_{sp}$$

for some (known) constant M_{sp}. Then we have:

$$R\left(|\psi|^s,\theta\right) \leq 2^{2s-1}\left(M_{sp} + C^{sp}\right).$$

Remark 2.2. Note the role of the various parameters in Theorem 2.7. The parameter $s > 1$ is the order of the moment of ψ we assume to exist. When s becomes larger, the hypothesis $R(|\psi|^s,\theta) < +\infty$ for any θ becomes more restrictive. However, we should take the largest possible s: we will see in the next subsection that we expect $\lambda \ll N$ and so a large s will help to make the moment term (Equation 2.1) the smaller possible. The parameter λ plays the same role than in Theorem 2.3 page 29; a union bound and an optimization with respect to λ is recommended in general. Finally, the parameter α appears in the change of variable used to obtain the moment bound. A union bound and an optimization with respect to α can also be performed. However, note that in some parts of this work we will see that we can arbitrarily choose $\alpha = 1/2$ with no loss on the order of magnitude of the bound.

2.5. **Bounds on the integrated risk.** In this subsection we are going to show that it is possible to upper bound the mean risk of our method:

$$P^{\otimes N}\Big\{\rho\left[R\left(\psi,\theta\right)\right]\Big\}.$$

Theorem 2.9. *For any $\lambda \in \mathbb{R}_+^*$, for any data-dependent posterior distribution $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$ we have:*

$$P_N\left\{\rho\left[R\left(\psi\wedge\frac{\lambda}{N},.\right)\right]\right\} \leq P_N\left[\Phi_{\frac{\lambda}{N}}\left\{\rho\left[R\left(\psi\wedge\frac{\lambda}{N},.\right)\right]\right\}\right]$$
$$\leq P_N\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}+\frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\}.$$

Proof. We follow the proof of Theorem 2.3 until we obtain:

$$P_N\exp\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\lambda\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi\wedge\frac{N}{\lambda},.\right)\right]\right\}\right.$$
$$\left.-\frac{\lambda}{N}\sum_{i=1}^N\rho\left[\Phi_{\frac{\lambda}{N}}\left(\psi_i(.)\wedge\frac{N}{\lambda}\right)\right]-\mathcal{K}(\rho,\pi)-\log\frac{1}{\varepsilon}\right\}=\varepsilon.$$

At this time, let us take $\varepsilon = 1$ and note that, by Jensen's inequality, we have:

$$\exp P_N\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\lambda\rho\left\{\Phi_{\frac{\lambda}{N}}\left[R\left(\psi\wedge\frac{N}{\lambda},.\right)\right]\right\}\right.\right.$$
$$\left.\left.-\frac{\lambda}{N}\sum_{i=1}^N\rho\left[\Phi_{\frac{\lambda}{N}}\left(\psi_i(.)\wedge\frac{N}{\lambda}\right)\right]-\mathcal{K}(\rho,\pi)\right\}\right]\leq 1.$$

This implies the theorem. □

Now, let us examine some consequences of Theorem 2.9 on a toy example.

Example 2.4 (Bounded linear least square regression)**.** Let us assume that we are in the i. i. d. case, $P_N = P^{\otimes N}$. Let us assume that $\mathcal{Y} = \mathbb{R}$, and that P is such that $P[Y \in (-1/2, 1/2)] = 1$. Let $\|.\|$ denote the euclidian norm on $\mathbb{R}^d$ and $\langle .,.\rangle$ the associated scalar product. We put, for any $d \in \mathbb{N}\setminus\{0\}$, $x \in \mathbb{R}^d$ and $\delta \geq 0$:

$$B_d(x,\delta) = \left\{x' \in \mathbb{R}^d, \|x - x'\| \leq \delta\right\}.$$

Let λ_d denote the Lebesgue measure on $\mathbb{R}^d$. Now, we assume that $\mathcal{X} = \Theta = B_d(0, 1/\sqrt{2})$, and that $f_\theta(x) = \langle\theta, x\rangle$. Note that this implies that, for any $(x,\theta) \in \mathcal{X}\times\Theta$ we have:

$$\frac{-1}{2} \leq f_\theta(x) \leq \frac{1}{2}.$$

Finally, we assume that $\psi = l^2$, so we have, for any $\theta \in \Theta$:

$$P\left[\psi\left(f_\theta(X), Y\right) \leq 1\right] = 1.$$

Note that we have, for any $\lambda \leq N/2$ and $p \in [0,1]$:

$$p \leq \Phi_{\frac{\lambda}{N}}(p) \leq p + \frac{\lambda}{2N}p^2$$

so we have almost surely, for any $\theta \in \Theta$:

$$r\left[\Phi_{\frac{\lambda}{N}}\circ\left(l^2\wedge\frac{N}{\lambda}\right),\theta\right] \leq r\left(l^2,\theta\right)+\frac{\lambda}{2N}r\left(l^4,\theta\right).$$

Moreover, let us assume that $\left\|\overline{\theta}\right\| < 1/2$. Let us finally choose the prior distribution π absolutely continuous with respect to λ_d and such that:

$$\frac{d\pi}{d\lambda}(\theta) = \frac{\mathbb{1}_\Theta(\theta)}{\lambda_d(\Theta)} = \frac{\mathbb{1}_\Theta(\theta)\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}\left(\frac{1}{\sqrt{2}}\right)^d}.$$

In this case, the theorem becomes, for any $\lambda \in [1, N/2]$, for any data-dependent posterior distribution $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$ we have:

$$\begin{aligned} P^{\otimes N}\left\{\rho\left[R\left(l^2,.\right)\right]\right\} &\leq P^{\otimes N}\left\{\rho\left[r\left(l^2,\theta\right) + \frac{\lambda}{2N} r\left(l^4,\theta\right)\right] + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\} \\ &\leq P^{\otimes N}\left\{\left(1+\frac{\lambda}{2N}\right)\rho\left[r\left(l^2,\theta\right)\right] + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\}. \end{aligned}$$

The optimal choice for ρ is the following Gibbs posterior:

$$\pi_{\exp[-\lambda r(l^2,.)]}.$$

However, in order to provide explicit computations, let use another posterior distribution. We put:

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d} r(l^2,\theta),$$

and:

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^N \left\langle \theta - \hat{\theta}, X_i\right\rangle^2.$$

Remark that the Gibbs posterior tends to concentrate, for large values of λ, around $\hat{\theta}$, at least when it belongs to Θ, otherwise it concentrates around $\tilde{\theta}$. The posterior we propose to use is build in order to mimic this point. Let us put, for any $\delta \in (0, 1/\sqrt{2})$:

$$\tilde{\theta}^\delta = \left(1 \wedge \frac{\frac{1}{\sqrt{2}} - \delta}{\|\tilde{\theta}\|}\right)\tilde{\theta}.$$

Now, let us put define the posterior $\tilde{\rho}^\delta$ such that:

$$\frac{\tilde{\rho}^\delta}{d\lambda}(\theta) = \frac{\mathbb{1}_{B_d\left(\tilde{\theta}^\delta,\delta\right)}(\theta)}{\lambda_d\left[B_d\left(\tilde{\theta}^\delta,\delta\right)\right]}.$$

Note that $\tilde{\theta}^\delta$ is chosen such that $B_d\left(\tilde{\theta}^\delta,\delta\right) \subset B_d(0,1/\sqrt{2})$ for any $\delta \in (0,1/\sqrt{2})$. So $\tilde{\rho}^\delta$ is absolutely continuous with respect to θ and we have:

$$\mathcal{K}(\rho,\pi) = d\log\frac{1}{\delta\sqrt{2}}.$$

Moreover, note that:

$$\begin{aligned} \rho\left[r\left(l^2,\theta\right)\right] &= \frac{1}{N}\sum_{i=1}^N \rho\left\{\left[Y_i - \langle .,X_i\rangle\right]^2\right\} \\ &\leq \frac{1}{N}\sum_{i=1}^N \left\{\left[Y_i - \left\langle\tilde{\theta}^\delta, X_i\right\rangle\right]^2 + \rho\left[\left\langle\tilde{\theta}^\delta - .,X_i\right\rangle^2\right]\right\} \\ &= \frac{1}{N}\sum_{i=1}^N \left[Y_i - \left\langle\hat{\theta},X_i\right\rangle\right]^2 + \frac{1}{N}\sum_{i=1}^N \left\langle\tilde{\theta}^\delta - \hat{\theta}, X_i\right\rangle^2 + \frac{\delta^2}{2} \\ &\leq \frac{1}{N}\sum_{i=1}^N \left[Y_i - \left\langle\hat{\theta},X_i\right\rangle\right]^2 + \frac{1}{N}\sum_{i=1}^N \left\langle\tilde{\theta} - \hat{\theta}, X_i\right\rangle^2 + \delta^2 \end{aligned}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - \left\langle \hat{\theta}, X_i \right\rangle\right]^2 + \frac{1}{N}\sum_{i=1}^{N}\left\langle \overline{\theta} - \hat{\theta}, X_i \right\rangle^2 + \delta^2 = r\left(l^2, \overline{\theta}\right) + \delta^2.$$

So we obtain, for any $\lambda \in [1, N/2]$, for any function $\delta : \mathcal{Z}^N \to (0, 1/\sqrt{2})$ (we will write $\delta(Z_1, ..., Z_N) = \delta$ for short):

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2, .\right)\right]\right\} \leq \left(1 + \frac{\lambda}{2N}\right)\left\{P^{\otimes N}\left[r\left(l^2, \overline{\theta}\right)\right] + \delta^2\right\} + \frac{d}{\lambda}\log\frac{1}{\delta\sqrt{2}}$$
$$\leq \left(1 + \frac{\lambda}{2N}\right)\left[R\left(l^2, \overline{\theta}\right) + \delta^2\right] + \frac{d}{\lambda}\log\frac{1}{\delta\sqrt{2}}.$$

Note that the optimal value for δ is:

$$\sqrt{\frac{d}{2\lambda\left(1 + \frac{\lambda}{2N}\right)}}.$$

However, as we are going to take λ of order $\sqrt{N}$, let us take:

$$\delta^* = \sqrt{\frac{d}{2\lambda}}.$$

We have, for any $\lambda \in [1, N/2]$:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta^*}\left[R\left(l^2, .\right)\right]\right\} \leq \left(1 + \frac{\lambda}{2N}\right)\left[R\left(l^2, \overline{\theta}\right) + \frac{d}{2\lambda}\right] + \frac{d}{\lambda}\log\sqrt{\frac{\lambda}{d}}$$
$$= R\left(l^2, \overline{\theta}\right) + \frac{d}{4N} + \frac{\lambda R\left(l^2, \overline{\theta}\right)}{2N} + \frac{d}{\lambda}\left[\frac{1}{2} + \log\sqrt{\frac{\lambda}{d}}\right].$$

If $R\left(l^2, \overline{\theta}\right) = 0$ we obtain:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta^*}\left[R\left(l^2, .\right)\right]\right\} \leq \frac{d}{4N} + \frac{d}{\lambda}\left[\frac{1}{2} + \log\sqrt{\frac{\lambda}{d}}\right]$$

and the choice $\lambda = N/2$ leads to the bound:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta^*}\left[R\left(l^2, .\right)\right]\right\} \leq \frac{d}{2N}\left[\frac{9}{2} + \log\frac{N}{2d}\right].$$

However, in most practical cases we expect to have $R\left(l^2, \overline{\theta}\right) > 0$. We then propose in order to explicit the computations to take the optimal value of λ when the logarithmic term is neglected:

$$\lambda = \sqrt{\frac{Nd}{R\left(l^2, \overline{\theta}\right)}} \wedge \frac{N}{2}.$$

For N large enough we obtain the bound:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta^*}\left[R\left(l^2, .\right)\right]\right\} \leq R\left(l^2, \overline{\theta}\right) + \sqrt{\frac{dR\left(l^2, \overline{\theta}\right)}{N}}\left[1 + \frac{1}{4}\log\frac{N}{dR\left(l^2, \overline{\theta}\right)}\right] + \frac{d}{4N}.$$

For a given d we obtain the rate of convergence $N^{\frac{-1}{2}}\log N$ except if there is no noise, in this case the rate is $N^{-1}\log N$. Actually, it is a well-known fact in learning theory that if one wants to achieve a better rate of convergence, one has to use relative bounds (bounds that compare the risk of an estimator to a given value of the risk). A detailed explanation is given by Audibert [2]. We are going to give relative bounds later in this work. However, let us insist on the fact that deviations bounds like corollary 2.5 are interesting even if they lead to sub-optimal rate of convergence as they can give better bounds for "small" values of N.

For the general case, we give the following corollary of Theorem 2.9.

Corollary 2.10. *For any $\lambda \in \mathbb{R}_+^*$, for any data-dependent posterior distribution $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$ we have:*

$$P_N\left\{\rho\left[R\left(\psi\wedge\frac{\lambda}{N},.\right)\right]\right\} \leq \frac{1}{\lambda}P_N\left[\mathcal{K}\left(\rho,\pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}\right)\right] - \frac{1}{\lambda}\log\pi\exp\left\{-\lambda R\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}.$$

In particular, if we take:

$$\hat{\rho} = \pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}$$

then we have:

$$P_N\left\{\hat{\rho}\left[R\left(\psi\wedge\frac{\lambda}{N},.\right)\right]\right\} \leq -\frac{1}{\lambda}\log\pi\exp\left\{-\lambda R\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}.$$

Proof. From Theorem 2.9 we have, for any $\lambda \in \mathbb{R}_+^*$, for any data-dependent posterior distribution $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned}
&P_N\left\{\rho\left[R\left(\psi\wedge\frac{\lambda}{N},.\right)\right]\right\} \\
&\qquad\leq P_N\left\{\rho\left\{r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}+\frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\} \\
&\qquad= P_N\left\{\frac{1}{\lambda}\mathcal{K}\left(\rho,\pi_{\exp\left\{-\lambda r\left[\Phi_{\lambda/N}\circ(\psi\wedge(N/\lambda)),.\right]\right\}}\right)\right. \\
&\qquad\qquad\left.-\frac{1}{\lambda}\log\pi\exp\left\{-\lambda r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}\right\}.
\end{aligned}$$

Finally, note that:

$$\begin{aligned}
&P_N\left\{-\log\pi\exp\left\{-\lambda r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}\right\} \\
&\qquad= P_N\inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\lambda\rho r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]+\mathcal{K}(\rho,\pi)\right\} \\
&\qquad\leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)}P_N\left\{\lambda\rho r\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]+\mathcal{K}(\rho,\pi)\right\} \\
&\qquad= \inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\lambda\rho R\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]+\mathcal{K}(\rho,\pi)\right\} \\
&\qquad\qquad= -\log\pi\exp\left\{-\lambda R\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}.
\end{aligned}$$

□

2.6. **Relative bounds.** The idea of relative bounds is to choose a particular parameter $\theta_0 \in \Theta$ and to upper bound the relative risk $R(\psi,\theta) - R(\psi,\theta_0)$ instead of $R(\psi,\theta)$. Usually, θ_0 is chosen as:

$$\theta_0 = \overline{\theta} = \arg\min_{\theta\in\Theta} R(\psi,\theta)$$

and in this case the relative risk is refered as the excess risk, but in this work we are going to adopt a more general point of view.

The idea is to use a prior μ and a posterior ν on the measurable space $(\Theta^2, \mathcal{T}^{\otimes 2})$ and to upper bound:

$$\int_{\Theta^2} [R(\psi,t) - R(\psi,t')]\, \nu[d(t,t')].$$

Definition 2.3. Let μ be a probability measure on the measurable space $(\Theta^2, \mathcal{T}^{\otimes 2})$. Let also (θ, θ') denote the canonical process on $(\Theta^2, \mathcal{T}^{\otimes 2})$. This notation allows us to write, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\nu\left[R(\psi,\theta) - R(\psi,\theta')\right] = \int_{\Theta^2} [R(\psi,t) - R(\psi,t')]\, \nu[d(t,t')]$$

Theorem 2.11. *For any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:*

$$\Phi_{\frac{\lambda}{N}}\left\{\nu \frac{1}{N}\sum_{i=1}^N P_N\left[\Big(\psi_i(\theta) - \psi_i(\theta')\Big) \wedge \frac{N}{\lambda}\right]\right\} \leq \nu\left\{\frac{1}{N}\sum_{i=1}^N \Phi_{\frac{\lambda}{N}}\left[\Big(\psi_i(\theta) - \psi_i(\theta')\Big) \wedge \frac{N}{\lambda}\right]\right\} + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}.$$

For the "moment case", let us choose $s > 1$. Then, for any $\alpha \in (0,1)$, for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\nu\left[R(\psi,\theta) - R(\psi,\theta')\right] \leq \frac{1}{\alpha}\Phi_{\frac{\lambda}{N}}^{-1}\Bigg\{\nu\frac{1}{N}\sum_{i=1}^N \Phi_{\frac{\lambda}{N}}\Bigg[\alpha\psi_i(\theta) - \alpha\psi_i(\theta') - \frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\left|\psi_i(\theta) - \psi_i(\theta')\right|^s\Bigg] + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}\Bigg\} + \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\nu P\Big\{\left[\psi\left(f_\theta(X),Y\right) - \psi\left(f_\theta(X),Y\right)\right]^s\Big\}.$$

Proof. The first assertion is proved in the same way than Theorem 2.3, replacing the measurable parameter space $(\Theta, \mathcal{T})$ by $(\Theta^2, \mathcal{T}^{\otimes 2})$ and the loss $\psi[f_\theta(X), Y]$ by $\psi[f_\theta(X), Y] - \psi[f_{\theta'}(X), Y]$. For the second part, follow the proof of Theorem 2.7. □

Definition 2.4. For the sake of simplicity, we introduce the following notation for the variance term:

$$V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta') = \frac{2N}{\alpha\lambda}\Bigg\{-\Big[r(\psi,\theta) - r(\psi,\theta')\Big] + \frac{1}{\alpha N}\sum_{i=1}^N \Phi_{\frac{\lambda}{N}}\Bigg[\alpha\psi_i(\theta) - \alpha\psi_i(\theta') - \frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\left|\psi_i(\theta) - \psi_i(\theta')\right|^s\Bigg]\Bigg\}.$$

Note that for any $\beta \in \mathbb{R}_+^*$ and any t we have $\Phi_\beta^{-1}(t) \leq t$ and so the moment bound becomes:

$$\nu\left[R(\psi,\theta) - R(\psi,\theta')\right] \leq \nu\left[r(\psi,\theta) - r(\psi,\theta')\right] + \frac{\alpha\lambda}{2N}\nu\Big[V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta')\Big] + \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\nu P\Big\{\left[\psi\left(f_\theta(X),Y\right) - \psi\left(f_{\theta'}(X),Y\right)\right]^s\Big\} + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\alpha\lambda}.$$

Now, we may wonder how to deal with the V term if we want to explicit the bound. Note that we have:

$$\alpha\psi_i(\theta) - \alpha\psi_i(\theta') - \frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi_i(\theta)-\psi_i(\theta')|^s \leq \alpha$$

and so, if we take $\alpha \leq 1/2$, as we have, for any $x \leq 1/2$ and $\lambda \leq N$:

$$x \leq \Phi_{\frac{\lambda}{N}}(x) \leq x + \frac{x^2}{2}$$

we obtain the following result.

Lemma 2.12. *We have, for any $\alpha \leq 1/2$ and $\lambda \in (0, N]$:*

$$V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta') \leq \frac{1}{N}\sum_{i=1}^{N}\left[\psi_i(\theta) - \psi_i(\theta') - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi_i(\theta)-\psi_i(\theta')|^s\right]^2.$$

Let us also remark that it is easy to deduce from Theorem 2.11 the bounds on the excess risk that we mentioned previously. Actually, let us choose $\theta_0 \in \Theta$, $\pi \in \mathcal{M}_+^1(\Theta)$ and $\mu = \pi \otimes \delta_{\theta_0}$. We restrict the theorem to posteriors $\nu = \rho \otimes \delta_{\theta_0}$ for any $\rho \in \mathcal{M}_+^1(\Theta)$. Note that:

$$\mathcal{K}(\nu,\mu) = \mathcal{K}(\rho\otimes\delta_{\theta_0}, \pi\otimes\delta_{\theta_0}) = \mathcal{K}(\rho,\pi) + \mathcal{K}(\delta_{\theta_0},\delta_{\theta_0}) = \mathcal{K}(\rho,\pi).$$

We obtain the following result.

Corollary 2.13. *Let $\theta_0 \in \Theta$. For any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\Phi_{\frac{\lambda}{N}}\left\{\rho\frac{1}{N}\sum_{i=1}^{N}P_N\left[\Big(\psi_i(.)-\psi_i(\theta_0)\Big)\wedge\frac{N}{\lambda}\right]\right\}$$
$$\leq \rho\left\{\frac{1}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\Big(\psi_i(.)-\psi_i(\theta_0)\Big)\wedge\frac{N}{\lambda}\right]\right\} + \frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\lambda}.$$

Moreover, let us choose $s > 1$. Then, for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\rho\left[R\left(\psi,.\right)\right] - R\left(\psi,\theta_0\right) \leq \rho\left[r\left(\psi,.\right)\right] - r\left(\psi,\theta_0\right) + \rho\left[\frac{\alpha\lambda}{2N}V_{\psi,\frac{\lambda}{N},\alpha,s}(.,\theta_0)\right]$$
$$+\frac{\lambda^{s-1}}{N^{s-1}}\left(\frac{(2s-2)^{s-1}}{s^s}\right)\rho P\Big\{\left[\psi\left(f_.(X),Y\right)-\psi\left(f_{\theta_0}(X),Y\right)\right]^s\Big\}$$
$$+\frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\alpha\lambda}.$$

Now, the same problem is raised than in the case of direct (non-relative) bounds: there are non-observable terms in the right-hand side. We can give here analogous of lemma 2.2 and 2.8. First, let us give the following definition.

Definition 2.5. Let us put, for any $(\theta,\theta') \in \Theta^2$, for any $\alpha \in \mathbb{R}$:

$$\Delta_\alpha(\psi,\theta,\theta') = \left[R\left(\psi,\theta\right) - R\left(\psi,\theta'\right)\right] - \frac{1}{N}\sum_{i=1}^{N}P_N\left[\Big(\psi_i(\theta)-\psi_i(\theta')\Big)\wedge\alpha\right].$$

Let us remark that in the case where ψ takes values in $[0,1]$ we have, for any $\lambda \leq N$, for any $(\theta,\theta') \in \Theta^2$:

$$\Delta_{\frac{\lambda}{N}}(\psi,\theta,\theta') = 0.$$

For the moment hypothesis, we propose the following.

Definition 2.6. Let us put:

$$M_{\psi,s} : \Theta^2 \to \mathbb{R}$$
$$(\theta,\theta') \mapsto \frac{1}{N}\sum_{i=1}^{N} P_N\left[\left|\psi_i(\theta)-\psi_i(\theta')\right|^s\right].$$

Note that $M_{\psi,s}(\theta,\theta')$ should be seen as the theoretical counterpart of the empirical quantity $V_{\psi,\lambda/N,\alpha,s}(\theta,\theta')$ (Definition 2.4 page 39). Remark that in the i. i. d. case,

$$M_{\psi,s}(\theta,\theta') = P\Big\{\left[\psi\left(f_\theta(X),Y\right)-\psi\left(f_\theta(X),Y\right)\right]^s\Big\}.$$

Let us assume (until the end of this subsection) that Θ is a normed space, with norm $\|.\|_\Theta$. In some cases, we can assume that π-almost surely:

$$M_{\psi,s}(\theta,\theta') \le C(s)\|\theta-\theta'\|_\Theta^s,$$

where $C(s)$ is known to the statistician.

Example 2.5 (Non-linear regression estimation)**.** Let us assume that we are in the i. i. d. case. In this example we take $\psi = l^2$. We assume Θ is actually an Hilbert space with scalar product $\langle .,. \rangle_\Theta$, that $P(|Y|^s) \le m_s$, that $f_\theta(.) = F(\langle \theta, \Psi(.)\rangle_\Theta)$ with $\Psi : \mathcal{X} \to \Theta$ and $F : \mathbb{R} \to [0,1]$, that F is derivable, that F' is continuous and $\|F'\|_\infty \le D$, and finally that Ψ is such that for any $x \in \mathcal{X}$, $\|\Psi(x)\|_\Theta \le K$. Then we have:

$$M_{\psi,s}(\theta,\theta') \le 2^{2s-1}\left(1+m_s\right)K^sD^s\|\theta-\theta'\|_\Theta^s.$$

Example 2.6 (Linear regression estimation)**.** Here, we still assume that we are in the i. i. d. case, that $\psi = l^2$, that Θ is an Hilbert space with scalar product $\langle .,. \rangle_\Theta$, that $P(|Y|^s) \le m_s$ and we take $f_\theta(.) = \langle \theta, \Psi(.)\rangle_\Theta$ with $\Psi : \mathcal{X} \to \Theta$ and for any $x \in \mathcal{X}$, $\|\Psi(x)\|_\Theta \le K$. We also assume that $\pi(\{\theta \in \Theta, \|\theta\|_\Theta \le \kappa\}) = 1$. Then we have:

$$M_{\psi,s}(\theta,\theta') \le 2^{2s-1}K^s\left(m_s+\kappa^sK^s\right)\|\theta-\theta'\|_\Theta^s.$$

Moreover, in the case where $\theta_0 = \overline{\theta}$, the upper-bound is not observable because of the presence of θ_0. We propose the following trivial inequality:

$$\rho\left\{\frac{1}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\left(\psi_i(\theta)-\psi_i(\theta_0)\right)\wedge\frac{N}{\lambda}\right]\right\}$$
$$\le \sup_{\theta\in\Theta}\rho\left\{\frac{1}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\left(\psi_i(\theta)-\psi_i(\theta)\right)\wedge\frac{N}{\lambda}\right]\right\}.$$

Using Jensen's inequality, we obtain the following result in the bounded case.

Corollary 2.14. *Let us assume that ψ takes values in $[0,C]$ for any $C \ge 1$. Let us put:*

$$\hat{\theta} = \arg\min_{\theta\in\Theta} r(\psi,\theta).$$

Let $\theta_0 \in \Theta$. For any $\varepsilon > 0$, for any $\lambda \in [0, N/(2C)]$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\Phi_{\frac{\lambda}{N}}\left\{\rho\left[R(\psi,.)\right]-R(\psi,\theta_0)\right\}$$
$$\le \rho\left\{\frac{1}{2N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[2\left(\psi_i(.)-\psi_i(\hat{\theta})\right)\right]\right\}$$
$$+\sup_{\theta\in\Theta}\left\{\frac{1}{2N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[2\left(\psi_i(\hat{\theta})-\psi_i(\theta)\right)\right]\right\}+\frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\lambda}.$$

Note that here the right-hand side is fully observable.

Example 2.7 (Bounded linear least square regression, Example 2.4 continued)**.** We use the notations and context of Example 2.4, remember that in this case:

$$\psi_i(\theta) = (Y_i - \langle X_i, \theta \rangle)^2 \in (0,1).$$

Also remember that we used the notation $\tilde{\theta}$ instead of $\hat{\theta}$. Then we have, for any $\lambda < N/2$:

$$\begin{aligned}\frac{1}{2N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[2\left(\psi_i(\tilde{\theta}) - \psi_i(\theta)\right)\right] &\leq r(\psi,\tilde{\theta}) - r(\psi,\theta) + \frac{\lambda}{N^2}\sum_{i=1}^{N}\left[\psi_i(\tilde{\theta}) - \psi_i(\theta)\right]^2 \\ &\leq r(\psi,\tilde{\theta}) - r(\psi,\theta) + \frac{4\lambda}{N^2}\sum_{i=1}^{N}\left\langle X_i, \tilde{\theta} - \theta\right\rangle^2 \\ &\leq r(\psi,\tilde{\theta}) - r(\psi,\theta) + \frac{4\lambda}{N}\left[r(\psi,\theta) - r(\psi,\tilde{\theta})\right].\end{aligned}$$

So, as soon as $\lambda < N/4$ we have:

$$\begin{aligned}&\sup_{\theta\in\Theta}\left\{\frac{1}{2N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[2\left(\psi_i(\tilde{\theta}) - \psi_i(\theta)\right)\right]\right\} \\ &\qquad\leq \sup_{\theta\in\Theta}\left(1 - \frac{\lambda}{4N}\right)\left[r(\psi,\tilde{\theta}) - r(\psi,\theta)\right] \\ &\qquad= \left(1 - \frac{\lambda}{4N}\right)\left[r(\psi,\tilde{\theta}) - \inf_{\theta\in\Theta} r(\psi,\theta)\right] = 0.\end{aligned}$$

So the corollary becomes in this case, for any $\theta_0 \in \Theta$, for any $\varepsilon > 0$, for any $\lambda \in [0, N/4]$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned}&\Phi_{\frac{\lambda}{N}}\left\{\rho\left[R\left(\psi,.\right)\right] - R(\psi,\theta_0)\right\} \\ &\qquad\leq \rho\left\{\frac{1}{2N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[2\left(\psi_i(.) - \psi_i(\tilde{\theta})\right)\right]\right\} + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}.\end{aligned}$$

2.7. **Relative bounds on the integrated risk.** In this subsection, we give the integrated version of the relative bounds (corollary 2.13), and try to examine its consequences in terms of rate of convergence.

Theorem 2.15. *For any $\lambda \in \mathbb{R}_+^*$, for any data-dependent posterior distribution $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$ we have:*

$$\begin{aligned}&P_N\left\{\rho\frac{1}{N}\sum_{i=1}^{N} P_N\left[\left(\psi_i(.) - \psi_i(\overline{\theta})\right) \wedge \frac{N}{\lambda}\right]\right\} \\ &\qquad\leq P_N\left\{\rho\left\{\frac{1}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\left(\psi_i(.) - \psi_i(\overline{\theta})\right) \wedge \frac{N}{\lambda}\right]\right\} + \frac{\mathcal{K}(\rho,\pi)}{\lambda}\right\}.\end{aligned}$$

Let us examine some consequences of Theorem 2.15 on the toy example we already studied.

Example 2.8 (Bounded linear least square regression, Example 2.4 continued)**.** We take the notations and the context of Example 2.4. Let us remember that in particular, we are in the i. i. d. case, with $P_N = P^{\otimes N}$. The theorem becomes, for any $\lambda \in [1, N/2]$, for function $\delta : \mathcal{Z}^N \to [0, 1/\sqrt{2}]$ we have:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq P^{\otimes N}\Bigg\{\tilde{\rho}^{\delta}\left[r\left(l^2,.\right)\right]-r\left(l^2,\overline{\theta}\right)$$
$$+\tilde{\rho}^{\delta}\left\{\frac{\lambda}{2N^2}\sum_{i=1}^{N}\left[\left(Y_i-\langle .,X_i\rangle\right)^2-\left(Y_i-\left\langle\overline{\theta},X_i\right\rangle\right)^2\right]^2\right\}+\frac{d}{\lambda}\log\frac{1}{\delta\sqrt{2}}\Bigg\}.$$

Now, note that:

$$\tilde{\rho}^{\delta}\left\{\frac{\lambda}{2N^2}\sum_{i=1}^{N}\left[\left(Y_i-\langle .,X_i\rangle\right)^2-\left(Y_i-\left\langle\overline{\theta},X_i\right\rangle\right)^2\right]^2\right\}$$
$$=\tilde{\rho}^{\delta}\left\{\frac{\lambda}{2N^2}\sum_{i=1}^{N}\left[2Y_i-\left\langle .+\overline{\theta},X_i\right\rangle\right]^2\left\langle .-\overline{\theta},X_i\right\rangle^2\right\}$$
$$\leq\tilde{\rho}^{\delta}\left[\frac{2\lambda}{N^2}\sum_{i=1}^{N}\left\langle .-\overline{\theta},X_i\right\rangle^2\right]$$

and so, for $\lambda\leq N/2$:

$$\tilde{\rho}^{\delta}\left[r\left(l^2,.\right)-r\left(l^2,\overline{\theta}\right)+\frac{2\lambda}{N^2}\sum_{i=1}^{N}\left\langle .-\overline{\theta},X_i\right\rangle^2\right]\leq 2\tilde{\rho}^{\delta}\frac{1}{N}\sum_{i=1}^{N}\left\langle .-\hat{\theta},X_i\right\rangle^2$$
$$\leq 2\delta^2+2\frac{1}{N}\sum_{i=1}^{N}\left\langle\tilde{\theta}-\hat{\theta},X_i\right\rangle^2\leq 2\delta^2+2\frac{1}{N}\sum_{i=1}^{N}\left\langle\overline{\theta}-\hat{\theta},X_i\right\rangle^2,$$

and note that usual computations on the linear model lead to:

$$P\left[\frac{1}{N}\sum_{i=1}^{N}\left\langle\overline{\theta}-\hat{\theta},X_i\right\rangle^2\right]\leq\frac{\sigma^2 d}{N},$$

where:

$$\sigma^2=\sup_{x\in\mathcal{X}}P\left[\left(Y_i-\left\langle\overline{\theta},X_i\right\rangle\right)^2\middle|X_i=x\right].$$

So, we put $\lambda=\frac{N}{2}$ and we obtain:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq\delta^2+\frac{2\sigma^2 d}{N}+\frac{2d}{N}\log\frac{1}{\delta\sqrt{2}}.$$

We choose $\delta=\sqrt{\frac{2d}{N}}$ and we obtain:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq\frac{2}{N}\left[d\left(1+\sigma^2\right)+\frac{1}{2}\log\frac{N}{d4\sqrt{2}}\right].$$

Note that we obtain a bound in $N^{-1}\log N$ (for a given d). This is not completely satisfying as in this case, we would like to achieve a bound in N^{-1}. Actually, the same problem would arise if we choose the optimal Gibbs posterior for ρ. The technique to remove the log term was suggested by Catoni [9, 10, 11] (it is called "localization" in [10] and [11]). The idea is to replace π by:

$$\pi_{\exp[-\beta R(l^2,.)]}$$

for some $\beta>0$. The reason for the choice of this particular prior distribution is discussed later (in the next subsection). The theorem becomes, for any $\beta\in\mathbb{R}_+^*$, $\lambda\in[1,N/2]$ and δ:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq P^{\otimes N}\Bigg\{\tilde{\rho}^{\delta}\left[r\left(l^2,.\right)\right]-r\left(l^2,\overline{\theta}\right)$$

$$+\tilde{\rho}^{\delta}\left\{\frac{\lambda}{2N^2}\sum_{i=1}^{N}\left[\left(Y_i-\langle .,X_i\rangle\right)^2-\left(Y_i-\left\langle\overline{\theta},X_i\right\rangle\right)^2\right]^2\right\}$$
$$+\frac{1}{\lambda}\mathcal{K}\left(\tilde{\rho}^{\delta},\pi_{\exp[-\beta R(l^2,.)]}\right)\Bigg\}.$$

Now, note that:

$$\mathcal{K}\left(\tilde{\rho}^{\delta},\pi_{\exp[-\beta R(l^2,.)]}\right)=d\log\frac{1}{\delta\sqrt{2}}+\beta\tilde{\rho}^{\delta}\left[R(l^2,.)\right]+\log\pi\exp\left[-\beta R(l^2,.)\right]$$
$$\leq d\log\frac{1}{\delta\sqrt{2}}+\frac{\beta\delta^2}{2}+\log\frac{\lambda_d\left\{\exp\left[-\beta R(l^2,.)+\beta R(l^2,\overline{\theta})\right]\right\}}{\lambda_d\left[B_d\left(0,1/\sqrt{2}\right)\right]}$$
$$+\beta\left\{\tilde{\rho}^{\delta}\left[R(l^2,.)\right]-R(l^2,\overline{\theta})\right\}$$

and so:

$$\mathcal{K}\left(\tilde{\rho}^{\delta},\pi_{\exp[-\beta R(l^2,.)]}\right)-\beta\left\{\tilde{\rho}^{\delta}\left[R(l^2,.)\right]-R(l^2,\overline{\theta})\right\}$$
$$\leq d\log\frac{1}{\delta\sqrt{2}}+\frac{\beta\delta^2}{2}+\log\frac{\Gamma\left(\frac{d}{2}+1\right)2^{\frac{d}{2}}\pi^{\frac{d}{2}}}{\pi^{\frac{d}{2}}\beta^{\frac{d}{2}}\sqrt{\det M}}$$
$$=d\log\frac{1}{\delta\sqrt{\beta}}+\frac{\beta\delta^2}{2}+\log\frac{\Gamma\left(\frac{d}{2}+1\right)}{\sqrt{\det M}}$$

where M is the variance-covariance matrix of X under P. This leads to:

$$\left(1-\frac{\beta}{\lambda}\right)P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq\delta^2\left(1+\frac{\beta}{2\lambda}\right)+\frac{2\sigma^2 d}{N}$$
$$+\frac{d}{\lambda}\log\frac{1}{\delta\sqrt{\beta}}+\frac{1}{\lambda}\log\frac{\Gamma\left(\frac{d}{2}+1\right)}{\sqrt{\det M}},$$

or equivalentely:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq\frac{1}{1-\frac{\beta}{\lambda}}\Bigg\{\delta^2\left(1+\frac{\beta}{2\lambda}\right)+\frac{2\sigma^2 d}{N}$$
$$+\frac{d}{\lambda}\log\frac{1}{\delta\sqrt{\beta}}+\frac{1}{\lambda}\log\frac{\Gamma\left(\frac{d}{2}+1\right)}{\sqrt{\det M}}\Bigg\}.$$

Now, we take the (suboptimal) values $\lambda=N/2$, $\delta=\sqrt{2d/N}$ and $\beta=d/(2\delta^2)=N/4$, and we use the following bound (see [43] for example):

$$\log\Gamma\left(\frac{d}{2}\right)\leq-\frac{1}{2}\log\frac{d}{2}+\frac{d}{2}\log\frac{d}{2}-\frac{d}{2}+\frac{1}{2}\log 2\pi+\frac{1}{6d}$$

and so:

$$\log\Gamma\left(\frac{d}{2}+1\right)\leq\frac{1}{2}\log\frac{d}{2}+\frac{d}{2}\log\frac{d}{2}-\frac{d}{2}+\frac{1}{2}\log 2\pi+\frac{1}{6d}$$

and we obtain:

$$P^{\otimes N}\left\{\tilde{\rho}^{\delta}\left[R\left(l^2,.\right)\right]-R\left(l^2,\overline{\theta}\right)\right\}\leq\frac{d}{N}\left(5+4\sigma^2\right)+\frac{4}{N}\left[\frac{1}{3d}-d+\log\frac{\pi d}{\det M}\right].$$

So, from now, we will try, as soon as possible, to give localized bounds. However, in deviation inequalities, the idea to replace π by the Gibbs distribution $\pi_{\exp[-\beta R(l^2,.)]}$ leads to non-observable bounds, whereas the estimation methods we propose often require an explicit computation of the bound. We will see that some techniques allow to obtain observable localized bounds.

2.8. **Relative bounds with respect to a Gibbs distribution.** The computations in the preceding example (Example 2.8) showed that a way to improve the bounds is clearly to work on the divergence term $\mathcal{K}(\rho,\pi)$, especially to try to make a good choice for the prior π. Remember that, once π is chosen, we know that the optimal ρ is a Gibbs distribution: $\pi_{\exp[-\beta r(\psi,.)]}$ for some $\beta \in \mathbb{R}_+^*$.

Let us remark that, for any $(\rho,\pi) \in \mathcal{M}_+^1(\Theta)$ we have:

$$P\Big[\mathcal{K}(\rho,\pi)\Big] = P\Big[\mathcal{K}(\rho,P(\rho))\Big] + \mathcal{K}(P(\rho),\pi). \tag{2.2}$$

This implies that, for a given data-dependent ρ, the optimal deterministic measure π is $P(\rho)$ in the sense that it minimizes the expectation of $\mathcal{K}(\rho,\pi)$ (left-hand side of Equation 2.2), making it equal to the expectation of $\mathcal{K}(\rho,P(\rho))$. This last quantity is the mutual information between the estimator and the sample.

So, for $\rho = \pi_{\exp[-\beta r(\psi,.)]}$, this is an incitation to replace the prior π with $P\left(\pi_{\exp[-\beta r(\psi,.)]}\right)$. It is then natural to approximate this distribution by $\pi_{\exp[-\beta R(\psi,.)]}$.

Note that this choice is exactly the one we made in Example 2.8, and that in this particular case it led to an improvement of the rate of convergence, by a $\log N$ term. So we can do what we proposed in this example (to try to localize systematically our bounds) by replacing π with $\pi_{\exp[-\beta R(\psi,.)]}$ for some β in every bound.

But then the following problem emerges: a term $\mathcal{K}(\rho,\pi_{\exp[-\beta R(\psi,.)]})$ appears in the bounds, and this term is not observable, so our willing to obtain empirical bounds can be fulfilled only if we are able to upper bound $\mathcal{K}(\rho,\pi_{\exp[-\beta R(\psi,.)]})$ by an empirical term. Actually, we have the following, for any $(\rho,\pi) \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned}&\mathcal{K}\left(\rho,\pi_{\exp[-\beta R(\psi,.)]}\right)\\ &\qquad = \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]},\pi\right) + \beta\left[\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)\right].\end{aligned}$$

So, the problem of obtaining an empirical bound on $\mathcal{K}(\rho,\pi_{\exp[-\beta R(\psi,.)]})$ is linked with the of getting an empirical bound for:

$$\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.).$$

In this subsection and in the next one, we focus on such bounds. More particularly, the next result (Theorem 2.16) is an empirical bound on:

$$\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)$$

for any $\beta \in \mathbb{R}_+^*$. The function $\beta \mapsto \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)$ being nonincreasing, a bound under the form:

$$\rho R(\psi,.) \leq \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.) + \text{empirical} \quad \text{terms}$$

gives a kind of scale to choose a posterior distribution ρ. An algorithm based on this idea is given after this main result. Finally the next subsection includes empirical bounds on $\mathcal{K}(\rho,\pi_{\exp[-\beta R(\psi,.)]})$.

We will need the following notation.

Definition 2.7. For the sake of shortness, let us put, for any $(\theta,\theta') \in \Theta^2$:

$$v_{\psi,\frac{\lambda}{N}}(\theta,\theta') = \frac{2N}{\lambda}\left\{\Phi_{\frac{\lambda}{N}}\left[r(\psi,\theta) - r(\psi,\theta')\right] - \left[r(\psi,\theta) - r(\psi,\theta')\right]\right\}.$$

Note that the variance term $v_{\psi,\lambda/N}(\theta,\theta')$ is the analogous of $V_{\psi,\lambda/N,\alpha,s}(\theta,\theta')$ introduced in Definition 2.4 page 39 in the bounded case.

Theorem 2.16. *Let us assume that for any $\theta \in \Theta$, $P\{\psi[f_\theta(X),Y] \leq 1\} = 1$. Then we have, for any $\varepsilon > 0$, for any $(\lambda,\beta,\gamma) \in (0,N)^3$ such that $\beta < \gamma < \lambda$, with P_N-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\left(\lambda\Phi_{\frac{\lambda}{N}} - \beta Id\right)\left[\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)\right]$$

$$\leq (\lambda-\gamma)\left[\rho r(\psi,.)-\pi_{\exp[-\gamma r(\psi,.)]}r(\psi,.)\right]$$
$$+\frac{\beta(\lambda-\gamma)}{\lambda(\gamma-\beta)}\log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\pi_{\exp[-\gamma r(\psi,d\theta)]}v_{\psi,\frac{\lambda}{N}}(\theta,\theta')\right]$$
$$+\log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\rho_{(d\theta)}v_{\psi,\frac{\lambda}{N}}(\theta,\theta')\right]+\mathcal{K}\left(\rho,\pi_{\exp[-\gamma r(\psi,.)]}\right)+\log\frac{1}{\varepsilon}.$$

This result can be interpreted in the same terms as the simplest PAC-Bayesian theorems given in the beginning of this work. The mean error of any estimator ρ when compared to the "almost optimal" Gibbs distribution is controlled by its empirical counterpart together with a variance terms (expressed in terms of $v_{\psi,\lambda/N}$) and a complexity term (the Kullbak divergence). The parameter λ plays exactly the same role as in the basic relative PAC-Bayesian theorem (Theorem 2.11 page 39). The parameter γ is the localization parameter. We have seen in the previous subsection how an appropriate choice for γ can help to remove extra $\log N$ terms in the rate of convergence of our estimators. Finally the parameter β is the parameter of the Gibbs distribution ρ is to be compared with. Note that the function $\beta\mapsto\pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)$ is nonincreasing, and converges in general to $R(\psi,\overline{\theta})$ as $\beta\to+\infty$.

Before we give the proof of this theorem, we propose a simple way to use it to build an estimator. In a first time, in order to optimize the values of λ, β and γ we have to use a union bound argument (as in the first section). Let Λ be a grid of values, we already proposed of $a>1$:

$$\Lambda=\left\{a^l,\quad 0\leq l\leq\left\lfloor\frac{\log N}{\log a}\right\rfloor\right\},$$

and let us put, for any posterior ρ:

$$\mathcal{B}(\rho,\beta)=\frac{1}{\lambda-\beta}\arg\min_{\substack{(\lambda,\gamma)\in\Lambda^2\\ \beta\leq\gamma\leq\lambda}}\left\{(\lambda-\gamma)\left[\rho r(\psi,.)-\pi_{\exp[-\gamma r(\psi,.)]}r(\psi,.)\right]\right.$$
$$+\frac{\beta(\lambda-\gamma)}{\lambda(\gamma-\beta)}\log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\pi_{\exp[-\gamma r(\psi,d\theta)]}v_{\psi,\frac{\lambda}{N}}(\theta,\theta')\right]$$
$$\left.+\log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\rho_{(d\theta)}v_{\psi,\frac{\lambda}{N}}(\theta,\theta')\right]+\mathcal{K}\left(\rho,\pi_{\exp[-\gamma r(\psi,.)]}\right)+\log\frac{|\Lambda|^3}{\varepsilon}\right\}.$$

Now, let us choose for ρ a Gibbs distribution:

$$\rho=\pi_{\exp[-\zeta r(\psi,.)]}$$

and let us put for any $\zeta\in\mathbb{R}_+^*$:

$$B(\zeta,\beta)=\mathcal{B}\left(\pi_{\exp[-\zeta r(\psi,.)]},\beta\right).$$

So we have, with probability at least $1-\varepsilon$, for any $\zeta\in\mathbb{R}_+^*$ and $\beta\in\Lambda$:

$$\pi_{\exp[-\zeta r(\psi,.)]}R(\psi,.)\leq\pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)+B(\zeta,\beta).$$

Note that for any (ζ,β), $B(\zeta,\beta)$ is observable, but we still have to choose the pair (ζ,β). Let us put:

$$\hat{\beta}(\zeta)=\sup\left\{\beta\in\Lambda:\quad B(\zeta,\beta)\leq 0\right\}$$

that leads to the following, with probability at least $1-\varepsilon$, for any ζ:

$$\pi_{\exp[-\zeta r(\psi,.)]}R(\psi,.)\leq\pi_{\exp\left[-\hat{\beta}(\zeta)R(\psi,.)\right]}R(\psi,.).$$

In other words, $\hat{\beta}(\zeta)$ is the largest β such that $\pi_{\exp[-\zeta r(\psi,.)]}$ is "better" (with large probability) than $\pi_{\exp\left[-\hat{\beta}(\zeta)R(\psi,.)\right]}$. So we propose to choose:

$$\hat{\zeta} \in \arg\max_{\zeta\in\lambda} \hat{\beta}(\zeta).$$

We are now going to give the proof of Theorem 2.16. We will need the following lemma (which proof is taken from Catoni [11]).

Lemma 2.17. *For any $\pi \in \mathcal{M}_+^1(\Theta)$ and any measurable functions $h, H : \Theta \mapsto \mathbb{R}$ we have:*

$$\pi_{\exp(-h)}(h) - \pi_{\exp(-H)}(h) \leq \pi_{\exp(-h)}(H) - \pi_{\exp(-H)}(H).$$

Proof. We have:

$$\begin{aligned}
\mathcal{K}\left(\pi_{\exp(-h)}, \pi_{\exp(-H)}\right) &= \pi_{\exp(-h)}(H) + \log \pi \exp(-H) + \mathcal{K}\left(\pi_{\exp(-h)}, \pi\right) \\
&= \pi_{\exp(-h)}(H) - \pi_{\exp(-H)}(H) - \mathcal{K}\left(\pi_{\exp(-H)}, \pi\right) + \mathcal{K}\left(\pi_{\exp(-h)}, \pi\right) \\
&= \pi_{\exp(-h)}(H) - \pi_{\exp(-H)}(H) - \mathcal{K}\left(\pi_{\exp(-H)}, \pi\right) \\
&\qquad - \pi_{\exp(-h)}(h) - \log \pi \exp(-H) \\
&\leq \pi_{\exp(-h)}(H) - \pi_{\exp(-H)}(H) - \mathcal{K}\left(\pi_{\exp(-H)}, \pi\right) \\
&\qquad - \pi_{\exp(-h)}(h) + \pi_{\exp(-H)}(h) + \mathcal{K}\left(\pi_{\exp(-H)}, \pi\right),
\end{aligned}$$

and so:

$$\begin{aligned}
\pi_{\exp(-h)}(H) - \pi_{\exp(-H)}(H) &\geq \pi_{\exp(-h)}(h) + \pi_{\exp(-H)}(h) \\
&\quad + \mathcal{K}\left(\pi_{\exp(-h)}, \pi_{\exp(-H)}\right) \geq \pi_{\exp(-h)}(h) + \pi_{\exp(-H)}(h),
\end{aligned}$$

that is the conclusion. □

We are now ready to give the proof of the theorem.

Proof of Theorem 2.16. Let us apply Theorem 2.11 with:

$$\mu = \pi_{\exp[-\beta R(\psi,.)]} \otimes \pi_{\exp[-\beta R(\psi,.)]}$$

and:

$$\nu = m \otimes \pi_{\exp[-\beta R(\psi,.)]},$$

where $m \in \mathcal{M}_+^1(\Theta)$. We obtain, with P_N-probability at least $1-\varepsilon$, for any $m \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned}
(2.3)\quad \lambda\left[mR(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)\right] &\leq \lambda\left[mr(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}r(\psi,.)\right] \\
&\quad + \frac{\lambda^2}{2N}\rho \otimes \pi_{\exp[-\beta R(\psi,.)]} v_\psi(.,.) + \mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) + \log\frac{1}{\varepsilon}.
\end{aligned}$$

In a first time, we apply inequality 2.3 with $m = \rho$ and we remark that:

$$\begin{aligned}
&\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) \\
&\qquad = \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right) + \beta\left[\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)\right].
\end{aligned}$$

This leads to:

$$\begin{aligned}
(2.4)\quad &\left(\lambda\Phi_{\frac{\lambda}{N}} - \beta Id\right)\left[\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}R(\psi,.)\right] \\
&\leq \lambda\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}r(\psi,.)\right] + \frac{\lambda^2}{2N}\rho_{(d\theta)} \otimes \pi_{\exp[-\beta R(\psi,d\theta')]} v_\psi(\theta,\theta') \\
&\qquad + \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right) + \log\frac{1}{\varepsilon} \\
&= \gamma\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]}r(\psi,.)\right] + \frac{\lambda^2}{2N}\rho_{(d\theta)} \otimes \pi_{\exp[-\beta R(\psi,d\theta')]} v_\psi(\theta,\theta')
\end{aligned}$$

$$+ \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]},\pi\right) + \log\frac{1}{\varepsilon} + (\lambda-\gamma)\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right].$$

First of all, let us remark that:

$$\begin{aligned}(2.5)\quad & \gamma\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] + \frac{\lambda^2}{2N}\rho_{(d\theta)}\otimes\pi_{\exp[-\beta R(\psi,d\theta')]} v_\psi(\theta,\theta') \\ & \qquad - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]},\pi\right) \\ & \leq \sup_{m\in\mathcal{M}_+^1(\Theta)}\left\{\gamma\left[\rho r(\psi,.) - m r(\psi,.)\right] + \frac{\lambda^2}{2N}\rho_{(d\theta)}\otimes m_{d\theta'} v_\psi(\theta,\theta') - \mathcal{K}(m,\pi)\right\} \\ & \qquad = \log\pi_{(d\theta')}\exp\left[-\gamma r(\psi,.) + \frac{\lambda^2}{2N}\rho_{(d\theta)} v_\psi(\theta,\theta')\right] + \gamma\rho r(\psi,.) \\ & \qquad = \log\pi_{(d\theta')}\exp\left[-\gamma r(\psi,.) + \frac{\lambda^2}{2N}\rho_{(d\theta)} v_\psi(\theta,\theta')\right] + \mathcal{K}\left(\rho,\pi_{\exp[-\gamma r(\psi,.)]}\right) \\ & \qquad\qquad - \mathcal{K}(\rho,\pi) - \log\pi\exp[-\gamma r(\psi,.)] \\ & \qquad = \log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\rho_{(d\theta)} v_\psi(\theta,\theta')\right] + \mathcal{K}\left(\rho,\pi_{\exp[-\gamma r(\psi,.)]}\right) - \mathcal{K}(\rho,\pi).\end{aligned}$$

For the other term of Inequation 2.4 let us remark that:

$$\begin{aligned}(2.6)\quad & (\lambda-\gamma)\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] \\ & \qquad = (\lambda-\gamma)\left[\rho r(\psi,.) - \pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.)\right] \\ & \qquad\qquad + (\lambda-\gamma)\left[\pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right].\end{aligned}$$

By lemma 2.17 we have:

$$\begin{aligned}& (\lambda-\gamma)\left[\pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] \\ & \qquad \leq \frac{\beta(\lambda-\gamma)}{\gamma}\left[\pi_{\exp[-\gamma r(\psi,.)]} R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} R(\psi,.)\right]\end{aligned}$$

and, if we apply inequality 2.3 with $m = \pi_{\exp[-\beta r(\psi,.)]}$ we obtain:

$$\begin{aligned}& \lambda\left[\pi_{\exp[-\gamma r(\psi,.)]} R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} R(\psi,.)\right] \\ & \qquad \leq \lambda\left[\pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] \\ & + \frac{\lambda^2}{2N}\pi_{\exp[-\beta r(\psi,.)]}\otimes\pi_{\exp[-\beta R(\psi,.)]} v_\psi(.,.) + \mathcal{K}\left(\pi_{\exp[-\beta r(\psi,.)]},\pi_{\exp[-\beta R(\psi,.)]}\right) + \log\frac{1}{\varepsilon}\end{aligned}$$

and so, using exactly the same method that we used into inequalities 2.4 and 2.5 we obtain:

$$\begin{aligned}& (\lambda-\beta)\left[\pi_{\exp[-\gamma r(\psi,.)]} R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} R(\psi,.)\right] \\ & \qquad \leq (\lambda-\gamma)\left[\pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] \\ & \qquad\qquad + \log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\pi_{\exp[-\gamma r(\psi,d\theta)]} v_\psi(\theta,\theta')\right] + \log\frac{1}{\varepsilon}.\end{aligned}$$

Plugging this result into inequality 2.6 leads to:

$$\begin{aligned}& (\lambda-\gamma)\left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.)\right] \\ & \qquad \leq (\lambda-\gamma)\left[\rho r(\psi,.) - \pi_{\exp[-\gamma r(\psi,.)]} r(\psi,.)\right] \\ & \quad + \frac{\beta(\lambda-\gamma)}{\lambda(\gamma-\beta)}\left\{\log\pi_{\exp[-\gamma r(\psi,d\theta')]}\exp\left[\frac{\lambda^2}{2N}\pi_{\exp[-\gamma r(\psi,d\theta)]} v_\psi(\theta,\theta')\right] + \log\frac{1}{\varepsilon}\right\}.\end{aligned}$$

Plugging this last result with inequality 2.5 into inequality 2.4 leads to the announced result. □

2.9. **Comparison of two posterior distributions and model selection.** The idea of this subsection is to give a bound on:

$$\rho^1 R(\psi,.) - \rho^2 R(\psi,.),$$

where both ρ^1 and ρ^2 are observable, allowing a choice between two posteriors distribution (and so between two randomized estimators). Note that if we have submodels of Θ, $\Theta_1 \subset \Theta$ and $\Theta_2 \subset \Theta$ and $\rho^1(\Theta_1) = \rho^2(\Theta_2) = 1$ then the result also allows to perform model selection, we give a detailed procedure later (the reason why we prefer to consider two priors on two different models instead of a unique prior on the union of both models is detailed by Catoni [11]: the localization of a prior on a reunion of models usually doesn't lead to optimal rates of convergence). The result given here (and its proof) is an adaptation of results previously obtained by Audibert [2] and Catoni [11] in the context of classification.

Here again, we are going to give localized bounds. As we have seen in the previous subsection, replacing π^i by $\pi^i_{\exp[-\beta_i R(\psi,.)]}$ raises the problem of upper bounding the divergence $\mathcal{K}(\rho^i, \pi^i_{\exp[-\beta_i R(\psi,.)]})$, $i = 1, 2$, where we have chosen two prior distributions $\pi^i \in \mathcal{M}^1_+(\Theta)$, $i = 1, 2$. Let us start with some empirical bounds for these terms.

Theorem 2.18. *Let us choose a prior distribution $\pi \in \mathcal{M}^1_+(\Theta)$. Let us assume that for any $\theta \in \Theta$, $P\{\psi[f_\theta(X), Y] \leq 1\} = 1$. For any $\varepsilon > 0$, for any $(\gamma, \beta) \in (O, N)^2$ such that $\beta < \gamma$, with P_N-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}^1_+(\Theta)$,*

$$\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) \leq \left(1 - \frac{\beta}{\gamma}\right)^{-1} \left\{ \mathcal{K}\left(\rho, \pi_{\exp[-\beta r(\psi,.)]}\right) - \frac{\beta}{\gamma} \log \varepsilon + \log \left[\pi_{\exp[-\beta r(\psi, d\theta')]} \exp\left(\frac{\beta\gamma}{2N} \rho_{(d\theta)} v_\psi(\theta, \theta') \right) \right] \right\}.$$

Note that the "theoretical" entropy term is controlled by its empirical counterpart together with a variance term.

Proof. Note that:

$$\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) = \beta \left(\rho R(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} R(\psi,.) \right) + \mathcal{K}(\rho, \pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right).$$

Let us apply Theorem 2.11 with:

$$\mu = \pi_{\exp[-\beta R(\psi,.)]} \otimes \pi_{\exp[-\beta R(\psi,.)]}$$

and $\nu = \rho \otimes \pi_{\exp[-\beta R(\psi,.)]}$ to obtain with probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}^1_+(\Theta)$:

$$\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) \leq \beta \left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.) + \frac{\gamma}{2N} \rho \otimes \pi_{\exp[-\beta R(\psi,.)]} v_\psi(.,.) + \frac{\log \frac{1}{\varepsilon} + \mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right)}{\gamma} \right] + \mathcal{K}(\rho, \pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right).$$

Replacing in the right-hand side of this inequality $\pi_{\exp[-\beta R(\psi,.)]}$ with a supremum over all possible distributions leads to the announced result. □

We will need also a variant of this theorem using a moment bound.

Theorem 2.19. *Let us choose $s > 1$, $\alpha \in (0,1)$ and a prior distribution $\pi \in \mathcal{M}_+^1(\Theta)$. Let us assume that for some known constant $\mathcal{D}_s$ and for some measurable function of the observations $D = D(Z_1, ..., Z_N$ independent of the parameter θ,*

$$\sup_{\theta,\theta' \in \Theta} M_{\psi,s}(\theta,\theta') \leq \mathcal{D}_s,$$

$$\sup_{\theta,\theta' \in \Theta} V_{\psi,\frac{\gamma}{N},\alpha,s}(\theta,\theta') \leq D.$$

For any $\varepsilon > 0$, $(\gamma,\beta) \in (0,N)^2$ such that $\beta < \alpha\gamma$,

$$\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) \leq \left(1 - \frac{\beta}{\alpha\gamma}\right)^{-1} \left\{ \mathcal{K}\left(\rho, \pi_{\exp[-\beta r(\psi,.)]}\right) - \frac{\beta}{\alpha\gamma} \log \varepsilon + \log\left[\pi_{\exp[-\beta r(\psi,d\theta')]} \exp\left(\frac{\alpha\beta\gamma}{2N} \rho_{(d\theta)} V_{\psi,\frac{\gamma}{N},\alpha,s}(\theta,\theta') + \frac{\beta}{s}\left(\frac{(s-1)\gamma}{sN}\right)^{s-1} \rho_{(d\theta)} M_{\psi,s}(\theta,\theta') \right) \right] \right\}.$$

Proof. Let us apply the moment bound in Theorem 2.11 with:

$$\mu = \pi_{\exp[-\beta R(\psi,.)]} \otimes \pi_{\exp[-\beta R(\psi,.)]}$$

and $\nu = \rho \otimes \pi_{\exp[-\beta R(\psi,.)]}$ to obtain:

$$\begin{aligned}
\mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right) &\leq \frac{\beta}{\alpha} \Phi_{\frac{\gamma}{N}}^{-1} \left\{ \alpha \left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.) + \frac{\alpha\gamma}{2N} \rho \otimes \pi_{\exp[-\beta R(\psi,.)]} V_{\psi,\frac{\gamma}{N},\alpha,s}(.,.) \right] + \frac{\log\frac{1}{\varepsilon} + \mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right)}{\gamma} \right\} \\
&\quad + \frac{\beta}{s}\left(\frac{(s-1)\gamma}{sN}\right)^{s-1} \rho_{(d\theta)} \otimes \pi_{\exp[-\beta R(\psi,.)]} M_{\psi,s}(\theta,\theta') \\
&\quad + \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right) \\
&\leq \beta \left[\rho r(\psi,.) - \pi_{\exp[-\beta R(\psi,.)]} r(\psi,.) + \frac{\alpha\gamma}{2N} \rho \otimes \pi_{\exp[-\beta R(\psi,.)]} V_{\psi,\frac{\gamma}{N},\alpha,s}(.,.) + \frac{\log\frac{1}{\varepsilon} + \mathcal{K}\left(\rho, \pi_{\exp[-\beta R(\psi,.)]}\right)}{\alpha\gamma} \right] \\
&\quad + \frac{\beta}{s}\left(\frac{(s-1)\gamma}{sN}\right)^{s-1} \rho_{(d\theta)} \otimes \pi_{\exp[-\beta R(\psi,.)]} M_{\psi,s}(\theta,\theta') \\
&\quad + \mathcal{K}(\rho,\pi) - \mathcal{K}\left(\pi_{\exp[-\beta R(\psi,.)]}, \pi\right).
\end{aligned}$$

Replacing in the right-hand side of this inequality the local prior $\pi_{\exp[-\beta R(\psi,.)]}$ with a supremum over all possible posterior distributions proves the announced result. □

Let us procede now to the comparison of two posteriors ρ^1 and ρ^2.

Theorem 2.20. *Let us choose two prior distributions π^1 and π^2 in $\mathcal{M}_+^1(\Theta)$. Let us assume that for any $\theta \in \Theta$, $P\{\psi[f_\theta(X), Y] \leq 1\} = 1$. Then we have, for any $\varepsilon > 0$, for any $(\lambda, \gamma_1, \gamma_2, \beta_1, \beta_2) \in (0,N)^5$ such that $\beta_1 < \gamma_1$ and $\beta_2 < \gamma_2$, with P_N-probability at least $1 - \varepsilon$, for any $(\rho^1, \rho^2) \in \left[\mathcal{M}_+^1(\Theta)\right]^2$:*

$$\Phi_{\frac{\lambda}{N}}\left[\rho^1 R(\psi,.) - \rho^2 R(\psi,.)\right] \leq \rho^1 r(\psi,.) - \rho^2 r(\psi,.) + \frac{\lambda}{2N} \rho^1 \otimes \rho^2 v_\psi(.,.)$$

$$+\frac{(\gamma_1\gamma_2-\beta_1\beta_2)\log\frac{3}{\varepsilon}}{\lambda(\gamma_1-\beta_1)(\gamma_2-\beta_2)}+\sum_{i=1}^{2}\frac{\gamma_i}{\lambda(\gamma_i-\beta_i)}\left\{\mathcal{K}\left(\rho^i,\pi^i_{\exp[-\beta_i r(\psi,.)]}\right)\right.$$
$$\left.+\log\pi^i_{\exp[-\beta_i r(\psi,d\theta')]}\exp\left[\frac{\beta_i\gamma_i}{2N}\rho^i_{(d\theta)}v_\psi(\theta,\theta')\right]\right\}.$$

Remark 2.3. Before the proof, let us just give a comment on the role of the parameters β_i and γ_i (here again the parameter λ plays the same role than in Theorem 2.11 page 39). The parameter β_i is the localization parameter for the prior distribution compared with ρ^i. We have seen in the previous subsections how a good choice for such parameters may help to get rid of extra $\log N$ terms in the bound. Note as a limit case that $\beta_i=0$ leads to no localization at all, and in this case, the theorem is exactly the same than Theorem 2.11. The parameters γ_i are involved in the control of the theoretical entropy term by its empirical counterpart (we will see in the proof that it is a reference to the use of Theorem 2.18 page 49). The choice $\gamma_1=\gamma_2=\lambda$ will be made in some parts of this thesis with no incidence on the rate of convergence, but of course there is no reason for this choice to be optimal.

Proof. We just apply Theorem 2.11 with $\mu=\pi^1_{\exp[-\beta_1 R(\psi,.)]}\otimes\pi^2_{\exp[-\beta_2 R(\psi,.)]}$ and $\nu=\rho^1\otimes\rho^2$. We obtain:

$$\Phi_{\frac{\lambda}{N}}\left[\rho^1R(\psi,.)-\rho^2R(\psi,.)\right]\leq\rho^1r(\psi,.)-\rho^2r(\psi,.)+\frac{\lambda}{2N}\rho^1\otimes\rho^2v_\psi(.,.)$$
$$+\frac{1}{\lambda}\left[\log\frac{1}{\varepsilon}+\mathcal{K}\left(\rho^1,\pi^1_{\exp[-\beta_1R(\psi,.)]}\right)+\mathcal{K}\left(\rho^2,\pi^2_{\exp[-\beta_2R(\psi,.)]}\right)\right].$$

Combining this inequality with Theorem 2.18 (page 49) ends the proof. □

Let us propose here a way to use this bound to perform model selection. Let us assume that we are given submodels of Θ, namely a family of $\mathcal{F}$-measurable sets $(\Theta_i)_{i\in I}$ where I is at most countable and $\Theta_i\subset\Theta$ for any i. Let us choose coefficients $(p_i)_{i\in I}\in[0,1]^I$ such that:

$$\sum_{i\in I}p_i=1$$

and prior distribution π^i over each Θ_i: $\pi^i\in\mathcal{M}^1_+(\Theta_i)$; the idea being that we have a prior distribution

$$\pi=\sum_{i\in I}p_i\pi^i$$

over Θ. Let us also choose some atomic prior probability measure on the positive real line $m\in\mathcal{M}^1_+(\mathbb{R}_+)$, which will serve to make the inequality uniform in λ, γ_i, and β_i. Let us put, for any $\rho^1,\rho^2\in\mathcal{M}^1_+(\Theta)$, any $(i_1,i_2)\in I^2$:

$$b(\rho^1,\rho^1,i^1,i^2,\beta_1,\beta_2,\gamma_1,\gamma_2,\lambda)=\Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho^1r(\psi,.)-\rho^2r(\psi,.)+\frac{\lambda}{2N}\rho^1\otimes\rho^2v_\psi(.,.)\right.$$
$$+\frac{(\gamma_1\gamma_2-\beta_1\beta_2)}{\lambda(\gamma_1-\beta_1)(\gamma_2-\beta_2)}\log\left(\frac{3p_{i_1}p_{i_2}m(\lambda)\prod_{k=1}^2m(\beta_k)\,m(\gamma_k)}{\varepsilon}\right)$$
$$+\sum_{k=1}^{2}\frac{\gamma_k}{\lambda(\gamma_k-\beta_k)}\left\{\mathcal{K}\left(\rho^k,\pi^{i_k}_{\exp[-\beta_k r(\psi,.)]}\right)\right.$$
$$\left.\left.+\log\pi^{i_k}_{\exp[-\beta_k r(\psi,d\theta')]}\exp\left[\frac{\beta_k\gamma_k}{2N}\rho^k_{(d\theta)}v_\psi(\theta,\theta')\right]\right\}\right\},$$

and:

$$B(\rho^1,\rho^2)=\inf\Big\{b(\rho^1,\rho^1,i_1,i_2,\beta_1,\beta_2,\gamma_1,\gamma_2,\lambda);\\ i_k\in I,\beta_k<\gamma_k\in\mathbb{R}_+,k=1,2,\lambda\in\mathbb{R}_+\Big\}.$$

We propose the following idea for model selection (generalizing the selection scheme described in [11], section 1.5.7). Let us consider without great loss of generality some finite subset $\mathcal{P}$ of posterior distributions. We may for instance set:

$$\mathcal{P}=\{\pi^i_{\exp[-\beta_i r(\psi,.)]},i\in I,\beta_i\in\mathrm{supp}(m)\}$$

(assuming to get a finite set $\mathcal{P}$ that I is finite and m finitely supported). It should in particular be understood that $\mathcal{P}$ is allowed, as in this example, to be a random set of distributions on Θ.

Let us consider for any $\rho\in\mathcal{P}$ some complexity function $\mathcal{C}(\rho)$, which we may for example take to be, for some fixed real constant $\zeta>1$,

$$\mathcal{C}(\rho)=\inf\Bigg\{\left(\frac{\beta}{\gamma-\beta}+\frac{1}{\zeta-1}+1\right)\log\left[3\varepsilon^{-1}p_i m(\beta)m(\gamma)\right]\\ +\frac{\gamma}{\gamma-\beta}\Bigg\{\mathcal{K}\left(\rho,\pi^i_{\exp[-\beta r(\psi,.)]}\right)\\ +\log\pi^i_{\exp[-\beta r(\psi,.)]}\exp\left[\frac{\beta\gamma}{2N}\rho_{(d\theta)}v_\psi(\theta,\theta')\right]\Bigg\};i\in I,\beta,\gamma\in\mathbb{R}_+,\zeta\beta\leq\gamma\Bigg\}.$$

With this choice of complexity we see that

$$B(\rho^1,\rho^2)\leq\inf_{\lambda\in\mathbb{R}_+}\Phi^{-1}_{\frac{\lambda}{N}}\Bigg\{\rho^1 r(\psi,.)-\rho^2 r(\psi,.)+\frac{\lambda}{2N}\rho^1\otimes\rho^2 v_\psi(.,.)\\ +\frac{\zeta+1}{\lambda(\zeta-1)}\log\left[m(\lambda)\frac{\varepsilon}{3}\right]+\frac{\mathcal{C}(\rho^1)+\mathcal{C}(\rho^2)}{\lambda}\Bigg\}.$$

As a consequence, using the fact that $\Phi^{-1}_{\frac{\lambda}{N}}$ is concave, we see also that

$$B(\rho^1,\rho^2)+B(\rho^2,\rho^1)\leq 2\inf_{\lambda\in\mathbb{R}_+}\Phi^{-1}_{\frac{\lambda}{N}}\Bigg\{\frac{\lambda}{2N}\frac{\rho^1\otimes\rho^2+\rho^2\otimes\rho^1}{2}v_\psi(.,.)\\ +\frac{\zeta+1}{\lambda(\zeta-1)}\log\left[m(\lambda)\frac{\varepsilon}{3}\right]+\frac{\mathcal{C}(\rho^1)+\mathcal{C}(\rho^2)}{\lambda}\Bigg\}. \tag{2.7}$$

This shows that the symmetric part of B has an upper bound which contains only variance and complexity factors.

We can then apply to B and $\mathcal{C}$ the selection scheme described in [11]. We have first to chain the bound B on $\mathcal{P}$, defining for any $\rho,\rho'\in\mathcal{P}$

$$\tilde{B}(\rho,\rho')=\inf\left\{\sum_{k=1}^h B(\rho^{k-1},\rho^k),h\geq 1,(\rho^0,...,\rho^h)\in\mathcal{P}^{h+1},\rho^0=\rho,\rho^h=\rho'\right\}.$$

Let us also put by convention $\tilde{B}(\rho,\rho)=0$ for any $\rho\in\mathcal{P}$. Let us notice that $\tilde{B}(\rho,\rho')+\tilde{B}(\rho',\rho)\leq B(\rho,\rho')+B(\rho',\rho)$, and therefore still can be bounded in terms of variance and complexity factors only.

Let us then consider an indexation of $\mathcal{P}$ according to increasing complexities: let

$$\mathcal{P}=\{\rho^1,...,\rho^M\},$$

where $\mathcal{C}(\rho^{k+1})\geq\mathcal{C}(\rho^k)$, $1\leq k<M$.

Now, let us put for every $k \in \{1, ..., M\}$:

$$t(k) = \max\Big\{ j \in \{1, ..., M\}, \qquad \forall \ell \in \{1, ..., j\}, \tilde{B}(\rho^k, \rho^\ell) \leq 0 \Big\}.$$

Thus $t(k)$ is the largest starting interval of $\mathcal{P}$ which can be proved to perform worse than ρ^k. Let us choose now as our best estimator, $\rho_{\hat{k}}$ defined as

$$\hat{k} = \min(\arg\max t).$$

This means that we choose between the posterior distributions indexed by $\arg\max t$ a distribution with minimal complexity.

It is easy to prove that the following result, proved in [11], still holds in our context:

Theorem 2.21. *Let us put $\hat{t} = t(\hat{k})$. For any $\varepsilon > 0$, with P_N-probability at least $1 - \varepsilon$,*

$$\rho^{\hat{k}} R(\psi, .) \leq \rho^j R(\psi, .) + \begin{cases} 0, & 1 \leq j \leq \hat{t}, \\ \min\{\tilde{B}(\rho^\ell, \rho^j); 1 \leq \ell \leq \hat{t}\}, & \hat{t} < j < \hat{k}, \\ \tilde{B}(\rho^{\hat{k}}, \rho^{\hat{t}+1}) + \tilde{B}(\rho^{\hat{t}+1}, \rho^j), & j \in (\arg\max t) \\ \tilde{B}(\rho^{\hat{k}}, \rho^j), & otherwise. \end{cases}$$

Moreover for any $k \in (\arg\max t)$,

$$\begin{aligned} \tilde{B}(\rho^k, \rho^{\hat{t}+1}) &> 0 \\ \tilde{B}(\rho^{\hat{t}+1}, \rho^k) &> 0 \\ \tilde{B}(\rho^j, \rho^k) &> 0 \qquad j \notin (\arg\max t), \end{aligned}$$

and for any j such that $\hat{t} < j < \hat{k}$, since $j \notin (\arg\max t)$, there is $\ell \leq \hat{t}$ such that $\tilde{B}(\rho^j, \rho^\ell) > 0$. Thus

$$\rho^{\hat{k}} R(\psi, .) \leq \rho^j R(\psi, .) + \begin{cases} 0, & 1 \leq j \leq \hat{t}, \\ [B(\rho^\ell, \rho^j) + B(\rho^j, \rho^\ell)] & \\ \qquad \mathbb{1}[\min_{\ell \leq \hat{t}} \tilde{B}(\rho^j, \rho^\ell) > 0], & \hat{t} < j < \hat{k}, \\ B(\rho^j, \rho^{\hat{t}+1}) + B(\rho^{\hat{t}+1}, \rho^j) & \\ \qquad + B(\rho^{\hat{k}}, \rho^{\hat{t}+1}) + B(\rho^{\hat{t}+1}, \rho^{\hat{k}}) & j \in (\arg\max t), \\ B(\rho^j, \rho^{\hat{k}}) + B(\rho^{\hat{k}}, \rho^j), & otherwise, \end{cases}$$

showing that, according to (2.7) page 52, $(\rho^j - \rho^{\hat{k}}) R(\psi, .)$ can be bounded by variance and complexity terms relative to posterior distributions with a complexity not greater than $\mathcal{C}(\rho^j)$, and an empirical loss in any case not much larger than the one of ρ^j, as it is further developed in [11].

Finally, we give the moment version of Theorem 2.20 (page 50).

Theorem 2.22. *Let us choose $s > 1$ and $\alpha \in (0, 1)$. Let us choose two prior distributions π^1 and π^2 in $\mathcal{M}_+^1(\Theta)$. We assume that for some constant $\mathcal{D}_s$,*

$$\sup_{\theta, \theta' \in \Theta} M_{\psi, s}(\theta, \theta') \leq \mathcal{D}_s.$$

We also assume that for some measurable function $D = D(\lambda, Z_1, ..., Z_N)$ of the observation and of λ we have:

$$\sup_{\theta, \theta' \in \Theta} V_{\psi, \frac{\lambda}{N}, \alpha, s}(\theta, \theta') \leq D,$$

Then we have, for any $\varepsilon > 0$, for any $(\lambda, \beta_1, \beta_2, \gamma_1, \gamma_2) \in \mathbb{R}_+^5$ such that $\beta_1 < \alpha\gamma_1$ and $\beta_2 < \alpha\gamma_2$, with P_N-probability at least $1 - \varepsilon$, for any $(\rho^1, \rho^2) \in \left[\mathcal{M}_+^1(\Theta)\right]^2$:

$$\rho^1 R(\psi,.) - \rho^2 R(\psi,.)$$
$$\leq \frac{1}{\alpha}\Phi^{-1}_{\frac{\lambda}{N}}\left\{\alpha\left[\rho^1 r(\psi,.) - \rho^2 r(\psi,.) + \frac{\alpha\lambda}{2N}\rho^1\otimes\rho^2 V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta')\right]\right.$$
$$+\sum_{i=1}^{2}\frac{\alpha\gamma_i}{\lambda(\alpha\gamma_i-\beta_i)}\left[\log\pi^i_{\exp[-\beta_i r(\psi,d\theta')]}\exp\left(\frac{\alpha\gamma_i\beta_i}{2N}\rho^i_{d\theta}V_{\psi,\frac{\gamma_i}{N},\alpha,s}(\theta,\theta')\right.\right.$$
$$\left.+\frac{\beta_i}{s}\left(\frac{(s-1)\gamma_i}{sN}\right)^{s-1}\rho^i_{d\theta}M_{\psi,s}(\theta,\theta')\right)$$
$$\left.+\mathcal{K}\left(\rho^i,\pi^i_{\exp[-\beta_i r(\psi,.)]}\right)\right]$$
$$\left.+\frac{1}{\lambda}\left(1+\frac{\beta_1}{\alpha\gamma_1-\beta_1}+\frac{\beta_2}{\alpha\gamma_2-\beta_2}\right)\log\frac{3}{\varepsilon}\right\}$$
$$+\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\rho^1\otimes\rho^2 M_{\psi,s}(\theta,\theta').$$

Before we give the proof, note that the additional hypothesis (existence of $\mathcal{D}_s$) is implied by the existence of the bound discussed at the end of subsection 2.6 ($M_{\psi,s}(\theta,\theta')\leq C(s)\|\theta-\theta'\|_\Theta$) together with the assumption that $\pi_i(\{\theta\in\Theta,\|\theta\|\leq k\})=1$ for $i\in\{1,2\}$. Note that this hypotheses, together with $\alpha\leq 1/2$, ensures in classical cases that there is a constant D such that for any $\lambda\leq N$,

$$V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta')\leq D,$$

we will see it in detail in the last section about the linear case.

The proof of this theorem uses the same guideline as the one of Theorem 2.20.

Proof. We just apply the second part (moment bound) of Theorem 2.11 with:

$$\mu=\pi^1_{\exp[-\beta_1 R(\psi,.)]}\otimes\pi^2_{\exp[-\beta_2 R(\psi,.)]}$$

and:

$$\nu=\rho^1\otimes\rho^2$$

and we obtain:

$$\rho^1 R(\psi,.) - \rho^2 R(\psi,.)$$
$$\leq \frac{1}{\alpha}\Phi^{-1}_{\frac{\lambda}{N}}\left\{\alpha\left[\rho^1 r(\psi,.) - \rho^2 r(\psi,.) + \frac{\alpha\lambda}{2N}\rho^1\otimes\rho^2 V_{\psi,\frac{\lambda}{N},\alpha,s}(.,.)\right]\right.$$
$$\left.+\frac{1}{\lambda}\left[\log\frac{1}{\varepsilon}+\mathcal{K}\left(\rho^1,\pi^1_{\exp[-\beta_1 R(\psi,.)]}\right)+\mathcal{K}\left(\rho^2,\pi^2_{\exp[-\beta_2 R(\psi,.)]}\right)\right]\right\}$$
$$+\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\rho^1\otimes\rho^2 M_{\psi}(\theta,\theta').$$

Combining this inequality with Theorem 2.19 (page 50) ends the proof. □

3. PAC-Bayesian regression in the transductive setting

3.1. **Additional definitions and notations.** In this section, we focus on the transductive setting, as described in the introduction.

Let us choose $k\in\mathbb{N}^*$ and $P_{(k+1)N}$ is some probability measure on the space:

$$\left((\mathcal{X}\times\mathcal{Y})^{(k+1)N},(\mathcal{B}_\mathcal{X}\otimes\mathcal{B}_\mathcal{Y})^{\otimes(k+1)N}\right)=\left(\mathcal{Z}^{(k+1)N},\mathcal{B}_\mathcal{Z}^{\otimes(k+1)N}\right).$$

Let $(X_i, Y_i)_{i=1...(k+1)N} = (Z_i)_{i=1...(k+1)N}$ we the canonical process on this space.

Definition 3.1. We will call $(Z_i)_{i=1...N}$ the learning sample (we assume that the statistician observes it) and $(Z_i)_{i=N+1...(k+1)N}$ the test sample (we assume that the statistician knows only $(X_i)_{i=N+1...(k+1)N}$ and wants to predict $(Y_i)_{i=N+1...(k+1)N}$).

Definition 3.2. For $i \in \{1, ..., N\}$ let $\tau_i : \mathcal{Z}^{(k+1)N} \to \mathcal{Z}^{(k+1)N}$ be defined for any $z = (z_i)_{i=1...(k+1)N} \in \mathcal{Z}^{(k+1)N}$ by:

$$\begin{cases} \tau_i(z)_{i+jN} = z_{i+(j-1)N}, \quad j \in \{1, ..., k\}, \\ \\ \tau_i(z)_i = z_{i+kN}, \\ \\ \tau_i(z)_{m+jN}, \quad m \neq i, \quad m \in \{1, ..., N\}, \quad j \in \{0, ..., k\}. \end{cases}$$

Definition 3.3. We say that $P_{(k+1)N}$ is partially exchangeable if for any $i \in \{1, ..., N\}$,

$$P_{(k+1)N} \circ \tau_i^{-1} = P_{(k+1)N}.$$

In the same way, any function g from $\mathcal{Z}^{(k+1)N}$ to any space will be said to be partially exchangeable if for any $i \in \{1, ..., N\}$, $g \circ \tau_i = g$.

Let us remark that this implies that the distribution of $(X_i, Y_i)_{i=1...(k+1)N}$ is unchanged by circular permutations of $(X_{i+jN}, Y_{i+jN})_{j=0...k}$ for any $i \in \{1, ..., N\}$. From now we assume that $P_{(k+1)N}$ satisfies Definition 3.3. Note that this implies that for any i and j, Z_i and Z_{i+jN} have the same marginal distribution. For the sake of coherence with the inductive case, we will let p_i denote this distribution. So note that of course we can have:

$$P_{(k+1)N} = \left[\bigotimes_{i=1}^{N} p_i\right]^{\otimes(k+1)},$$

but we will not assume that we are in this particular case, unless in some cases where we explicitly mention it. Moreover, another particular case of interest is the i. i. d. case with every p_i equal to P and $P_{(k+1)N} = P^{\otimes(k+1)N}$.

Definition 3.4. For any bounded measurable function $h : \left(\mathcal{Z}^{(k+1)N}, \mathcal{B}_{\mathcal{Z}}^{\otimes(k+1)N}\right) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, where $\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$ we put:

$$T_i(h) = \frac{1}{k+1} \sum_{j=0}^{k} h \circ \tau_i^j.$$

Let $T = T_N \circ ... \circ T_1$.

Note that under our assumption that $P_{(k+1)N}$ is partially exchangeable, we have, for any h:

$$P_{(k+1)N}(h) = P_{(k+1)N}[T_i(h)] = P_{(k+1)N}[T(h)]$$

for any $i \in \{1, ..., N\}$.

Remember that we took, for any measurable nonnegative function $\psi : \mathcal{Y}^2 \to \mathbb{R}$ for any $\theta \in \Theta$, $\psi_i(\theta) = \psi\left(f_\theta(X_i), Y_i\right)$ and:

$$r_1(\psi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \psi_i(\theta),$$

$$r_2(\psi, \theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \psi_i(\theta).$$

We will also use the notation:

$$\overline{r}(\psi,\theta)=\frac{1}{(k+1)N}\sum_{i=1}^{(k+1)N}\psi_i(\theta)=\frac{r_1(\psi,\theta)+kr_2(\psi,\theta)}{k+1}.$$

3.2. **Main lemma and deviation inequality.** We proceed now exactly as in the inductive case: we give a deviation inequality as a lemma, then introduce a prior measure π and combine both to obtain the theorem.

Lemma 3.1. *For any partially exchangeable measurable functions $\lambda:\mathcal{Z}^{(k+1)N}\to\mathbb{R}_+^*$ and $\eta:\mathcal{Z}^{(k+1)N}\to\mathbb{R}$, for any $\theta\in\Theta$ we have:*

$$T\exp\left\{-\lambda r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),\theta\right]+\lambda\Phi_{\frac{\lambda}{N}}\left[\overline{r}\left(\psi\wedge\frac{N}{\lambda},\theta\right)\right]-\eta\right\}\leq\exp(-\eta)$$

where $\eta=\eta(Z_1,...,Z_{(k+1)N})$ and $\lambda=\lambda(Z_1,...,Z_{(k+1)N})$ for short.

Proof. We have:

$$\begin{aligned}T\exp\left\{-\lambda r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),\theta\right]\right\}&=T\left[\prod_{i=1}^N\left(1-\psi_i(\theta)\wedge\frac{N}{\lambda}\right)\right]\\&=\prod_{i=1}^N T_i\left(1-\psi_i(\theta)\wedge\frac{N}{\lambda}\right)=\prod_{i=1}^N\left[1-T_i\left(\psi_i(\theta)\wedge\frac{N}{\lambda}\right)\right]\\&=\exp\left\{\frac{-\lambda}{N}\sum_{i=1}^N\Phi_{\frac{\lambda}{N}}\left[T_i\left(\psi_i(\theta)\wedge\frac{N}{\lambda}\right)\right]\right\}\\&\leq\exp\left\{-\lambda\Phi_{\frac{\lambda}{N}}\left[\frac{1}{N}\sum_{i=1}^N T_i\left(\psi_i(\theta)\wedge\frac{N}{\lambda}\right)\right]\right\}\end{aligned}$$

by the convexity of Φ. Now, remark that:

$$\frac{1}{N}\sum_{i=1}^N T_i\left(\psi_i(\theta)\wedge\frac{N}{\lambda}\right)=\overline{r}\left(\psi\wedge\frac{N}{\lambda},\theta\right).$$

□

Definition 3.5. We choose a prior distribution π as a partially exchangeable function: $\mathcal{Z}^{(k+1)N}\to\mathcal{M}_+^1(\Theta)$. For the sake of simplicity, we will write π instead of $\pi(Z_1,...,Z_{(k+1)N})$.

This means that the prior is allowed to be data-dependent, but in a partially exchangeable way only. Of course we have in particular the possibility to choose π that does not actually depend of the data, as we did in the inductive setting.

Theorem 3.2. *For any partially exchangeable measurable function $\lambda:\mathcal{Z}^{(k+1)N}\to\mathbb{R}_+^*$, for any $\varepsilon>0$ we have, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any posterior $\rho\in\mathcal{M}_+^1(\Theta)$:*

$$\rho\left[\overline{r}\left(\psi\wedge\frac{N}{\lambda},.\right)\right]\leq\Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho\left\{r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}+\frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\lambda}\right\}$$

and equivalently:

$$\begin{aligned}&\rho\left[r_2\left(\psi\wedge\frac{N}{\lambda},.\right)\right]\\&\qquad\leq\frac{k+1}{k}\Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho\left\{r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}+\frac{\mathcal{K}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\lambda}\right\}\end{aligned}$$

$$- \frac{1}{k}\rho\left[r_1\left(\psi \wedge \frac{N}{\lambda}, .\right)\right].$$

Proof. Let us put:

$$\begin{aligned} H &= H(Z_1, ..., Z_{(k+1)N}) \\ &= -\lambda r_1\left[\Phi_{\frac{\lambda}{N}}\left(\psi \wedge \frac{N}{\lambda}\right), \theta\right] + \lambda\Phi_{\frac{\lambda}{N}}\left[\overline{r}\left(\psi \wedge \frac{N}{\lambda}, \theta\right)\right] - \eta \end{aligned}$$

for short. We apply Lemma 1.1 to obtain:

$$\begin{aligned} P_{(k+1)N}\exp\left\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)}\left[\rho(H) - \mathcal{K}(\rho,\pi)\right]\right\} &= P_{(k+1)N}\pi\exp(H) \\ &= P_{(k+1)N}T\left[\pi\exp(H)\right] = P_{(k+1)N}\pi T\left[\exp(H)\right] \leq \exp(-\eta). \end{aligned}$$

by Lemma 3.1. Now, let us choose $\eta = -\log\varepsilon$ to end the proof. □

We now give the moment version of Theorem 3.2.

Theorem 3.3. *For any $(\alpha, s) \in (0,1)\times(1,+\infty)$, for any partially exchangeable measurable function $\lambda : \mathcal{Z}^{(k+1)N} \to \mathbb{R}_+^*$, for any $\varepsilon > 0$ we have, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\begin{aligned} \rho\left[\overline{r}\left(\psi, .\right)\right] \leq \frac{1}{\alpha}\Phi_{\frac{\lambda}{N}}^{-1}\Bigg\{\rho r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\alpha\psi - \frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s, .\right)\right] \\ + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\Bigg\} + \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\rho\left[\overline{r}\left(|\psi|^s, .\right)\right]. \end{aligned}$$

Here again we have a non-observable term, $\rho\left[\overline{r}\left(|\psi|^s, .\right)\right]$, in the right-hand side. We have to make some hypothesis in order to upper-bound it by an observable quantity. We give an example with the following lemma.

Lemma 3.4. *Let us assume that we are in the case where $\mathcal{Y} = \mathbb{R}$ and where $\mathcal{B}_\mathcal{Y}$ is the Borel σ-algebra on $\mathbb{R}$, and that $\psi = l^p$ for a $p \in [1,+\infty)$. Let us assume that we know that the true regression function f is such that $\|f\|_\infty \leq C/2$ for some $C \geq 0$, and that π is chosen in such a way that $\pi(\{\theta \in \Theta, \|f_\theta\|_\infty \leq C/2\}) = 1$, $P_{(k+1)N}$-almost surely. Let us moreover assume that there are two constants $b > 0$ and $B < +\infty$ such that for any $x \in \mathcal{X}$:*

$$P\left\{\exp\left[b|Y - f(X)|\right]\middle| X = x\right\} \leq B.$$

Then we have, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, π-almost surely:

$$\rho\left[\overline{r}\left(|\psi|^s, .\right)\right] \leq \frac{1}{k+1}\rho\left[r_1\left(|\psi|^s, .\right)\right] + \frac{k}{k+1}\left[C + \frac{1}{b}\log\frac{kBN}{\varepsilon}\right]^{sp}.$$

Proof. For any $\theta \in \Theta$ and $\beta \in \mathbb{R}_+^*$:

$$\begin{aligned} P_{(k+1)N}\left[\exists i \in \{N+1, ..., (k+1)N\}, \quad \psi_{i,\theta}^s \geq \beta\right] \\ \leq \sum_{i=N+1}^{(k+1)N} P_{(k+1)N}\left(|f_\theta(X_i) - Y_i|^{sp} \geq \beta\right) \\ = kNP_{(k+1)N}\left(|f_\theta(X_1) - Y_1| \geq \beta^{\frac{1}{sp}}\right) \\ \leq kNP_{(k+1)N}\left(|f(X_1) - Y_1| \geq \beta^{\frac{1}{sp}} - C\right) \end{aligned}$$

$$\leq kNP_{(k+1)N}\exp\left(b|f(X_1)-Y_1|-b\beta^{\frac{1}{sp}}+bC\right)$$
$$\leq kNB\exp\left[b\left(C-\beta^{\frac{1}{sp}}\right)\right].$$

Now, remark that:

$$kNB\exp\left[b\left(C-\beta^{\frac{1}{sp}}\right)\right]\leq\varepsilon$$

if:

$$\beta\geq\left(C+\frac{1}{b}\log\frac{BkN}{\varepsilon}\right)^{sp}.$$

This ends the proof. □

We can combine Theorem 3.3 and Lemma 3.4 by a union bound argument to obtain the following corollary.

Corollary 3.5. *Let us assume that the conditions given in Lemma 3.4 are satisfied. For any $s>1$, for any partially exchangeable measurable function $\lambda:\mathcal{Z}^{(k+1)N}\to\mathbb{R}_+^*$, for any $\varepsilon>0$ we have, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any posterior $\rho\in\mathcal{M}_+^1(\Theta)$:*

$$\rho\left[r_2\left(\psi,.\right)\right]\leq\frac{k+1}{k\alpha}\Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\alpha\psi-\frac{\alpha}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s,.\right)\right]\right.$$
$$\left.+\frac{\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon}}{\lambda}\right\}+\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\left[C+\frac{1}{b}\log\frac{2kBN}{\varepsilon}\right]^{sp}$$
$$+\frac{1}{k}\rho\left[r_1\left(\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}|\psi|^s-\psi,.\right)\right].$$

3.3. **Inductive bounds obtained by integration with respect to the test sample.** It is possible to use deduce bounds for the inductive setting by using Lemma 3.1, we just need to integrate with respect to the test sample Z_{N+1}, ..., $Z_{(k+1)N}$. We will assume in this whole subsection that:

$$P_{(k+1)N}=\left[\bigotimes_{i=1}^{N}p_i\right]^{\otimes(k+1)}.$$

Definition 3.6. We put:

$$\mathbf{P}_{(k+1)N}=P_{(k+1)N}\left[.\,|Z_1,...,Z_N\right].$$

Theorem 3.6. *In the case where:*

$$P_{(k+1)N}=\left[\bigotimes_{i=1}^{N}p_i\right]^{\otimes(k+1)},$$

for any $\lambda\in\mathbb{R}_+^*$, for any $\varepsilon>0$ we have, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any posterior $\rho\in\mathcal{M}_+^1(\Theta)$:

$$\rho\left[R\left(\psi,.\right)\right]$$
$$\leq\frac{k+1}{k}\Phi_{\frac{\lambda}{N}}^{-1}\left\{\rho\left\{r_1\left[\Phi_{\frac{\lambda}{N}}\circ\left(\psi\wedge\frac{N}{\lambda}\right),.\right]\right\}+\frac{\mathbf{P}_{(k+1)N}\left[\mathcal{K}(\rho,\pi)\right]+\log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$-\frac{1}{k}\rho\left[r_1\left(\psi,.\right)\right]+\rho\left[\Delta_{\frac{\lambda}{N}}\left(\psi,.\right)\right].$$

Remark 3.1. Note that this is exactly Theorem 3.2 where the risk on the test sample, r_2, is replaced by the mean risk R.

Proof. We follow the proof of Theorem 3.2. We still take: Let us put:

$$H = -\lambda r_1 \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), \theta \right] + \lambda \Phi_{\frac{\lambda}{N}} \left[\overline{r} \left(\psi \wedge \frac{N}{\lambda}, \theta \right) \right] - \eta.$$

We have:

$$\begin{aligned} P_{(k+1)N} \exp \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho(\mathbf{P}_{(k+1)N} H) - \mathbf{P}_{(k+1)N}[\mathcal{K}(\rho, \pi)] \right] \right\} \\ \leq P_{(k+1)N} \exp \mathbf{P}_{(k+1)N} \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho(H) - \mathcal{K}(\rho, \pi) \right] \right\} \\ \leq P_{(k+1)N} \mathbf{P}_{(k+1)N} \exp \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho(H) - \mathcal{K}(\rho, \pi) \right] \right\} \\ = P_{(k+1)N} \exp \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\rho(H) - \mathcal{K}(\rho, \pi) \right] \right\} \end{aligned}$$

by Jensen's inequality. Using Lemma 1.1 and Lemma 3.1 exactly as we did in the proof of Theorem 3.2 we upper bound this last quantity by $\exp(-\eta)$, and we choose $\eta = -\log \varepsilon$ to end the proof. □

Note that the bound is not as good as the one given in the inductive setting (Theorem 2.5). However, the bound becomes exactly the same as $k \to +\infty$. Moreover, we will see later that is is more convenient to use this result as we are allowed to choose a data dependent prior π, but in this case we can wonder what k to choose? The idea to make a union bound over all possibles values for $k \in \mathbb{N}^*$ (using prior $1/(k+k^2)$ on k) and to choose the k that gives the best upper bound is due to Catoni [11]. Here, it leads to the following result.

Corollary 3.7. *For any $\lambda \in \mathbb{R}_+^*$, for any $\varepsilon > 0$ we have, with P_N-probability at least $1-\varepsilon$, for any posterior $\rho \in \mathcal{M}_+^1(\Theta)$:*

$$\rho\left[R(\psi, .)\right] \leq \inf_{k \in \mathbb{N}^*} \left\{ \frac{k+1}{k} \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho r_1 \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), . \right] + \frac{\mathbf{P}_{(k+1)N}\left[\mathcal{K}(\rho, \pi)\right] + \log \frac{k(k+1)}{\varepsilon}}{\lambda} \right\} - \frac{1}{k} \rho \left[\Delta_{\frac{\lambda}{N}}(\psi, .) \right] \right\}.$$

3.4. **Relative bounds in the transductive setting.** We now give relative bounds in the transducive setting (Theorem 3.8) as well as their integrated version for the inductive setting (3.9).

Definition 3.7. We choose a prior distribution μ as a partially exchangeable function: $\mathcal{Z}^{(k+1)N} \to \mathcal{M}_+^1(\Theta^2)$ (as previously, for the sake of simplicity, we will write μ instead of $\mu(Z_1, ..., Z_{(k+1)N})$).

Theorem 3.8. *For any $\varepsilon > 0$, for any partially exchangeable measurable function λ taking values $\mathbb{R}_+^*$, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:*

$$\begin{aligned} \Phi_{\frac{\lambda}{N}} \left\{ \nu \left[\frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} (\psi_i(\theta) - \psi_i(\theta')) \wedge \frac{N}{\lambda} \right] \right\} \\ \leq \nu \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[(\psi_i(\theta) - \psi_i(\theta')) \wedge \frac{N}{\lambda} \right] \right\} + \frac{\mathcal{K}(\nu, \mu) + \log \frac{1}{\varepsilon}}{\lambda}. \end{aligned}$$

For the "moment case", let us choose $s > 1$. Then, for any $\alpha \in (0,1)$ and $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\nu\left[r_2\left(\psi,\theta\right) - r_2\left(\psi,\theta'\right)\right] \leq \left[\frac{k+1}{k}\right] \frac{1}{\alpha} \Phi_{\frac{\lambda}{N}}^{-1}\left\{\nu \frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[\alpha\left(\psi_i(\theta) - \psi_i(\theta')\right.\right.\right.$$
$$\left.\left.\left. - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1} \left|\psi_i(\theta) - \psi_i(\theta')\right|^s\right)\right] + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$- \frac{1}{k}\nu\left[r_1\left(\psi,\theta\right) - r_1\left(\psi,\theta'\right)\right]$$
$$+ \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1} \nu\left[\frac{1}{kN}\sum_{i=1}^{(k+1)N} \left|\psi_i(\theta) - \psi_i(\theta')\right|^s\right].$$

Theorem 3.9. *In the case where:*

$$P_{(k+1)N} = \left[\bigotimes_{i=1}^{N} p_i\right]^{\otimes(k+1)},$$

for any $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\nu\left\{\frac{1}{N}\sum_{i=1}^{N} p_i\left[\left[\psi\left(f_\theta(X),Y\right) - \psi\left(f_{\theta'}(X),Y\right)\right] \wedge \frac{N}{\lambda}\right]\right\}$$
$$\leq \left[\frac{k+1}{k}\right] \Phi_{\frac{\lambda}{N}}^{-1}\left\{\nu\left\{\frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[\left(\psi_i(\theta) - \psi_i(\theta')\right) \wedge \frac{N}{\lambda}\right]\right\} + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$- \nu\left[\frac{1}{kN}\sum_{i=1}^{N}\left(\psi_i(\theta) - \psi_i(\theta')\right) \wedge \frac{N}{\lambda}\right].$$

Let us choose $s > 1$. Then, for any $\alpha \in (0,1)$ and $\varepsilon > 0$, for any $\lambda \in \mathbb{R}_+^*$, with P_N-probability at least $1-\varepsilon$, for any $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\nu\left[R\left(\psi,\theta\right) - R\left(\psi,\theta'\right)\right] \leq \left[\frac{k+1}{k}\right] \frac{1}{\alpha} \Phi_{\frac{\lambda}{N}}^{-1}\left\{\nu \frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[\alpha\left(\psi_i(\theta) - \psi_i(\theta')\right.\right.\right.$$
$$\left.\left.\left. - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1} \left|\psi_i(\theta) - \psi_i(\theta')\right|^{s'}\right)\right] + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$
$$- \frac{1}{k}\nu\left[r\left(\psi,\theta\right) - r\left(\psi,\theta'\right)\right]$$
$$+ \frac{1}{sk}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1} \nu\left[kM_{\psi,s}(\theta,\theta') + \frac{1}{N}\sum_{i=1}^{N} \left|\psi_i(\theta) - \psi_i(\theta')\right|^s\right].$$

In order to make the formulas more explicit, let us compare this last inequality with the one in Theorem 2.11 page 39 (the analogous of this result obtained without using transductive bounds). The variance term (last line) is not the same but plays exactly the same role. The generic term:

$$\frac{1}{\alpha} \Phi_{\frac{\lambda}{N}}^{-1}\left\{\nu \frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[\alpha\left(\psi_i(\theta) - \psi_i(\theta')\right.\right.\right.$$
$$\left.\left.\left. - \frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1} \left|\psi_i(\theta) - \psi_i(\theta')\right|^{s'}\right)\right] + \frac{\mathcal{K}(\nu,\mu) + \log\frac{1}{\varepsilon}}{\lambda}\right\}$$

is the same but is here multiplied by $(k+1)/k$, but this is "almost" compensated by the presence of the term:

$$-\frac{1}{k}\nu\left[r\left(\psi,\theta\right)-r\left(\psi,\theta'\right)\right].$$

Of course, it is easy to see that the bound is better when k is large, and as a limit case, the ancient bound (Theorem 2.11, obtained without using the transductive bounds as an intermediary step) is better. However, we insist on the fact that Theorem 3.9 allows to use a data-dependant prior, while Theorem 2.11 does not. For the role of the various parameters, we refer the reader to previous discussions (as Remark 2.2 page 34).

Finally, we give a localized version of this bound: the analogous of Theorem 2.22 (page 53) obtained by integration of transductive bounds.

Theorem 3.10. *In the case where:*

$$P_{(k+1)N}=\left[\bigotimes_{i=1}^{N}p_i\right]^{\otimes(k+1)},$$

let us choose $s>1$ and $\alpha\in(0,1)$. Let us choose two prior distributions π^1 and π^2 in $\mathcal{M}_+^1(\Theta)$. Then we have, for any $\varepsilon>0$, for any $(\lambda,\beta_1,\beta_2,\gamma_1,\gamma_2)\in\mathbb{R}_+^5$ such that $\beta_1<\alpha\gamma_1$ and $\beta_2<\alpha\gamma_2$, with P_N-probability at least $1-\varepsilon$, for any $(\rho^1,\rho^2)\in\left[\mathcal{M}_+^1(\Theta)\right]^2$:

$$\begin{aligned}
&\rho^1R(\psi,.)-\rho^2R(\psi,.)\\
&\qquad\leq\rho^1r(\psi,.)-\rho^2r(\psi,.)+\left[\frac{k+1}{k}\right]\frac{\alpha\lambda}{2N}\rho^1\otimes\rho^2V_{\psi,\frac{\lambda}{N},\alpha,s}(\theta,\theta')\\
&\quad+[\tfrac{k+1}{k}]\sum_{j=1}^{2}\left\{\frac{\gamma_j}{\lambda(\alpha\gamma_j-\beta_j)}\log\pi^j_{\exp[-\beta_jr(\psi,d\theta')]}\exp\left[\frac{\alpha\lambda\beta_j\gamma_j}{2N}\rho^j_{d\theta}V_{\psi,\frac{\gamma_j}{N},\alpha,s}(\theta,\theta')\right.\right.\\
&\qquad+\tfrac{\beta_j}{s}\left(\tfrac{(s-1)\gamma_j}{sN}\right)^{s-1}\rho^j_{d\theta}\left(\left[\tfrac{k}{k+1}\right]M_{\psi,s}(\theta,\theta')+\left[\tfrac{1}{k+1}\right]\frac{1}{N}\sum_{i=1}^{N}\left|\psi_i(\theta)-\psi_i(\theta')\right|^s\right)\Bigg]\\
&\qquad\left.\left.+\frac{\gamma_j}{\lambda(\alpha\gamma_j-\beta_j)}\mathcal{K}\left(\rho^j,\pi^j_{\exp[-\beta_jr(\psi,.)]}\right)\right\}\\
&\qquad+\left[\frac{k+1}{k}\right]\left(1+\frac{\beta_1}{\alpha\gamma_1-\beta_1}+\frac{\beta_2}{\alpha\gamma_2-\beta_2}\right)\frac{\log\frac{3}{\varepsilon}}{\alpha\lambda}\\
&\quad+\frac{1}{s}\left(\frac{(s-1)\lambda}{sN}\right)^{s-1}\left[\rho^1\otimes\rho^2M_{\psi,s}(\theta,\theta')+\frac{1}{k}\rho^1\otimes\rho^2\frac{1}{N}\sum_{i=1}^{N}\left|\psi_i(\theta)-\psi_i(\theta')\right|^s\right].
\end{aligned}$$

Proof. The proof this theorem is similar to the one of its analogous obtained in the inductive case, Theorem 2.22 page 53, using Theorem 3.9 (page 60) instead of Theorem 2.11 (page 39). □

In order to help the reader to interpret the formulas let us give some comments about this theorem. We invite the reader to compare this theorem with Theorem 2.22. Note that it is almost the same with some additional variance terms, and a factor $(k+1)/k$. Note that here again, k is chosen by the statistician and that large values leads to better bounds. However, the choice of the limit case ($k=+\infty$) leads to the inductive bound (Theorem 2.22) in which we are not allowed to choose a data-dependant prior. A discussion about the other parameters can be found in Remark 2.3 page 51. Remember that β_i indicates "how much we localize" the prior π^i. The limit case $\beta_i=0$ leads to no localization at all and in this case the theorem

is similar to the previous one (Theorem 3.9 page 60). In the applications of this theorem, we will see that we can take $\gamma_i = \lambda$ and $\alpha = 1/2$ with no loss on the order of magnitude of the bound. Finally, s is the order of the largest moment we assume to exist for ψ.

4. A first application: compression schemes, and extensions

4.1. Presentation of compression schemes.

Compression schemes were introduced by Littlestone and Warmuth [27] as a way to implement Rissanen's minimum description length (MDL, [45, 4]) principle in the context of classification, which principle can be seen as a control of the complexity of a model by an exigence on its ability to compress the data.

The idea is to use a small subset $Z_I = (Z_i, i \in I)$ of the sample $(Z_i, i \in \{1, ..., (k+1)N\})$, called the compression set, and to build an estimator only on the basis of Z_I. We can then control the difference between the performance of this estimator on the training sample and on the test sample by a bound with a complexity term depending on $h = |I|$.

Definition 4.1. Let us assume that we have a function (the compression scheme):

$$t : \bigcup_{i=0}^{+\infty} (\mathcal{X} \times \mathcal{Y})^i \to \Theta$$
$$z = ((x_1, y_1), ..., (x_i, y_i)) \mapsto t(z)$$

such that, for any $i \in \mathbb{N}*$, for any permutation σ of $\{1, ..., i\}$ we have:

$$t((x_1, y_1), ..., (x_i, y_i)) = t((x_{\sigma(1)}, y_{\sigma(1)}), ..., (x_{\sigma(i)}, y_{\sigma(i)})).$$

Let us put, for any $I \subset \{1, ..., (k+1)N\}$:

$$Z_I = (Z_i)_{i \in I}.$$

Note that this notations are due to Catoni [10], and Audibert [2], who adapted the work of Littlestone and Warmuth to the PAC-Bayesian setting.

4.2. An extension of compression schemes: indexed compression schemes.

The motivation for this extention of compression schemes is the following classical example.

Example 4.1 (Support Vector Machines as compression schemes)**.** Let us assume that we are given a measurable function:

$$K : \mathcal{X}^2 \to \mathcal{Y}$$
$$(x, x') \mapsto K(x, x').$$

Then we can propose the following compression scheme:

$$f_{t(Z_I)}(.) = \sum_{i \in I} \hat{\alpha}_i K(X_i, .),$$

where $\hat{\alpha} = (\hat{\alpha}_i, i \in I)$ is given by:

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^I} \sum_{j \in I} \psi \left[Y_j, \sum_{i \in I} \alpha_i K(X_i, X_j) \right].$$

When $\mathcal{Y} \subset \mathbb{R}$ and $K(.,.)$ is a Mercer's kernel, we obtain an estimator that looks like support vector machine (defined by Boser, Guyon and Vapnik [7] in the classification case $|\mathcal{Y}| = 2$, and extended by Vapnik [41] to the case where $\mathcal{Y} = \mathbb{R}$; note that in most of the generalization bounds given for SVM, the complexity is controlled through the margin of the classifier, however, the compression scheme point of view has already been studied, see for example Fung, Mangasarian and Smola [20]).

Actually, one of the problems with SVM estimators is the choice of the function K (note that here, K is not restricted to a Mercer's kernel). In the case where $(\mathcal{X}, d)$ is a metric space we can use the Gaussian kernel:

$$K(x, x') = \exp\left(-\frac{d^2(x, x')}{2\gamma^2}\right)$$

but the choice of γ^2 remains a problem. A possibility is to use several kernels (indexed by a finite set D) and to let the bound determinate which kernel is the most appropriate for each point.

Definition 4.2. Let us assume that we have a finite set of index D and a function (indexed compression scheme):

$$t : \bigcup_{i=0}^{+\infty} (\mathcal{X} \times \mathcal{Y} \times D)^i \to \Theta$$
$$z = ((x_1, y_1, d_1), ..., (x_i, y_i, d_i)) \mapsto t(z)$$

such that, for any $i \in \mathbb{N}*$, for any permutation σ of $\{1, ..., i\}$ we have:

$$t((x_1, y_1, d_1), ..., (x_i, y_i, d_i)) = t((x_{\sigma(1)}, y_{\sigma(1)}, d_{\sigma(1)}), ..., (x_{\sigma(i)}, y_{\sigma(i)}, d_{\sigma(i)})).$$

Let us put, for any $h \in \{1, ..., N\}$, $I = (I_1, ..., I_N) \in \{1, ..., N\}^h$ and any $d \in D^{|I|}$:

$$Z_{h,I,d} = \left((X_{I_1}, Y_{I_1}, d_1), ..., (X_{I_h}, Y_{I_h}, d_h)\right).$$

Remark 4.1. Note that in the case where $|D| = 1$, we obtain a compression scheme in the meaning of Definition 4.2.

Example 4.2 (SVM, continued)**.** In our example, we choose a family of functions $K_1, ..., K_l : \mathcal{X}^2 \to \mathcal{Y}$ and we put, for $h \in \{1, ..., N\}$, $I = (i_1, ..., i_h) \in \{1, ..., N\}^h$ and $d \in \{1, ..., l\}^h$:

$$f_{t(Z_{h,I,d})}(.) = \sum_{i=1}^{h} \hat{\alpha}_i K_{d_i}(X_{I_i}, .),$$

where $\hat{\alpha} = (\hat{\alpha}_i, i = 1, ..., h)$ is given by:

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^I} \sum_{j=1}^{h} \psi\left[Y_j, \sum_{k=1}^{h} \alpha_k K_{d_k}(X_{i_k}, X_{i_j})\right].$$

Another possibility of extension is to allow a choice of the coefficients α_i on the basis of the whole sample. In order to do this, we propose an interval for the values of α_i: $[-C, C]$ with a given $C > 0$ (note that in many SVM algorithms, constraints impose the coefficients to lie in a compact interval, see Cristianini and Shawe-Taylor [16]). Then let us choose D as a discretization of this interval of size $2M + 1$ (with a given $M \geq 1$):

$$D = \left\{C\left(1 - \frac{i}{M}\right), \quad i \in \{0, ..., 2M\}\right\},$$

and define, for $h \in \{1, ..., N\}$, $I = (i_1, ..., i_h) \in \{1, ..., N\}^h$ and $d \in D^h$:

$$f_{t(Z_{h,I,d})}(.) = \sum_{i=1}^{h} C\left(1 - \frac{d_i}{M}\right) K(X_{I_i}, .).$$

Of course, we will see in the next subsection that we have to pay for the precision of the grid D in this case (the upper bound on the risk of $t(Z_{h,I,d})$ will be increasing with M. Note that this is related with margin-based bounds for SVM in the context of classification: if the data can be separated with a large margin, we can choose a small value for M, and we obtain an upper bound smaller than in the case where the data cannot be separated with a large margin, in which we have to choose a large

value for M. The link between PAC-Bayesian inference and margin is examined in detail by Langford and Shawe-Taylor in [25].

In what follows, we give upper bound on the risk of (indexed) compression schemes: direct bounds in subsection 4.3 and relative bounds in subsection 4.5. Most of the results are given in the transductive setting, note that the generalization to the inductive setting is easy by subsection 3.3. Note also that we give bounds on the truncated risk $r_2[\psi \wedge (N/\lambda), .]$; to derive the moment version from Theorem 3.3 is straightforward.

4.3. **Direct bounds.**

Theorem 4.1. *Let us choose $\alpha \in (0,1)$. For any partially exchangeable measurable functions $\lambda : \mathcal{Z}^{(k+1)N} \to \mathbb{R}_+^*$, for any $\varepsilon > 0$, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any $h \in \{0,...,N\}$, $I \in \{1,...,(k+1)N\}^h$ and $d \in D^h$ we have:*

$$r_2\left[\psi \wedge \frac{N}{\lambda}, t(Z_{h,I,d})\right] \leq \frac{k+1}{k}\Phi^{-1}_{\frac{\lambda}{N}}\left\{r_1\left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda}\right), t(Z_{h,I,d})\right] + \frac{1}{\lambda}\left[h\left(1+\log\frac{(k+1)N|D|}{h\alpha}\right) + \log\frac{1}{(1-\alpha)\varepsilon}\right]\right\} - \frac{1}{k}r_1\left[\psi \wedge \frac{N}{\lambda}, t(Z_{h,I,d})\right]. \tag{4.1}$$

Proof. Let us put, for any $h \in \{0,...,N\}$:

$$\mathcal{F}_h = \left\{t(Z_{h,I,d}), \quad I \in \{1,...,(k+1)N\}^h, d \in D^h\right\},$$

for $h \neq 0$, $\mathcal{E}_h = \mathcal{F}_h \setminus \mathcal{F}_{h-1}$ and $\mathcal{E}_0 = \mathcal{F}_0$, and:

$$\mathcal{C}_h = |\mathcal{E}_h|.$$

Remark that, by definition of the function t we have:

$$\mathcal{C}_h \leq \binom{(k+1)N|D|}{h}.$$

Let us choose a parameter $\alpha \in (0,1)$. We just apply Theorem 3.2 with the following choice of prior:

$$\pi = \sum_{h=0}^{N} \alpha^h(1-\alpha) \sum_{\theta \in \mathcal{E}_h} \frac{\delta_\theta}{\mathcal{C}_h} + \alpha^{(k+1)N}\delta_{\theta_0}$$

for an arbitrary $\theta_0 \in \Theta$ (the last term being here only to ensure that π is a probability measure). Note that this choice is admissible because π is an exchangeable function of $(Z_1,...,Z_{(k+1)N})$. Now, we use every posterior under the form:

$$\rho = \delta_\theta$$

for $\theta \in \mathcal{E}_h$ and $h \in \{0,...,N\}$. We can compute:

$$\begin{aligned}\mathcal{K}(\rho,\pi) &= -\log(1-\alpha) - h\log\alpha + \log\mathcal{C}_h \\ &\leq -\log(1-\alpha) - h\log\alpha + \log\binom{(k+1)N|D|}{h} \\ &\leq -\log(1-\alpha) - |I|\log\alpha + |I|\left(1+\log\frac{(k+1)N|D|}{|I|}\right).\end{aligned}$$

Finally, note that for any $h \in \{0,...,N\}$, $I \in \{1,...,(k+1)N\}^h$ and $d \in \{1,...,l\}^{|I|}$ there is (by definition) a $\theta \in \mathcal{E}_h$ such that:

$$\theta = t\left(Z_{h,I,d}\right).$$

So:

$$\delta_\theta \left\{ r_2 \left[\psi \wedge \frac{N}{\lambda}, . \right] \right\} = r_2 \left[\psi \wedge \frac{N}{\lambda}, t(Z_{h,I,d}) \right].$$

This ends the proof. □

We can of course give the version of this theorem in the inductive case. The idea is just to point out the difference with the transductive case, as we can only use data in the learning sample, so I is restricted to belong to $\{1, ..., N\}^h$. However, in the prior distribution, we shall give the same weight to every $I \in \{1, ..., (k+1)N\}^h$ in order to ensure the exchangeability of π.

Theorem 4.2. *We assume that we are in the case where:*

$$P_{(k+1)N} = \left[\bigotimes_{i=1}^{N} p_i \right]^{\otimes(k+1)}.$$

Let us choose $\alpha \in (0,1)$. For any $\lambda > 0$, for any $\varepsilon > 0$, with $P_{(k+1)N}$-probability at least $1 - \varepsilon$, for any $h \in \{0, ..., N\}$, $I \in \{1, ..., N\}^h$ and $d \in D^h$ we have:

$$R[\psi, t(Z_{h,I,d})] \leq \frac{k+1}{k} \Phi^{-1}_{\frac{\lambda}{N}} \left\{ r_1 \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), t(Z_{h,I,d}) \right] + \frac{1}{\lambda} \left[h \left(1 + \log \frac{(k+1)N|D|}{h\alpha} \right) + \log \frac{1}{(1-\alpha)\varepsilon} \right] \right\} - \frac{1}{k} r_1 \left[\psi \wedge \frac{N}{\lambda}, t(Z_{h,I,d}) \right] + \Delta_{\frac{\lambda}{N}} [\psi, t(Z_{h,I,d})].$$

4.4. **Basics algorithms for compression schemes.** Let $\mathcal{B}(Z_{h,I,d})$ denote the right-hand side of Equation 4.1. Let us remark that for reasonable values of N and even in the case of compression schemes ($|D| = 1$) the search for the exact minimum:

$$(h^*, I^*, d^*) = \arg \min_{(h,I,d)} \mathcal{B}(Z_{h,I,d})$$

may be not feasible.

However, note that the bound given by Theorem 4.1 has the advantage to be valid for every compression set of size $h \in \{0, ..., N\}$. This means that we are allowed to use every heuristic we want in order to select our compression set. We propose an example of heuristic based on thresholding in the case of SVM, and then propose a more general algorithm.

Example 4.3 (SVM, continued)**.** Remember that we proposed, for SVM (with one single kernel K):

$$f_{t(Z_I)}(.) = \sum_{i \in I} \hat{\alpha}_i K(X_i, .),$$

where $\hat{\alpha} = (\hat{\alpha}_i, i \in I)$ is given by:

$$\hat{\alpha} \in \arg \min_{\alpha \in \mathbb{R}^I} \sum_{j \in I} \psi \left[Y_j, \sum_{i \in I} \alpha_i K(X_i, X_j) \right].$$

We propose the following algorithm. First of all, choose $I = \{1, ..., N\}$ and so:

$$\hat{\alpha} \in \arg \min_{\alpha \in \mathbb{R}^N} \sum_{j=1}^{N} \psi \left[Y_j, \sum_{i=1}^{N} \alpha_i K(X_i, X_j) \right].$$

Then, choose a $\kappa > 0$ and replace I by:

$$I_\kappa = \{ i \in \{1, ..., N\}, \quad \hat{\alpha}_i \geq \kappa \}.$$

We can of course choose the value of κ in order to obtain a given value for $|I_\kappa|$ (and so a given compression rate), or try to minimize with respect to κ the upper bound on the risk of the estimator $t(Z_{I_\kappa})$ given by inequality 4.1.

Example 4.4 (Iterative selection in the general case)**.** We propose here a general iterative algorithm for (indexed) compression schemes:

- at step 0 start with $I_0 = \emptyset$ and $d_0 = ()$;
- at step h, we have $|I_h| = h$ and d_h, we define:
$$(j^*, \delta^*) = \arg \min_{\substack{j \in \{1, ..., N\} \\ \delta \in D}} \mathcal{B}(Z_{h+1,(I_h,j),(d_h,\delta)})$$
and $I_{h+1} = (I_h, j^*)$, $d_{h+1} = (d_h, \delta^*)$;
- stop when $h + 1 = N$ and then choose:
$$(h^*, I^*, d^*) = \arg \min_{h \in \{0,\dots,N\}} \mathcal{B}(Z_{h,I_h,d_h}).$$

At the end the guarantee is that the risk of the estimator defined by (h^*, I^*, d^*) does not exceed $\mathcal{B}(Z_{h,I_h,d_h})$ with probability at least $1 - \varepsilon$.

4.5. **Relative bounds and adaptation of the algorithm of Example 4.4.**

Theorem 4.3. *Let us choose $\alpha \in (0,1)$. For any $\varepsilon > 0$, for any partially exchangeable measurable function λ taking values $\mathbb{R}_+^*$, with $P_{(k+1)N}$-probability at least $1-\varepsilon$, for any $(h, h') \in \{1, ..., N\}^2$, $I \in \{1, ..., N\}^h$, $J \in \{1, ..., N\}^{h'}$, $d \in D^h$ and $d' \in D^{h'}$:*

$$\begin{aligned}(4.2)\quad &\left[\frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} (\psi_i[t(Z_{h,I,d})] - \psi_i[t(Z_{h',J,d'})]) \wedge \frac{N}{\lambda}\right] \\ &\leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{\frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[(\psi_i[t(Z_{h,I,d})] - \psi_i[t(Z_{h',J,d'})]) \wedge \frac{N}{\lambda}\right]\right. \\ &\qquad \left. + \frac{h\left(1 + \log \frac{(k+1)N|D|}{h\alpha}\right) + h'\left(1 + \log \frac{(k+1)N|D|}{h'\alpha}\right) + \log \frac{1}{(1-\alpha)\varepsilon}}{\lambda}\right\}.\end{aligned}$$

Proof. This is an application of Theorem 3.8 with prior:

$$\mu = \left[\sum_{h=0}^{N} \alpha^h (1-\alpha) \sum_{\theta \in \mathcal{E}_h} \frac{\delta_\theta}{\mathcal{C}_h} + \alpha^{(k+1)N} \delta_{\theta_0}\right]^{\otimes 2}$$

for a given $\theta_0 \in \Theta$ and:

$$\nu = \delta_{t(Z_{h,I,d})} \otimes \delta_{t(Z_{h',J,d'})}.$$

The computations of the entropy term is exactly the same than in the proof of Theorem 4.1. □

We now propose an algorithm for compression schemes using this new bound. For the sake of simplicity, let us assume that ψ takes values in $[0, 1]$. In this case, the bound in the theorem becomes:

$$\begin{aligned}(4.3)\quad &r_2[\psi, t(Z_{h,I,d})] - r_2[\psi, t(Z_{h',J,d'})] \\ &\leq \frac{k+1}{k} \Phi^{-1}_{\frac{\lambda}{N}}\left\{\nu\left[\frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}(\psi_i[t(Z_{h,I,d})] - \psi_i[t(Z_{h',J,d'})])\right]\right. \\ &\qquad \left. + \frac{h\left(1 + \log \frac{(k+1)N|D|}{h\alpha}\right) + h'\left(1 + \log \frac{(k+1)N|D|}{h'\alpha}\right) + \log \frac{1}{(1-\alpha)\varepsilon}}{\lambda}\right\}\end{aligned}$$

$$- \frac{1}{k} \left\{ r_1[\psi, t(Z_{h,I,d})] - r_1[\psi, t(Z_{h-1,I^-,d^-})] \right\}.$$

Let $B(h, I, d, h', J, d')$ denote the right-hand side of inequality 4.3. We can use B in the way described after Theorem 2.20 (page 50), using as complexity function $\mathcal{C}(h, I, d) = h$, to get the result stated in Theorem 2.21 (page 53).

5. A second application: linear regression estimation with quadratic loss

In this section we deal with the linear case in a more general setting than in Example 2.4 (but with similar techniques). The idea is to allow the comparison and selection of linear models by an application of Theorem 2.22.

5.1. **Notations and assumptions in the linear case.** We assume that we are in the i. i. d. case, with the distribution of every pair $Z_i = (X_i, Y_i)$, p_i, being equal to P, and so $P_N = P^{\otimes N}$. We assume that $\mathcal{Y} = \mathbb{R}$ and $\psi = l^2$. We take Θ as an Hilbert space with scalar product $\langle ., . \rangle_\Theta$ and associated norm $\|.\|_\Theta$. Let us put, for any $\delta \geq 0$ and $t \in \Theta$:

$$B_\Theta(t, \delta) = \{\theta \in \Theta, \quad \|\theta - t\|_\Theta \leq \delta\}.$$

We assume that $P(|Y|^s) \leq m_s$ and we take:

$$f_\theta = \langle \theta, \Psi(.) \rangle_\Theta,$$

with $\Psi : \mathcal{X} \to \Theta$ and for any $x \in \mathcal{X}$, $\|\Psi(x)\|_\Theta \leq K$. Note that we have proved that in this case, if θ and θ' are such that $\|\theta\| \leq k$ and $\|\theta'\| \leq \kappa$ we have:

$$M_{\psi,s}(\theta, \theta') \leq 2^{2s-1} K^s \left(m_s + \kappa^s K^s\right) \|\theta - \theta'\|_\Theta^s.$$

So the hypothesis discussed in subsection 2.6 ($M_{\psi,s}(\theta, \theta') \leq C(s) \|\theta - \theta'\|_\Theta^s$ for any θ and θ') is satisfied with:

$$C(s) = 2^{2s-1} K^s \left(m_s + \kappa^s K^s\right).$$

This also implies that, if θ and θ' are such that $\|\theta\| \leq \kappa$ and $\|\theta'\| \leq \kappa$ we have:

$$M_{\psi,s}(\theta, \theta') \leq 2^s \kappa^s C(s).$$

So the first hypothesis of Theorem 2.22 is satisfied with $\mathcal{D}_s = C(s) 2^s \kappa^s$.

From now, we consider only the case $s = 3$. Note that, using Lemma 2.5, as soon as $\alpha \leq \frac{1}{2}$, for any λ and $s \in \mathbb{N} \setminus \{0\}$:

$$\begin{aligned} V_{\psi, \frac{\lambda}{N}, \alpha, 3}(\theta, \theta') &\leq \frac{1}{N} \sum_{i=1}^N \left[\psi_i(\theta) - \psi_i(\theta') + \frac{1}{3} \left(\frac{(3-1)\lambda}{3N} \right)^3 |\psi_i(\theta) - \psi_i(\theta')|^3 \right]^2 \\ &\leq \frac{1}{N} \sum_{i=1}^N \left\{ [\psi_i(\theta) - \psi_i(\theta')]^2 + 2 \frac{64\lambda^6}{2187 N^6} [\psi_i(\theta) - \psi_i(\theta')]^4 \right. \\ &\qquad \left. + \left(\frac{64\lambda^6}{2187 N^6} \right)^2 [\psi_i(\theta) - \psi_i(\theta')]^6 \right\}. \end{aligned}$$

For short, note that $64/2187 \leq 1/30$. Now, note that:

$$\begin{aligned} |\psi_i(\theta) - \psi_i(\theta')| &= \left| Y_i - f_\theta(X_i))^2 - (Y_i - f_{\theta'}(X_i))^2 \right| \\ &= |2Y_i - f_\theta(X_i) - f_{\theta'}(X_i)| \, |f_\theta(X_i) - f_{\theta'}(X_i)| \\ &= |2Y_i - \langle \theta + \theta', \Psi(X_i) \rangle| \, |\langle \theta - \theta', \Psi(X_i) \rangle| \\ &\leq 2 \left(|Y_i| + \kappa \|\Psi(X_i)\|_\Theta \right) \|\Psi(X_i)\|_\Theta \|\theta - \theta'\|_\Theta \end{aligned}$$

Let us put:

$$\mathcal{C} = \mathcal{C}(Z_1, ..., Z_N) = \frac{1}{N}\sum_{i=1}^{N}\Bigg\{ 4\left(|Y_i| + \kappa \left\|\Psi(X_i)\right\|_\Theta\right)^2 \left\|\Psi(X_i)\right\|_\Theta^2 + \frac{4}{15}\kappa^2 \left(|Y_i| + \kappa \left\|\Psi(X_i)\right\|_\Theta\right)^4 \left\|\Psi(X_i)\right\|_\Theta^4 + \frac{16}{900}\kappa^4 \left(|Y_i| + \kappa \left\|\Psi(X_i)\right\|_\Theta\right)^6 \left\|\Psi(X_i)\right\|_\Theta^6 \Bigg\}.$$

and so we have the upper bound, as soon as $\lambda \leq N$:

$$V_{\psi,\frac{\lambda}{N},\alpha,3}(\theta,\theta') \leq \mathcal{C}\left\|\theta - \theta'\right\|_\Theta^2.$$

Note that this also proves that:

$$V_{\psi,\frac{\lambda}{N},\alpha,3}(\theta,\theta') \leq 2\kappa^2\mathcal{C}$$

so that the second condition of Theorem 2.22 is satisfied with constant:

$$D = 2\kappa^2\mathcal{C}.$$

Let us choose a finite family $(\theta_1, ..., \theta_m) \in \Theta^m$ and define the submodels, for any $I \subset \{1, ..., m\}$:

$$\Theta_I = span\{\theta_i, i \in I\}.$$

Let us choose a set of parameters $(p_I)_{I\subset\{1,...,m\}}$ such that $p_I \geq 0$ and:

$$\sum_{I\subset\{1,...,m\}} p_I = 1.$$

Definition 5.1. We take $k \in \mathbb{R}_+^*$. From now, we are going to work with the parameter space $B_\Theta(0,k)$. We put, for any $I \subset \{1, ..., m\}$:

$$\hat{\theta}_I = \arg\min_{\theta\in\Theta_I} r_1(l^2, \theta).$$

For any $\delta \leq \kappa$ we put:

$$\tilde{\theta}_I^\delta = \inf\left(1, \frac{\kappa - \delta}{\|\hat{\theta}_I\|_\Theta}\right)\hat{\theta}_I.$$

Note that if δ is small enough and $\hat{\theta}_I$ lies in the interior of $B(0,\kappa)$ then $\tilde{\theta}_I^\delta = \hat{\theta}_I$. However, we are sure that

$$B(\tilde{\theta}_I^\delta, \delta) \subset B(0,\kappa),$$

and this is not the case for $\hat{\theta}_I$ (we will need this later).

Definition 5.2. Let us put:

$$\mathcal{M}_I = \left[f_{\theta_j}(X_i)\right]_{\substack{i\in\{1,...,N\}\\ j\in I}},$$

$$M_I = \frac{1}{N}\mathcal{M}_I'\mathcal{M}_I,$$

and:

$$\mathcal{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}.$$

Moreover for any $d \in \mathbb{N}$ let $\mathcal{I}_d$ denote the identity matrix of size d.

It is well known that $\hat{\theta}_I$ can be expressed with the help of $\mathcal{M}_I$ and $\mathcal{Y}$.

Theorem 5.1. *We have:*

$$\hat{\theta}_I = \sum_{i \in I} \hat{\alpha}_i^I \theta_i,$$

with:

$$(\alpha_i)_{i \in I} = \left(\mathcal{M}_I' \mathcal{M}_I\right)^{-1} \mathcal{M}_I' \mathcal{Y}.$$

Definition 5.3. Let λ^I the image of the Lebesgue measure on $\mathbb{R}^I$ by the application:

$$T_I : \mathbb{R}^I \to \Theta_I$$
$$(\alpha_i)_{i \in I} \mapsto \sum_{i \in I} \alpha_i \theta_i.$$

Moreover, we define, for any $I \subset \{1, ..., m\}$ and $\delta \in (0, \kappa]$, $\tilde{\rho}_I^\delta$ such that:

$$\frac{d\tilde{\rho}_I^\delta}{d\lambda^I}(\theta) = \frac{\mathbb{1}_{B_\Theta(\tilde{\theta}_I^\delta, \delta) \cap \Theta_I}(\theta)}{\lambda^I \left[B_\Theta(\tilde{\theta}_I^\delta, \delta) \cap \Theta_I\right]}.$$

5.2. **Application of Theorem 2.22.**

Theorem 5.2. *For any $\alpha \in (0, 1/2]$, for any $\varepsilon > 0$, for any $\kappa > 0$, for any $(\lambda, \beta_1, \beta_2) \in (\mathbb{R}_+^*)^3$ such that $\beta_1 < \alpha\lambda$ and $\beta_2 < \alpha\lambda$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $I \subset \{1, ..., m\}$, $J \subset \{1, ..., m\}$ and $(\delta_1, \delta_2) \in (0, \kappa]^2$ such that $\hat{\theta}_I = \tilde{\theta}_I^{\delta_1}$ and $\hat{\theta}_J = \tilde{\theta}_J^{\delta_2}$ and that the matrices:*

$$M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right) \mathcal{I}_{|I|}$$

and

$$M_J - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right) \mathcal{I}_{|J|}$$

are definite positive we have:

$$\begin{aligned}
\tilde{\rho}_I^{\delta_1} R(l^2, .) - \tilde{\rho}_J^{\delta_2} R(l^2, .) &\leq b(\alpha, I, J, \delta_1, \delta_2, \lambda, \beta_1, \beta_2, \varepsilon) \\
&= \frac{1}{\alpha} \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \alpha \left[r\left(l^2, \hat{\theta}_I\right) - r\left(l^2, \hat{\theta}_J\right) + \frac{3\mathcal{C}\alpha\lambda}{2N} \left\| \hat{\theta}_I - \hat{\theta}_J \right\|_\Theta^2 \right.\right. \\
&+ \left(\frac{1}{N} \sum_{i=1}^N \|\Psi(X_i)\|_\Theta^2 \right) \left(\left(1 + \frac{1}{\alpha\lambda - \beta_1}\right) \delta_1^2 + \left(1 + \frac{1}{\alpha\lambda - \beta_2}\right) \delta_2^2 \right) \\
&+ \frac{1}{\alpha\lambda - \beta_1} \left(\frac{|I|}{2} \log \frac{1}{\delta_1^2 \beta_1} + \left(\frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2} \right) \delta_1^2 \right. \\
&\left. + \log \frac{\Gamma\left(\frac{|I|}{2} + 1\right)}{\sqrt{\det\left[M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right) \mathcal{I}_{|I|}\right]}} \right) \\
&+ \frac{1}{\alpha\lambda - \beta_2} \left(\frac{|J|}{2} \log \frac{1}{\delta_2^2 \beta_2} + \left(\frac{\alpha\lambda\beta_2\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2} \right) \delta_2^2 \right. \\
&\left.\left. + \log \frac{\Gamma\left(\frac{|J|}{2} + 1\right)}{\sqrt{\det\left[M_J - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right) \mathcal{I}_{|J|}\right]}} \right) \right] \\
&\left. + \left(1 + \frac{\beta_1}{\alpha\lambda - \beta_1} + \frac{\beta_2}{\alpha\lambda - \beta_2}\right) \frac{\log \frac{3p_I p_J}{\varepsilon}}{\lambda} \right\} + \frac{4\lambda^2 C(3)}{N^2} \left(\delta_1^3 + \delta_2^3 + \left\| \hat{\theta}_I - \hat{\theta}_J \right\|_\Theta^3 \right).
\end{aligned}$$

The proof is given at the end of this subsection. This theorem can be used in the same way as Theorem 2.20 (page 50).

A few lines may be helpful to interpret the theorem. The role of α, λ, β_1 and β_2 was discussed in detail previously. The values $\mathcal{C}$, $C(3)$ and κ are constants. The parameter δ_i represents the concentration of the posterior distribution in model i around the estimator. When δ_i is too small, the posterior distribution is too much concentrated around the estimator and can miss the optimal value of $\theta \in \Theta_I$. This is a typical situation of overlearning and the bound explodes. On the contrary, when δ_i is too large, the posterior distribution tends to become the uniform distribution on the model Θ_I. An optimization with respect to δ_i is necessary, and allowed as the theorem is valid uniformly for any δ_i.

Note the order of the bound given by the theorem. The comparison of two submodels Θ_I and Θ_J is interesting when the order of

$$\left\| \hat{\theta}_I - \hat{\theta}_J \right\|_\Theta^2$$

is $1/N$. In this case the optimal order for λ is N, as for β_1 and β_2 but with $\alpha\lambda - \beta_1$ and $\alpha\lambda - \beta_2$ of order N too (for example $\lambda = cN$ and $\beta_1 = \beta_2 = \alpha\lambda/2$, but there is no reason for these particular values to be optimal). Note also that an explicit optimization with respect to δ_1 is possible if we take into account only the first order terms. We obtain:

$$\delta_1^* = \sqrt{\frac{|I|}{\frac{\alpha\lambda-\beta_1+1}{N}\sum_{i=1}^N \|\Psi(X_i)\|_\Theta^2 + \frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}}}.$$

Note however it is possible that this value does not lead to

$$\tilde{\theta}_I^{\delta_1} = \hat{\theta}_I,$$

this forcing us to choose a smaller value. However asymptotically, the optimal order for δ_1 is $\sqrt{|I|/N}$. Let us see what is the order of the bound when we linearize it (we upper bound $\Phi^{-1}_{\lambda/N}$ by the identity) and take $\lambda = cN$, $\beta_1 = \beta_2 = \alpha\lambda/2$ and $\delta_1 = \delta_2 = 1/\sqrt{N}$, with c small enough to have the matrices:

$$M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}$$

and:

$$M_J - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|J|}$$

definite positive.

$$\begin{aligned} b\left(\alpha, I, J, \sqrt{\frac{|I|}{N}}, \sqrt{\frac{|J|}{N}}, cN, \frac{\alpha cN}{2}, \frac{\alpha cN}{2}, \varepsilon\right) &\leq r\left(l^2, \hat{\theta}_I\right) - r\left(l^2, \hat{\theta}_J\right) \\ &+ \frac{3\mathcal{C}\alpha c}{2}\left\|\hat{\theta}_I - \hat{\theta}_J\right\|_\Theta^2 + 4c^2 C(3)\left\|\hat{\theta}_I - \hat{\theta}_J\right\|_\Theta^3 \\ &+ \frac{|I|+|J|}{N}\left\{\left[\frac{1}{N}\sum_{i=1}^N \|\Psi(X_i)\|_\Theta^2\right]\left(1 + \frac{2}{\alpha cN}\right)\right. \\ &\left. + \alpha c\mathcal{C} + \frac{8\kappa c^2 C(3)}{9} - \frac{1}{2}\log\alpha c - \frac{1}{2}\right\} \\ + \frac{1}{N}\left\{\frac{3\log\frac{3p_I p_J}{\varepsilon}}{\alpha c} + \frac{1}{2}\log\frac{\pi|I|}{\det\left[M_I - \left(\frac{c\mathcal{C}}{2} - \frac{16\kappa c^2 C(3)}{9}\right)\mathcal{I}_{|I|}\right]} + \frac{1}{6|I|}\right. & \end{aligned}$$

$$+\frac{1}{2}\log\frac{\pi|J|}{\det\left[M_J-\left(\frac{c\mathcal{C}}{2}-\frac{16\kappa c^2C(3)}{9}\right)\mathcal{I}_{|J|}\right]}+\frac{1}{6|J|}\Bigg\}+4c^2C(3)\frac{|I|^{\frac{3}{2}}+|J|^{\frac{3}{2}}}{N^{\frac{3}{2}}}.$$

Note that an arbitrarily choice like $\alpha=1/2$ leads to the following order of magnitude:

$$r\left(l^2,\hat{\theta}_I\right)-r\left(l^2,\hat{\theta}_J\right)+\mathrm{Cst}_1.\mathrm{Variance}_{I,J}+\mathrm{Cst}_2.\frac{\mathrm{Dim}_{I,J}}{N},$$

where:

$$\mathrm{Variance}_{I,J}=\left\|\hat{\theta}_I-\hat{\theta}_J\right\|_\Theta^2$$

and

$$\mathrm{Dim}_{I,J}=|I|+|J|,$$

and Cst_1 and Cst_2 are random variables of the order of magnitude of a constant.

Let us now give the proof of the theorem.

Proof. Let us choose I and J as subsets of $\{1,...,m\}$. Let us choose π^1 absolutely continuous with respect to λ^I, π^2 absolutely continuous with respect to λ^J with:

$$\frac{d\pi^1}{d\lambda^I}(\theta)=\frac{\mathbb{1}_{B_\Theta(0,\kappa)\cap\Theta_I}(\theta)}{\lambda^I\left[B_\Theta(0,\kappa)\cap\Theta_I\right]}$$

and:

$$\frac{d\pi^2}{d\lambda^J}(\theta)=\frac{\mathbb{1}_{B_\Theta(0,\kappa)\cap\Theta_J}(\theta)}{\lambda^J\left[B_\Theta(0,\kappa)\cap\Theta_J\right]}.$$

We apply Theorem 2.22 with $s=3$, $\gamma_1=\gamma_2=\lambda$ and a given $\alpha\leq 1/2$. We obtain that for any $\varepsilon>0$, for any $(\lambda,\beta_1,\beta_2)\in(0,N)^3$ such that $\beta_1<\alpha\lambda$ and $\beta_2<\alpha\lambda$, with P-probability at least $1-\varepsilon$, for any $I\subset\{1,...,m\}$ and $J\subset\{1,...,m\}$, for any $(\rho^1,\rho^2)\in\mathcal{M}_+^1(\Theta_I)\times\mathcal{M}_+^1(\Theta_J)$:

$$\begin{aligned}
&\rho^1R(l^2,.)-\rho^2R(l^2,.)\\
&\quad\leq\frac{1}{\alpha}\Phi^{-1}_{\frac{\lambda}{N}}\Bigg\{\alpha\Bigg[\rho^1r(l^2,.)-\rho^2r(l^2,.)+\frac{\alpha\lambda}{2N}\rho^1\otimes\rho^2V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\\
&\qquad+\frac{1}{\alpha\lambda-\beta_1}\log\pi^1_{\exp[-\beta_1r(l^2,d\theta')]}\exp\Bigg(\frac{\alpha\lambda\beta_1}{2N}\rho^1_{d\theta}V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\\
&\qquad\qquad+\frac{4\beta_1\lambda^2}{9N^2}\rho^1_{d\theta}M_{l^2,3}(\theta,\theta')\Bigg)\\
&\qquad\frac{1}{\alpha\lambda-\beta_2}\log\pi^2_{\exp[-\beta_2r(l^2,d\theta')]}\exp\Bigg(\frac{\alpha\lambda\beta_2}{2N}\rho^2_{d\theta}V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\\
&\qquad\qquad+\frac{4\beta_2\lambda^2}{9N^2}\rho^2_{d\theta}M_{l^2,3}(\theta,\theta')\Bigg)\\
&\quad+\frac{1}{\alpha\lambda-\beta_1}\mathcal{K}\left(\rho^1,\pi^1_{\exp[-\beta_1r(l^2,.)]}\right)+\frac{1}{\alpha\lambda-\beta_2}\mathcal{K}\left(\rho^2,\pi^2_{\exp[-\beta_2r(l^2,.)]}\right)\Bigg]\\
&\qquad+\left(1+\frac{\beta_1}{\alpha\lambda-\beta_1}+\frac{\beta_2}{\alpha\lambda-\beta_2}\right)\frac{\log\frac{3}{\varepsilon}}{\lambda}\Bigg\}+\frac{4\lambda^2}{9N^2}\rho^1\otimes\rho^2M_{l^2,s}(\theta,\theta').
\end{aligned}$$

Now, we know that the optimal posteriors are Gibbs distributions, but we make the following choice in order to be obtain explicit computations. We choose $(\delta_1,\delta_2)\in(\mathbb{R}_+^*)^2$, and $\rho^1=\tilde{\rho}_I^{\delta_1}$, $\rho^2=\tilde{\rho}_J^{\delta_2}$. Note that we have:

$$\rho^1 r(l^2,\theta) = \rho^1 \frac{1}{N}\sum_{i=1}^{N}\left[\left(Y_i - f_{\tilde{\theta}_I^{\delta_1}}(X_i)\right) + \left\langle \tilde{\theta}_I^{\delta_1} - \theta, \Psi(X_i)\right\rangle_\Theta\right]^2$$
$$\leq r\left(l^2,\tilde{\theta}_I^{\delta_1}\right) + \delta_1^2\left[\frac{1}{N}\sum_{i=1}^{N}\|\Psi(X_i)\|_\Theta^2\right].$$

Moreover, for the variance term we have, using the bound we proved in the beginning of the section:

$$\rho^1_{(d\theta)}\otimes\rho^2_{(d\theta')}\left[V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\right] \leq \mathcal{C}\rho^1_{(d\theta)}\otimes\rho^2_{(d\theta')}\|\theta-\theta'\|_\Theta^2$$
$$\leq 3\mathcal{C}\rho^1_{(d\theta)}\otimes\rho^2_{(d\theta')}\left[\left\|\theta-\tilde{\theta}_I^{\delta_1}\right\|_\Theta^2 + \left\|\tilde{\theta}_I^{\delta_1}-\tilde{\theta}_J^{\delta_2}\right\|_\Theta^2 + \left\|\tilde{\theta}_J^{\delta_2}-\theta'\right\|_\Theta^2\right]$$
$$\leq 3\mathcal{C}\left[\delta_1^2+\delta_2^2+\left\|\tilde{\theta}_I^{\delta_1}-\tilde{\theta}_J^{\delta_2}\right\|_\Theta^2\right]$$

and in the same way:

$$\rho^1_{(d\theta)}\otimes\rho^2_{(d\theta')}\left[M_{l^2,3}(\theta,\theta')\right] \leq C(3)\|\theta-\theta'\|_\Theta^3 \leq 9C(3)\left[\left\|\tilde{\theta}_I^{\delta_1}-\tilde{\theta}_J^{\delta_2}\right\|_\Theta^3 + \delta_1^3+\delta_2^3\right].$$

For the entropy term we have:

$$\mathcal{K}\left(\rho^1,\pi^1_{\exp[-\beta_1 r(l^2,.)]}\right) = \log\left[\left(\frac{\kappa}{\delta_1}\right)^{|I|}\right] + \beta_1\rho^1 r(l^2,.)$$
$$+\log\lambda^I\left\{\exp\left[-\beta_1 r(l^2,.)\right]\mathbb{1}_{B(0,\kappa)}(.)\right\} + \log\frac{\Gamma\left(\frac{|I|}{2}+1\right)}{\pi^{\frac{|I|}{1}}\kappa^{|I|}}$$

and we already noted that:

$$\beta_1\rho^1 r(l^2,.) \leq \beta_1\left\{r\left(l^2,\tilde{\theta}_I^{\delta_1}\right) + \delta_1^2\left[\frac{1}{N}\sum_{i=1}^{N}\|\Psi(X_i)\|_\Theta^2\right]\right\}.$$

Finally, we have to compute:

$$\log\pi^1_{\exp[-\beta_1 r(l^2,d\theta')]}\exp\left[\frac{\alpha\lambda\beta_1}{2N}\rho^1_{d\theta}V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\right.$$
$$\left.+\frac{4\beta_1\lambda^2}{9N^2}\rho^1_{d\theta}M_{l^2,3}(\theta,\theta')\right] \leq \log\lambda^I\exp\left[-\beta_1 r(l^2,\theta')\right.$$
$$\left.+\frac{\alpha\lambda\beta_1}{2N}\rho^1_{d\theta}V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta') + \frac{4\beta_1\lambda^2}{9N^2}\rho^1_{d\theta}M_{l^2,3}(\theta,\theta')\right]$$
$$-\log\lambda^I_{d\theta'}\left\{\exp\left[-\beta_1 r(l^2,\theta')\right]\mathbb{1}_{B(0,\kappa)}(.)\right\}.$$

Note that this last term vanishes with the analogous term in the expansion of the entropy term, while we have:

$$\log\lambda^I_{d\theta'}\exp\left[-\beta_1 r(l^2,\theta') + \frac{\alpha\lambda\beta_1}{2N}\rho^1_{d\theta}V_{l^2,\frac{\lambda}{N},\alpha,3}(\theta,\theta')\right.$$
$$\left.+\frac{4\beta_1\lambda^2}{9N^2}\rho^1_{d\theta}M_{l^2,3}(\theta,\theta')\right] \leq \log\lambda^I\exp\left[-\beta_1 r(l^2,\theta')\right.$$
$$\left.+\frac{\alpha\lambda\beta_1\mathcal{C}}{2N}\rho^1_{d\theta}\|\theta-\theta'\|_\Theta^2 + \frac{8\kappa\beta_1\lambda^2 C(3)}{9N^2}\rho^1_{d\theta}\|\theta-\theta'\|_\Theta^2\right]$$

$$\leq \log \lambda^I_{d\theta'} \exp\left[-\beta_1 r(l^2,\theta') + \frac{\alpha\lambda\beta_1\mathcal{C}}{N}\left\|\tilde{\theta}^{\delta_1}_I - \theta'\right\|^2_\Theta \right.$$
$$\left. + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}\left\|\tilde{\theta}^{\delta_1}_I - \theta'\right\|^2_\Theta\right] + \left(\frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}\right)\delta_1^2$$

And we have:

$$\log \lambda^I_{d\theta'} \exp\left[-\beta_1 r(l^2,\theta') + \frac{\alpha\lambda\beta_1\mathcal{C}}{N}\left\|\tilde{\theta}^{\delta_1}_I - \theta'\right\|^2_\Theta \right.$$
$$\left. + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}\left\|\tilde{\theta}^{\delta_1}_I - \theta'\right\|^2_\Theta\right] = -\beta_1 r\left(l^2,\tilde{\theta}^{\delta_1}_I\right)$$
$$+ \frac{1}{2}\log\frac{\pi^{|I|}}{\beta_1^{|I|}\det\left[M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}\right]}$$

as soon as we have $\hat{\theta}_I = \tilde{\theta}^{\delta_1}_I$ and the symmetric matrix in the det is definite positive. We end the proof by a union bound over all possibles values for I and J with weights p_I and p_J. □

5.3. **Extension to data-dependent models.** It is of course possible to allow the family $(\theta_1,...,\theta_m)$ to depend of the data, using transductive bounds and integrating over the test sample. So from now we use the transductive point of view, and so we assume that $(Z_1,...,Z_{(k+1)N})$ is drawn from $P_{(k+1)N}$. We keep the notations introduced for the linear regression case, but we allow every θ_i to be an exchangeable function of the data:

$$\forall i \in \{1,...,m\}, \quad \theta_i = F(\{Z_1,...,Z_{(k+1)N}\}).$$

By analogous computations we obtain the following result.

Theorem 5.3. *For any $k \in \mathbb{N}\setminus\{0,1\}$, for any $\alpha \in (0,1/2]$, for any $\varepsilon > 0$, for any $\kappa > 0$, for any $(\lambda,\beta_1,\beta_2) \in (\mathbb{R}^*_+)^3$ such that $\beta_1 < \alpha\lambda$ and $\beta_2 < \alpha\lambda$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $I \subset \{1,...,m\}$, $J \subset \{1,...,m\}$ and $(\delta_1,\delta_2) \in (0,\kappa]^2$ such that $\hat{\theta}_I = \tilde{\theta}^{\delta_1}_I$ and $\hat{\theta}_J = \tilde{\theta}^{\delta_2}_J$ and that the matrices:*

$$Mat(1) = M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa k\lambda^2 C(3)}{9(k+1)N^2} - \frac{8\kappa\mathcal{E}\lambda^2}{27(k+1)N^2}\right)\mathcal{I}_{|I|}$$

and

$$Mat(2) = M_J - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa k\lambda^2 C(3)}{9(k+1)N^2} - \frac{8\kappa\mathcal{E}\lambda^2}{27(k+1)N^2}\right)\mathcal{I}_{|J|}$$

are definite positive we have:

$$\tilde{\rho}^{\delta_1}_I R(l^2,.) - \tilde{\rho}^{\delta_2}_J R(l^2,.) \leq b(\alpha,I,J,\delta_1,\delta_2,\lambda,\beta_1,\beta_2,\varepsilon)$$
$$= r\left(l^2,\hat{\theta}_I\right) - r\left(l^2,\hat{\theta}_J\right) + \frac{3(k+1)\mathcal{C}\alpha\lambda}{2kN}\left\|\hat{\theta}_I - \hat{\theta}_J\right\|^2_\Theta$$
$$+ \left(\frac{1}{N}\sum_{i=1}^N \|\Psi(X_i)\|^2_\Theta\right)\left(\left(1 + \frac{k+1}{k(\alpha\lambda-\beta_1)}\right)\delta_1^2 + \left(1 + \frac{k+1}{k(\alpha\lambda-\beta_2)}\right)\delta_2^2\right)$$
$$+ \frac{k+1}{k(\alpha\lambda-\beta_1)}\left(\frac{|I|}{2}\log\frac{1}{\delta_1^2\beta_1}\right.$$
$$\left. + \left(\frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa k\beta_1\lambda^2 C(3)}{9(k+1)N^2} + \frac{8\kappa\mathcal{E}\beta_1\lambda^2}{27(k+1)N^2}\right)\delta_1^2\right.$$

$$- \frac{1}{2} \log \det[Mat(1)] + \log \Gamma \left(\frac{|I|}{2} + 1 \right) \Bigg)$$

$$+ \frac{k+1}{k(\alpha\lambda - \beta_2)} \Bigg(\frac{|J|}{2} \log \frac{1}{\delta_2^2 \beta_2}$$

$$+ \left(\frac{\alpha\lambda\beta_2\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2} + \frac{8\kappa\mathcal{E}\beta_2\lambda^2}{27(k+1)N^2} \right) \delta_2^2$$

$$- \frac{1}{2} \log \det[Mat(2)] + \log \Gamma \left(\frac{|J|}{2} + 1 \right) \Bigg)$$

$$+ \left(1 + \frac{\beta_1}{\alpha\lambda - \beta_1} + \frac{\beta_2}{\alpha\lambda - \beta_2} \right) \frac{k \log \frac{3p_I p_J}{\varepsilon}}{(k+1)\alpha\lambda} + \frac{4\lambda^2 C(3)}{N^2} \left(\delta_1^3 + \delta_2^3 + \left\| \hat{\theta}_I - \hat{\theta}_J \right\|_\Theta^3 \right)$$

$$+ \frac{3\mathcal{E}}{ks} \left(\frac{(s-1)\lambda}{sN} \right)^{s-1} \left[\delta_1^3 + \delta_2^3 + \left\| \hat{\theta}_I - \hat{\theta}_J \right\|_\Theta^3 \right].,$$

where:

$$\mathcal{E} = \frac{1}{N} \sum_{i=1}^{N} \|\Psi(X_i)\|_\Theta^3 \left[4Y_i^3 + 32\kappa^3 \|\Psi(X_i)\|_\Theta^3 \right].$$

Proof. We apply Theorem 3.10 in the same context. □

Example 5.1 (Support Vector Machines)**.** We choose $m = (k+1)N$ and want to obtain:

$$\{\Psi(X_1), ..., \Psi(X_{(k+1)N})\} = \{\theta_1, ..., \theta_{(k+1)N}\}.$$

In order to do this, let us choose a complete order on $\mathcal{H}$, say $\leq_{\mathcal{H}}$, and define $(\theta_1, ..., \theta_{(k+1)N})$ as $(\Psi(X_1), ..., \Psi(X_{(k+1)N}))$ reordered under $\leq_{\mathcal{H}}$. For a given I we are working with functions of the form:

$$\sum_{i \in I} \alpha_i \langle \Psi(X_i), \Psi(.) \rangle .$$

Let us put $K(x, x') = \langle \Psi(x), \Psi(x') \rangle$, we can see that the functions we work with are under the form of SVM:

$$\sum_{i \in I} \alpha_i K(X_i, .).$$

Here, we can choose a maximal size H for I and take, for any I such that $|I| \leq H$:

$$p_I = \frac{1}{H} \frac{1}{\binom{(k+1)N}{|I|}}$$

and 0 for any other I. Note that of course, the posterior $\tilde{\rho}_I^{\delta_1}$ is observable only for the models containing $\Psi(X_i)$ with $i \leq N$.

5.4. **A bound for a single model.** Here, we give another bound in the linear case that will we usefull in the second part of the thesis.

Definition 5.4. Let us put:

$$\overline{\theta}_I = \arg \min_{\theta \in \Theta_I} R(\theta).$$

We follow the proof of Theorem 5.2 but restrict the second model to $\{\overline{\theta}_I\}$, that means that instead of π^2 and ρ^2 we simply take $\delta_{\overline{\theta}_I}$ (and $\beta_2 = 0$). We obtain the following result.

For any $\alpha \in (0,1/2]$, for any $\varepsilon > 0$, for any $\kappa > 0$, for any $(\lambda, \beta_1) \in (\mathbb{R}_+^*)^2$ such that $\beta_1 < \alpha\lambda$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $I \subset \{1,...,m\}$, and $\delta_1 \in (0,\kappa]^2$ such that $\hat{\theta}_I = \tilde{\theta}_I^{\delta_1}$ and

$$M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}$$

is definite positive we have:

$$\begin{aligned}
&R(l^2,\hat{\theta}_I) - R(l^2,\overline{\theta}_I) \\
&\qquad \leq r(l^2,\hat{\theta}_I) - r(l^2,\overline{\theta}_I) + \frac{3\mathcal{C}\alpha\lambda}{2N}\left\|\hat{\theta_I} - \overline{\theta}_I\right\|_\Theta^2 \\
&\qquad + \left(\frac{1}{N}\sum_{i=1}^N \|\Psi(X_i)\|_\Theta^2\right)\left(1 + \frac{1}{\alpha\lambda - \beta_1}\right)\delta_1^2 \\
&\quad + \frac{1}{\alpha\lambda - \beta_1}\left(\frac{|I|}{2}\log\frac{1}{\delta_1^2\beta_1} + \left(\frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}\right)\delta_1^2\right. \\
&\qquad \left. + \log\frac{\Gamma\left(\frac{|I|}{2}+1\right)}{\sqrt{\det\left[M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}\right]}}\right) \\
&\qquad + \left(1 + \frac{\beta_1}{\alpha\lambda - \beta_1}\right)\frac{\log\frac{3p_I}{\varepsilon}}{\alpha\lambda} + \frac{4\lambda^2 C(3)}{N^2}\left(\delta_1^3 + \left\|\hat{\theta}_I - \overline{\theta}_I\right\|_\Theta^3\right).
\end{aligned}$$

Now, note that:

$$\left\|\hat{\theta_I} - \overline{\theta}_I\right\|_\Theta^3 \leq 2\kappa\left\|\hat{\theta_I} - \overline{\theta}_I\right\|_\Theta^2$$

and that:

$$\left\|\hat{\theta_I} - \overline{\theta}_I\right\|_\Theta^2 \leq C_I\left[R(l^2,\hat{\theta}_I) - R(l^2,\overline{\theta}_I)\right]$$

where:

$$C_I = \sup_{\|\theta\|=1}\frac{1}{P\left[\langle\theta, X\rangle^2\right]}.$$

In the second part of the thesis, at some point we assume that we know the marginal distribution of X and so C_I is known to the statistician. This implies the following theorem.

Theorem 5.4. *For any $\alpha \in (0,1/2]$, for any $\varepsilon > 0$, for any $\kappa > 0$, for any $(\lambda,\beta_1) \in (\mathbb{R}_+^*)^2$ such that $\beta_1 < \alpha\lambda$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $I \subset \{1,...,m\}$, and $\delta_1 \in (0,\kappa]^2$ such that $\hat{\theta}_I = \tilde{\theta}_I^{\delta_1}$,*

$$1 - \frac{3\mathcal{C}\alpha\lambda C_I}{2N} - \frac{8\kappa\lambda^2 C(3)C_I}{N^2} > 0$$

and

$$M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}$$

is definite positive we have:

$$\begin{aligned}
&R(l^2,\hat{\theta}_I) - R(l^2,\overline{\theta}_I) \\
&\leq \frac{1}{1 - \frac{3\mathcal{C}\alpha\lambda C_I}{2N} - \frac{8\kappa\lambda^2 C(3)C_I}{N^2}}\left\{\left(\frac{1}{N}\sum_{i=1}^N\|\Psi(X_i)\|_\Theta^2\right)\left(1 + \frac{1}{\alpha\lambda - \beta_1}\right)\delta_1^2\right. \\
&\quad \left. + \frac{1}{\alpha\lambda - \beta_1}\left(\frac{|I|}{2}\log\frac{1}{\delta_1^2\beta_1} + \left(\frac{\alpha\lambda\beta_1\mathcal{C}}{N} + \frac{16\kappa\beta_1\lambda^2 C(3)}{9N^2}\right)\delta_1^2\right.\right.
\end{aligned}$$

$$+ \log \frac{\Gamma\left(\frac{|I|}{2}+1\right)}{\sqrt{\det\left[M_I - \left(\frac{\alpha\lambda\mathcal{C}}{N} - \frac{16\kappa\lambda^2 C(3)}{9N^2}\right)\mathcal{I}_{|I|}\right]}}\Bigg)$$

$$+ \left(1 + \frac{\beta_1}{\alpha\lambda - \beta_1}\right)\frac{\log\frac{3p_I}{\varepsilon}}{\alpha\lambda} + \frac{4\lambda^2 C(3)}{N^2}\delta_1^3\Bigg\}.$$

Part 2. Iterative feature selection in least square regression estimation

In this part, we focus on regression estimation in both the inductive and the transductive case. We assume that we are given a set of features (which can be a base of functions, but not necessarily). We begin by giving a deviation inequality on the risk of an estimator in every model defined by using a single feature. These models are too simple to be useful by themselves, but we then show how this result motivates an iterative algorithm that performs feature selection in order to build a suitable estimator. We prove that every selected feature actually improves the performance of the estimator. We give all the estimators and results at first in the inductive case, which requires the knowledge of the distribution of the design, and then in the transductive case, in which we do not need to know this distribution.

6. The setting of the problem

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

6.1. **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 $\mathbb{R}$.

6.1.1. *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 $i \in \{1, ..., N\}$. Our objective is then to estimate the regression function on the basis of the observations.

Definition 6.1 (The regression function)**.** We denote:

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

6.1.2. *The transductive setting.* In the transductive case, we assume that P_{2N} is some exchangeable probability measure on the space $\left((\mathcal{X} \times \mathbb{R})^{2N}, (\mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})^{\otimes 2N}\right)$. We will write $(X_i, Y_i)_{i=1...2N} = (Z_i)_{i=1...2N}$ a random vector distributed according to P_{2N}.

Definition 6.2 (Exchangeable probability distribution)**.** Let $\mathfrak{S}_k$ denote the set of all permutations of $\{1, ..., k\}$. We say that P_{2N} is exchangeable if for any $\sigma \in \mathfrak{S}_{2N}$ we have: $(X_{\sigma(i)}, Y_{\sigma(i)})_{i=1...2N}$ has the same distribution under P_{2N} that $(X_i, Y_i)_{i=1...2N}$.

We assume that we observe $(X_i, Y_i)_{i=1...N}$ and $(X_i)_{i=N+1...2N}$; $(X_i, Y_i)_{i=1...N}$ is usually called the training sample and $(X_i, Y_i)_{i=N+1...2N}$ the test sample. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1...2N}$. This is why Vapnik [41] called this kind of inference "transductive inference".

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

6.2. **The model.** In both settings, we are going to use the same model to estimate the regression function. Let Θ be a vector space, and:

$$\begin{aligned} F : \Theta \times \mathcal{X} &\to \mathbb{R} \\ (\theta, x) &\mapsto F(\theta, x) = f_\theta(x) \end{aligned}$$

be such that, for any $x_0 \in \mathcal{X}$, the application $\theta \mapsto f_\theta(x_0)$ is linear. We define the model:

$$\mathcal{F} = \{f_\theta(.), \theta \in \Theta\}.$$

Remark that we do not assume that f belongs to $\mathcal{F}$.

6.3. **Presentation of the results.** In both settings, we give a concentration inequality on the risk of estimators in unidimensional models of the form:

$$\{\alpha\theta, \alpha \in \mathbb{R}\}$$

for a given θ.

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, it means that we estimate $f(.)$ by a function $\hat{f} \in \mathcal{F}$, 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, the estimation of $Y_{N+1}, ..., Y_{2N}$ can be performed by the procedure without any prior knowledge about the marginal distribution of X under P. We also give in this case some generalizations (like the case where the test sample has a different size).

We then briefly show that the technique used to obtain bounds in models of dimension 1 can also be used in more general models.

In a last section, we come back to the assertion that in our method, "every selected feature actually improves the current estimator" and show how this can be interpreted as an oracle inequality.

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

Hypothesis. In all this section, we assume that $\mathcal{F}$ and P are such that:

$$\forall \theta \in \Theta, P \exp\left[f_\theta(X)Y\right] < +\infty.$$

7.1. **Notations.** For any random variable T we put:

$$V(T) = P\left[(T - PT)^2\right]$$
$$M^3(T) = P\left[(T - PT)^3\right],$$

and we define for any $\gamma \geq 0$:

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

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

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

We give now notations that are specific to the inductive setting.

Definition 7.1. We put:

$$R(\theta) = P\left[(Y - f_\theta(X))^2\right]$$
$$r(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - f_\theta(X_i)\right)^2,$$

and in this setting, our objective is $f_{\overline{\theta}}$ where:

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

Now, we suppose that we are given a finite family of m vectors:

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

We are going to use the family Θ_0 to estimate the function f, the estimator will be under the form:

$$\hat{f}(x) = \sum_{k=1}^{m} \alpha_k f_{\theta_k}(x),$$

where every α_k will depend on the observations $Z_1, ..., Z_N$. We can think of Θ_0 as a basis of Θ, but actually there is no other assumption about Θ_0 than finiteness.

Every θ_k defines a unidimensional submodel of $\mathcal{F}$:

$$\{f_{\alpha\theta_k}(.), \alpha \in \mathbb{R}\} = \{\alpha f_{\theta_k}(.), \alpha \in \mathbb{R}\}.$$

In a first step, we are going to work on each of these submodels individually. So let us put, for any $k \in \{1, ..., m\}$:

$$\begin{aligned}
\overline{\alpha}_k &= \arg\min_{\alpha\in\mathbb{R}} R(\alpha\theta_k) = \frac{P[f_{\theta_k}(X)Y]}{P[f_{\theta_k}(X)^2]} \\
\hat{\alpha}_k &= \arg\min_{\alpha\in\mathbb{R}} r(\alpha\theta_k) = \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)Y_i}{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2} \\
\mathcal{C}_k &= \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2}{P[f_{\theta_k}(X)^2]}.
\end{aligned}$$

7.2. **Main result.** The following theorem gives a control of the excess risk of an estimator in the model $\{f_{\alpha\theta_k}(.), \alpha \in \mathbb{R}\}$ for each k. This estimator is not the usual least square estimator $\hat{\alpha}_k$ but $\mathcal{C}_k\hat{\alpha}_k$.

Theorem 7.1. *Let us put:*

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

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

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

where we have:

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

with:

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

The proof of the theorem is given at the end of this section, let us first show how we can use it in order to build an estimator under the form:

$$\hat{f}(.) = \sum_{k=1}^{m} \alpha_k f_{\theta_k}(.).$$

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

$$C_N(P,m,\varepsilon,\theta_k) \xrightarrow[N\to+\infty]{} \frac{\sqrt{2}\left[M^3\left(W_{\theta_k}\right)\right]^2}{9V\left(W_{\theta_k}\right)^{\frac{5}{2}} P\left[f_{\theta_k}(X)^2\right]}.$$

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 f_{θ_k} is bounded by C'_k then W_{θ_k} is bounded by $C_k = 2C_Y C'_k$ and we have (basically):

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

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

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

The following theorem justifies this method.

Theorem 7.2. *If we assume that there is a constant c such that:*

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

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

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

where we have:

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

and:

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

and:

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

7.3. **Application to regression estimation.**

7.3.1. *Interpretation of Theorems 7.1 and 7.2 in terms of confidence intervals.*

Definition 7.2. Let us put, for any $(\theta, \theta') \in \Theta^2$:

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

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

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

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:

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

where $\beta(\varepsilon, k)$ is the bound given by Theorem 7.1 or more likely by Theorem 7.2.

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 $\langle ., . \rangle_P$:

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

Definition 7.3. We define, for any k and ε:

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

Then the theorem is equivalent to the following corollary.

Corollary 7.3. *We have:*

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

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

Definition 7.4. We write $\Pi_P^{k,\varepsilon}$ the orthogonal projection into $\mathcal{CR}(k, \varepsilon)$ with respect to the distance d_P.

7.3.2. *The algorithm.* The previous formulation of Theorem 7.1 motivates the following iterative algorithm:

- choose $\theta(0) \in \Theta$, for example, $\theta(0) = 0$;
- at step $n \in \mathbb{N}^*$, we have: $\theta(0), ..., \theta(n-1)$. Choose $k(n) \in \{1, ..., m\}$ (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 7.5. Let n_0 denote the stopping step, and:

$$\hat{f}(.) = f_{\theta(n_0)}(.)$$

the corresponding function.

FIGURE 1. Detailed version of the feature selection algorithm.

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

Repeat:

- set $n \leftarrow n + 1$;
- set $best\_improvement \leftarrow 0$ and $\theta(n) \leftarrow \theta(n-1)$;
- for $k \in \{1, ..., m\}$, compute:

$$v_k = P\left[f_{\theta_k}(X)^2\right],$$

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

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

  and if $\delta_k > best\_improvement$, set:

$$best\_improvement \leftarrow \delta_k,$$

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

- if $best\_improvement > 0$:

$$\theta_{k(n)}(n) \leftarrow \theta_{k(n)}(n) + sgn(\gamma_k) \Big(|\gamma_k| - \beta(\varepsilon, k) \Big)_+;$$

until $best\_improvement < \kappa$ (where $sgn(x) = -1$ if $x \leq 0$ and 1 otherwise).

Return the estimator:

$$\hat{f}(.) = \sum_{k=1}^{m} \theta_k(n) f_{\theta_k}(.).$$

7.3.3. *Results and comments on the algorithm.*

Theorem 7.4. *We have:*

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

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

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

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

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

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:

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

that can be written:

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

□

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 $k \in \{1, ..., m\}$ (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 7.4 motivates the choice:

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

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

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

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

$$k(1) = 1, ..., k(m) = m, k(m+1) = 1, ..., k(2m) = m, ...$$

Such a procedure could look similar to the famous Widrow-Hoff algorithm [44] (also known as ADALINE), which estimates the function $f(.)$ by an estimator under the form:

$$\sum_{k=1}^{m} \alpha_k f_{\theta_k}(.),$$

and updates the α_k sequentially by a gradient descent strategy. Actually, there are two major differences: first, the gradient descent requires the calibration of a parameter $\eta > 0$, that is avoided here, then, ADALINE is only a way to compute the usual least square estimator, and has absolutely no guarantees against overlearning if the family $\{f_{\theta_1}, ..., f_{\theta_m}\}$ is too large.

Example 7.1. 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 and we simply take, for any x and θ:

$$f_\theta(x) = \theta(x).$$

The choice of m should not be a problem, the algorithm avoiding itself overlearning we can take a large value of m like $m = N$ (see later for more details). In this setting, the algorithm is a procedure for (soft) thresholding of coefficients. In the particular case of a wavelets basis, see Kerkyacharian and Picard [24] or Donoho and Johnstone [18] 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:

$$k(1) = 1, ..., k(m) = m,$$

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

Actually, in the case given in the example, 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:

$$N^{\frac{-2\beta}{2\beta+1}} \log N.$$

We prove this point in the last section of this part.

7.4. **An extension to the case of Support Vector Machines.** Thanks to a method due to Seeger [37], it is possible to extend this method to the case where the set Θ_0 is data dependent in the following way:

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

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

$$|\Theta_0(Z_1, ..., Z_N, N)| \leq N \, |\Theta_0(Z_i, N)| = N m'(N).$$

We put:

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

In this case, we need some adaptations of our previous notations. We put, for $i \in \{1, ..., N\}$:

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

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

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

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

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

Theorem 7.5. *We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $k \in \{1, ..., m'(N)\}$ and $i \in \{1, ..., N\}$:*

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

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

Example 7.2. Let us consider the case where Θ is a Hilbert space with scalar product $\langle ., . \rangle$, and:

$$f_\theta(x) = \langle \theta, \Psi(x) \rangle$$

for any $\theta \in \Theta$ and $x \in \mathcal{X}$, where Ψ is an application $\mathcal{X} \to \Theta$. Let us put $\Theta_0[(x, y), N] = \{\Psi(x)\}$. In this case we have $m'(N) = 1$ and:

$$\hat{f}(.) = \sum_{i=1}^{N} \alpha_{i,1} \left\langle \Psi(X_i), \Psi(.) \right\rangle.$$

Let us define,

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

the function K is called the kernel, and:

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

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

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

SVM where first introduced by Boser, Guyon and Vapnik [7] in the context of classification, and then generalized by Vapnik [41] to the context of regression estimation. For a general introduction to SVM, see also Cristianini and Shawe-Taylor [16] and Catoni [10].

Example 7.3. A widely used kernel is the Gaussian kernel:

$$K_\gamma(x, x') = \exp\left(-\gamma \frac{d^2(x, x')}{2}\right),$$

where $d(.,.)$ 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 $\mathcal{X} \to \mathbb{R}$, and for any x and θ:

$$f_\theta(x) = \theta(x).$$

Now, let us put:

$$\Theta_0[(x, y), N] = \left\{K_2(x, .), K_{2^2}(X, .), ..., K_{2^{m'(N)}}(x, .)\right\}.$$

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

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

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). For example, in order to imitate the oscillation of wavelets, we can introduce a more sophisticated SVM estimator, based on the kernel family:

$$K_{\gamma,\gamma'}(x, x') = \exp\left(-2^{2\gamma}(x - x')^2\right) \cos\left(2^{\gamma+\gamma'-1}(x - x')\right)$$

for $\gamma \in \{1, ..., m_1\}$, $\gamma' \in \{1, ..., m_2\}$ (note that $m'(N) = m_1 m_2$).

7.5. **Proof of the theorems.** In a first time, we prove a lemma that is the basis of proofs of Theorems 7.1 and 7.5.

Lemma 7.6. *We have, for any $\theta \in \Theta$, $\gamma > 0$ and $\eta \geq 0$:*

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

and

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

Proof. For the first equality, we write:

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

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

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

by the first equality of Lemma 7.6. By the same way, using the reverse inequality we obtain:

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

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

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

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:

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

We obtain:

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

and so:

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

Now, we put:

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

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

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

For short, we take the notation of the theorem:

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

Now, dividing both sides by:

$$P\left[f_{\theta_k}(X)^2\right]$$

we obtain:

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

In order to conclude, just remark that:

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

□

Proof of Theorem 7.2. Remark that, for any $\theta \in \Theta$:

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

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 7.6. For any $\theta \in \Theta$:

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

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

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

where:

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

Taking:

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

and:

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

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

$$P\left(W_{\theta_k}^2\right) \leq \frac{1}{N}\sum_{i=1}^{N} Y_i^2 f_{\theta_k}(X_i)^2 + \sqrt{\frac{2V\left(W_{\theta_k}^2\right)\log\frac{2m}{\varepsilon}}{N}} + \frac{\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}^2\right)} J_{\theta_k}\left(\sqrt{\frac{2\log\frac{2m}{\varepsilon}}{NV\left(W_{\theta_k}^2\right)}}\right) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 f_{\theta_k}(X_i)^2 + \mathcal{A}_k \tag{7.1}$$

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

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

Remember that in the proof of Theorem 7.1 we got the upper bound, with probability at least $1-\frac{\varepsilon}{2}$, for any k:

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

that gives:

$$\begin{aligned}
-P\left(W_{\theta_k}\right)^2 \leq &-\left(\frac{1}{N}\sum_{i=1}^{N} Y_i f_{\theta_k}(X_i)\right)^2 \\
&+\left\{\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}} + \frac{\log^{\frac{5}{2}}\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3} I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2\right\} \\
&\left\{2\left|\frac{1}{N}\sum_{i=1}^{N} Y_i f_{\theta_k}(X_i)\right| + \right. \\
&\left.\sqrt{\frac{2V\left(W_{\theta_k}\right)\log\frac{4m}{\varepsilon}}{N}} + \frac{\log^{\frac{5}{2}}\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)^3} I_{\theta_k}\left(\sqrt{\frac{2\log\frac{4m}{\varepsilon}}{NV\left(W_{\theta_k}\right)}}\right)^2\right\}
\end{aligned} \tag{7.2}$$

$$= - \left(\frac{1}{N} \sum_{i=1}^{N} Y_i f_\theta(X_i) \right)^2 + \mathcal{B}_k.$$

for short. Let us combine inequalities 7.1 and 7.2. We obtain that, with probability at least $1 - \varepsilon$, for every k we have:

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

□

Proof of Theorem 7.5. This proof is a variant of the proof of Theorem 7.1, the method it uses is due to Seeger [37]. Let us define, for any $i \in \{1, ..., N\}$:

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

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

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

by the first equality of Lemma 7.6. In the same way, we obtain the reverse inequality and, combining both results, for any $(i,k) \in \{1, ..., N\} \times \{1, ..., m'(N)\}$, for any $\lambda_{i,k} > 0$ and $\eta_{i,k} \geq 0$:

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

where:

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

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

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

We obtain:

$$P^{\otimes N} \sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} \exp\left\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j f_{\theta_{i,k}}(X_j) - P\left(Y f_{\theta_{i,k}}(X)\right) \right| \right.$$

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

$$= P^{\otimes N} \sum_{i=1}^{N} \sum_{k'=1}^{m'(N)} P_i \exp\Bigg\{ \lambda_{i,k} \left| \frac{1}{N-1} \sum_{j \neq i} Y_j f_{\theta_{i,k}}(X_j) - P\left(Y f_{\theta_{i,k}}(X)\right) \right|$$

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

Now, we put:

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

and achieve the proof exactly as for Theorem 7.1. □

8. Simulations in the inductive case

8.1. **Description of the example.** Here, we assume that we have:

$$Y_i = f(X_i) + \xi_i$$

for $i \in \{1, ..., N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0,1] \subset \mathbb{R}$ are i.i.d. from a uniform distribution $\mathcal{U}(0,1)$ (and we assume that the statistician knows this point), the η_i are i.i.d. from a Gaussian distribution $\mathcal{N}(0,\sigma)$ and independent from the X_i. The statistician observes $(X_1, Y_1), ..., (X_N, Y_N)$ and wants to estimate the regression function f.

We will use three estimations methods. The first one will be an SVM obtained by the algorithm described previously, the second one a thresholded wavelets estimate also obtained by this algorithm, and we will compare both estimators to a "classical" thresholded wavelet estimate, as given by Kerkyacharian and Picard [24].

8.2. **The estimators.**

8.2.1. *Thresholded wavelets estimators.* Let us describe briefly the thresholded wavelet estimator. Let (φ, ψ) be the father wavelet and the mother wavelet, and:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x + k)$$

for $k \in \{0, ..., 2^j - 1\} = S_j$. For the sake of simplicity, let us write:

$$\psi_{-1,k}(x) = \varphi(x)$$

for $k \in \{0\} = S_{-1}$.

Here, we will use the Haar basis, with:

$$\varphi(x) = \mathbb{1}_{[0,1]}(x)$$
$$\psi(x) = \mathbb{1}_{\left[0,\frac{1}{2}\right]}(x) - \mathbb{1}_{\left[\frac{1}{2},1\right]}(x).$$

In the general case, we should use warped wavelets (for more details, see Kerkyacharian and Picard [24]): we put $F(x) = P(X \leq x)$, and:

$$\hat{\beta}_{j,k} = \frac{1}{N} \sum_{i=1}^{N} Y_i \psi_{j,k}(F(X_i)).$$

Just remark that the use of this method implies some assumptions about F that are not required by our algorithm (here again, see Kerkyacharian and Picard [24]).

In the case of the example, we will have:

$$\hat{\beta}_{j,k} = \frac{1}{N}\sum_{j=1}^{N} Y_i \psi_{j,k}(X_i).$$

For a given $\kappa \geq 0$ and $J \in \mathbb{N}$, we take:

$$\tilde{f}_J(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} \hat{\beta}_{j,k} \mathbb{1}(|\hat{\beta}_{j,k}| \geq \kappa t_N) \psi_{j,k}(.)$$

where:

$$t_N = \sqrt{\frac{\log N}{N}}.$$

Actually, we must choose J in such a way that:

$$2^J \sim t_N^{-1}.$$

When $\kappa = 0$ we obtain a classical wavelet estimator, and when $\kappa > 0$ we obtain a thresholded wavelet estimator, this is what we are going to do.

Here, we choose $\kappa = 0.5$ and $J = 7$.

8.2.2. *Wavelet estimators with our algorithm.* Here, we use the same family of functions, and we apply the algorithm given in subsection 7.3. So we take:

$$m = 2^J = 128.$$

We change only one thing in the method in order to obtain faster computations: here, applying the central limit theorem, we replace the theoretical confidence interval by its asymptotic Gaussian approximation.

More precisely:

$$\sqrt{N} \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)Y_i - P\left[f_{\theta_k}(X)Y\right]}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N} f_{\theta_k}(X_j)Y_j\right)^2}} \rightsquigarrow \mathcal{N}(0,1).$$

We put:

$$v_{k,N} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(f_{\theta_k}(X_i)Y_i - \frac{1}{N}\sum_{j=1}^{N} f_{\theta_k}(X_j)Y_j\right)^2}}{\sqrt{N}}.$$

We obtain:

$$\left(\mathcal{C}_k \hat{\alpha}_k - \overline{\alpha}_k\right) \frac{P\left[f_{\theta_k}(X)^2\right]}{v_{k,N}} \rightsquigarrow \mathcal{N}(0,1),$$

or:

$$\left(R(\mathcal{C}_k \hat{\alpha}_k \theta_k) - R(\overline{\alpha}_k \theta_k)\right) \frac{P\left[f_{\theta_k}(X)^2\right]}{v_{k,N}^2} \rightsquigarrow \chi_1^2,$$

and so we use the confidence interval for $\overline{\alpha}_k$:

$$\left[\mathcal{C}_k \hat{\alpha}_k \pm \frac{v_{k,N}}{P\left[f_{\theta_k}(X)^2\right]} q_{1-\frac{\varepsilon}{2m'(N)}}\right]$$

where q_α is the α-quantile of $\mathcal{N}(0,1)$.

Remark that the numerical results are not very different if we use the confidence interval given by Theorem 7.1.

Moreover, let us remark that the union bound are always "pessimistic", and that we use a union bound argument over all the m models despite only a few of them are effectively used in the estimator. So, we propose to actually use the individual confidence interval for each model:

$$\left[\mathcal{C}_k \hat{\alpha}_k \pm \frac{v_{k,N}}{P\left[f_{\theta_k}(X)^2\right]} q_{1-\frac{\varepsilon}{2}}\right]$$

TABLE 1. Values of t_i and c_i in the fonction $Blocks(.)$.

i	1	2	3	4	5	6	7	8	9	10	11
c_i	4	-5	3	-4	5	4.2	-2.1	4.3	-3.1	2.1	-4.2
t_i	0.10	0.13	0.15	0.23	0.25	0.40	0.44	0.65	0.76	0.78	0.81

TABLE 2. Results of the experiments. For each experiment, we give the mean risk (R) and the mean excess risk ($R - \sigma^2$) for each estimator.

Function $f(.)$	s.d. σ	standard thresholded wavelets	thresh. wav. with our method	multiscale SVM
Doppler	0.3	0.158 / 0.068	0.151 / 0.061	0.149 / 0.059
HeaviSine	0.3	0.154 / 0.064	0.138 / 0.048	0.129 / 0.039
Blocks	0.3	0.150 / 0.060	0.146 / 0.056	0.159 / 0.069
Doppler	1	1.142 / 0.142	1.114 / 0.114	1.091 / 0.091
HeaviSine	1	1.156 / 0.156	1.084 / 0.084	1.055 / 0.055
Blocks	1	1.155 / 0.155	1.105 / 0.105	1.104 / 0.104

instead of the theoretical union bound interval.

8.2.3. *SVM estimator.* Here, we use the multiscale SVM estimator described in Example 7.3 of subsection 7.4, with kernel:

$$K_\gamma(x, x') = \exp\left(-(2^\gamma x - 2^\gamma x')^2\right) = \exp\left(-2^{2\gamma}(x - x')^2\right)$$

and $\gamma \in \{1, ..., m'(N)\}$ where $m'(N) = 6$.

We use the same Gaussian approximation than in the previous example, and the individuals confidence intervals.

8.3. **Experiments and results.** The simulations were realized with the R software [32].

For the experiments, we use the following functions f that are some of the functions used by Donoho and Johnstone for experiments on wavelets, for example in [18], and by a lot of authors since then:

$$Doppler(t) = u\sqrt{t(1-t)}\sin\frac{2\pi(1+v)}{t+v} \quad \text{where } u = 2 \text{ and } v = 0.05$$

$$HeaviSine(t) = \frac{1}{4}\Big[4\sin 4\pi t - sgn(t - 0.3) - sgn(0.72 - t)\Big]$$

$$Blocks(t) = \frac{1}{4}\sum_{i=1}^{11} c_i \mathbb{1}_{(t_i, +\infty)}(t)$$

where $sgn(t)$ is the sign of t (say -1 if $t \leq 0$ and $+1$ otherwise). The values of the c_i and t_i are given in Table 1.

We consider 6 experiments (for the three regression functions and two different values for σ, 0.3 and 1). We choose ε=10%. We repeat each experiment 20 times. We give the results in Table 2.

The result of thresholding wavelets following [24] or using our algorithm is comparable. However, our thresholding method gives best results, especially when the noise level is significant. The main advantage of our method is that is is self-contained: in the "standard" thresholding, we have to choose the parameter κ. Here, the choice $\kappa = 0.5$ seemed to give the better results, but this choice was

possible only because we knew the regression function in these simulations. In real life problems, the choice of κ could be more problematic.

SVM gave best results, except in the case where $f = Blocks$ (with low noise). But the main advantage of SVM is that it is much easier to generalize in the case where $\mathcal{X}$ is not $\mathbb{R}$ or an interval of $\mathbb{R}$, but for example in the case where $\mathcal{X} = \mathbb{R}^n$ with $n \geq 2$. More generally, let us assume that $\mathcal{X}$ is a metric space for some distance d. We can use SVM with the Gaussian kernel ($(x, x') \in \mathcal{X}^2$):

$$K_\gamma(x, x') = \exp\left(-2^{2\gamma}\frac{d^2(x, x')}{2}\right).$$

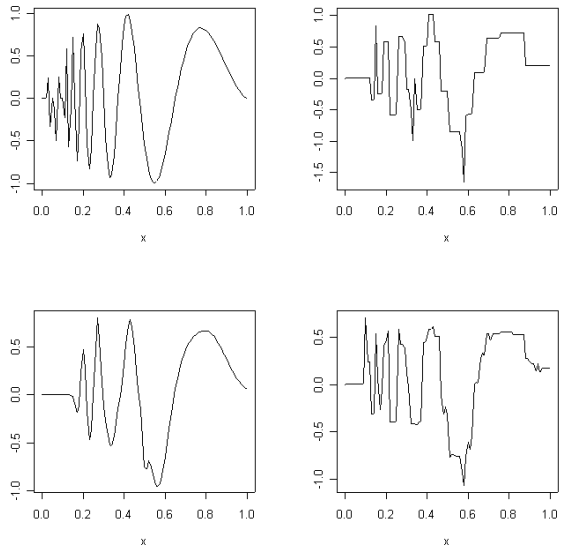

FIGURE 2. Experiment 1, $f = Doppler$ and $\sigma = 0.3$. Up-left: true regression function. Down-left: SVM. Up-right: wavelet estimate with our algorithm. Down-right: "classical" wavelet estimate.

9. THE TRANSDUCTIVE CASE

Remark that in this section, we make no longer assumptions about the existence of an exponential moment for $f_\theta(X)Y$.

9.1. **Notations.** Let us recall that we assume that P_{2N} is some exchangeable probability measure on the space $\left((\mathcal{X} \times \mathbb{R})^{2N}, (\mathcal{B} \times \mathcal{B}_\mathbb{R})^{\otimes 2N}\right)$. Let $(X_i, Y_i)_{i=1\ldots 2N} = (Z_i)_{i=1\ldots 2N}$ denote a random vector distributed according to P_{2N}.

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 $P_{2N} = P^{\otimes 2N}$, but what follows still holds for general exchangeable distributions P_{2N}.

We assume that we observe $(X_i, Y_i)_{i=1\ldots N}$ and $(X_i)_{i=N+1\ldots 2N}$. In this case, we only focus on the estimation of the values $(Y_i)_{i=N+1\ldots 2N}$.

Definition 9.1. We put, for any $\theta \in \Theta$:

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

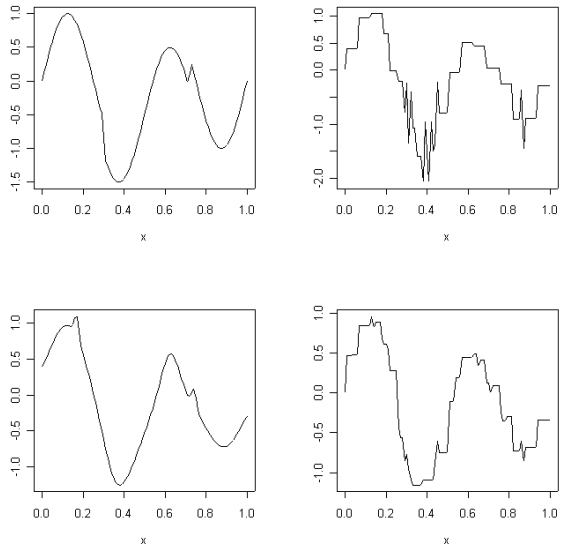

FIGURE 3. Experiment 2, $f = HeaviSine$ and $\sigma = 0.3$.

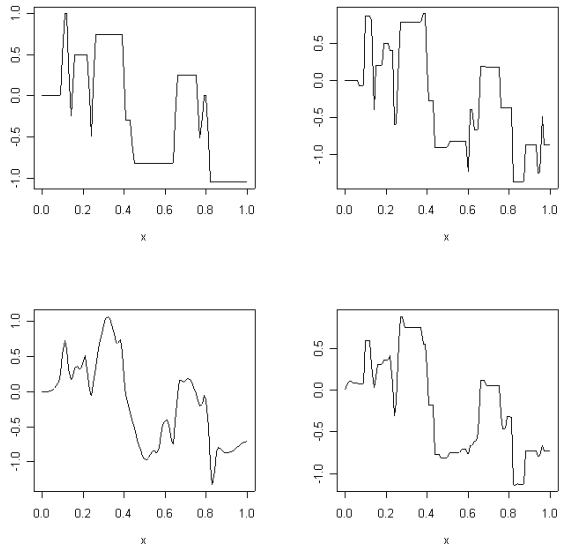

FIGURE 4. Experiment 3, $f = Blocks$ and $\sigma = 0.3$.

$$r_2(\theta) = \frac{1}{N} \sum_{i=N+1}^{2N} \left(Y_i - f_\theta(X_i)\right)^2 .$$

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:

$$\Theta_0 = \Theta_0(X_1, ..., X_{2N})$$

but we assume that the function $(X_1, ..., X_{2N}) \mapsto \Theta_0(X_1, ..., X_{2N})$ is exchangeable with respect to its $2N$ arguments, and is such that $m = m(N)$ depends only on N, not on $(X_1, ..., X_{2N})$.

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:

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

where

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

Remark that, in this case, every θ_k is an exchangeable function of $(X_1, ..., X_{2N})$. In some cases, we will use other indexations. For example, in the case of SVM, we will take $m = 2N$ and:

$$\Theta_0 = \{\Psi(X_1), ..., \Psi(X_{2N})\}.$$

Clearly, there is no reason for having $\theta_1 = \Psi(X_1)$. In such a case, if necessary we can use another notation, for example define $\theta_i^* = \Psi(X_i)$. Then we will have:

$$\Theta_0 = \{\theta_1^*, ..., \theta_m^*\}$$

where θ_i^* is not an exchangeable function of $(X_1, ..., X_{2N})$.

Now, let us write, for any $k \in \{1, ..., m\}$:

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

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

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

9.2. **Basics Results.**

Theorem 9.1. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

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

Remark 9.1. 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:

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

then we can get a looser observable upper bound:

$$P_{2N}\Bigg\{\forall k \in \{1, ..., m\}, \quad r_2\left[(\mathcal{C}^k \alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k)$$
$$\leq 4\left[B^2 + \frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2 Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2}\right]\frac{\log\frac{2m}{\varepsilon}}{N}\Bigg\} \geq 1-\varepsilon.$$

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

Theorem 9.2. *For any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:*

$$r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k) \leq \frac{8\log\frac{4m}{\varepsilon}}{N}\left[\frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4Y_i^4\log\frac{2m}{\varepsilon}}{2N}}\right].$$

Remark 9.2. Let us assume that Y is such that we know two constants b_Y and B_Y such that:

$$P\exp\left(b_Y Y\right) \leq B_Y < \infty.$$

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

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

So the bound of the theorem leads to a looser observable bound:

$$r_2\left[(\mathcal{C}^k\alpha_1^k).\theta_k\right] - r_2(\alpha_2^k.\theta_k) \leq \frac{8\log\frac{8m}{\varepsilon}}{N}\left[\frac{\frac{1}{N}\sum_{i=1}^{N} f_{\theta_k}(X_i)^2Y_i^2}{\frac{1}{N}\sum_{i=N+1}^{2N} f_{\theta_k}(X_i)^2} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{2N} f_{\theta_k}(X_i)^4\log\frac{4m}{\varepsilon}\log^4\frac{4NB_Y}{\varepsilon}}{2Nb_Y^4}}\right].$$

A proof of this assertion is given in the next section.

The proofs of both theorems are given in the next section: however, we are going to see at first how to apply this result.

Let us compare the first order term of this theorem to the analogous term in the inductive case (Theorems 7.1 and 7.2). 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 had the real variance of $Yf(X)$ 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 9.1 more than of its improved variants.

9.3. **Improvements and generalization of the bound.** The proof of all the theorems of this subsection is given in the next section.

9.3.1. *Relative bounds.* We introduce some new notations.

Definition 9.2. We write:

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

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

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

The we have the following theorem.

Theorem 9.3. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:*

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

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 $k \in \{1,...,m\}$:

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

So we can see that this theorem is an improvement on theorem 9.1 when some features $f_{\theta_k}(.)$ are well correlated with Y. But we loose another factor 2 by making the first-order term of the bound observable.

9.3.2. *Improvement of the variance term.*

Theorem 9.4. *We have, for any $\varepsilon > 0$, with P_{2N}-probability at least $1-\varepsilon$, for any $k \in \{1,...,m\}$:*

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

where:

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

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

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

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

Here again, we can make the bound fully observable under an exponential moment assumption about Y.

9.3.3. *Test sample of different size.* In the context of classification, Catoni [12] gave a method in order to be able to deal with the case where the test sample is of size kN where k is an integer greater than 0. More precisely, we assume that $P_{(k+1)N}$ is an exchangeable probability distribution on $(\mathcal{X}\times\mathbb{R})^{(k+1)N}$ and that we observe:

$$(X_1,Y_1),...,(X_N,Y_N) \quad \text{and} \quad X_{N+1},...,X_{(k+1)N}.$$

In the case where $k > 1$, the variance term will be better than in the case where $k = 1$. This method can be used in the setting of regression too.

Definition 9.3. From now, we will use the notation, when $k \neq 1$:

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

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

We still consider a family:

$$\Theta_0(X_1, ..., X_{(k+1)N}) = \{\theta_1, ..., \theta_m\}$$

that is data-dependent in an exchangeable way, with the same indexation convention than in the case where $k = 1$. Now, let us write, for any $h \in \{1, ..., m\}$:

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

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

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

Let us finally put:

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

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

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

Then we have the following theorem.

Theorem 9.5. *Let us assume that we have constants B_h and β_h such that, for any $h \in \{1, ..., m\}$:*

$$P \exp\left(\beta_h \left| f_{\theta_h}(X_i)Y_i \right|\right) \leq B_h.$$

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

$$r_2\left(\mathcal{C}^h \alpha_1^h \theta_h\right) - r_2\left(\alpha_2^h \theta_h\right) \leq \frac{\left(1 + \frac{1}{k}\right)^2}{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2} \left[\frac{2\mathbb{V}_{\theta_h} \log \frac{4m}{\varepsilon}}{N} \right.$$
$$\left. + \frac{16\left(\log \frac{4m}{\varepsilon}\right)^{\frac{3}{2}} \left(\log \frac{4(k+1)mNB_h}{\varepsilon}\right)^3}{3\beta_h^3 N^{\frac{3}{2}} \mathbb{V}_{\theta_h}^{\frac{1}{2}}} + \frac{64\left(\log \frac{4m}{\varepsilon}\right)^2 \left(\log \frac{4(k+1)mNB_h}{\varepsilon}\right)^6}{9\beta_h^6 N^2 \mathbb{V}_{\theta_h}^2} \right].$$

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

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

9.4. **Application to regression estimation.** 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 9.4. We take, for any $(\theta, \theta') \in \Theta^2$:

$$d_2(\theta, \theta') = \sqrt{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \left[f_\theta(X_i) - f_{\theta'}(X_i) \right]^2} = \sqrt{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \langle \theta - \theta', \Psi(X_i) \rangle^2}.$$

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

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

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

$$\mathcal{CR}(h, \varepsilon) = \left\{ \theta \in \Theta : \left| \langle \theta - \mathcal{C}^h \alpha_1^h \theta_h, \theta_h \rangle_2 \right| \leq \sqrt{\beta(\varepsilon, h)} \right\},$$

where $\beta(\varepsilon, h)$ is the upper bound in theorem 9.1 (or in the other theorems given previously).

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

Corollary 9.6. *We have:*

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

Definition 9.5. 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: $\theta(0), ..., \theta(n-1)$. Choose $h(n)$, for example:

  $$h(n) = \arg \max_{h \in \{1, ..., m\}} d_2(\theta(n-1), \mathcal{CR}(h, \varepsilon)),$$

  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 9.6. We write n_0 the stopping step, and:

$$\hat{f}(.) = f_{\theta(n_0)}(.)$$

the corresponding function.

Here again we give a detailed version of the algorithm, see Figure 5. 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 9.7. *We have:*

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

The proof of this theorem is exactly the same as the proof of Theorem 7.4.

Example 9.1 (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:

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

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

Example 9.2 (SVM and multiscale SVM)**.** Let us choose Θ as the set of all functions $\mathcal{X} \to \mathbb{R}$, $f_\theta(x) = \theta(x)$, a family of kernels $K_1, ..., K_{m'(N)}$ for a $m'(N) \geq 1$ and:

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

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

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

FIGURE 5. Detailed version of the feature selection algorithm in the transductive case.

We have $\varepsilon > 0$, $\kappa > 0$, N observations $(X_1, Y_1), ..., (X_N, Y_N)$ and also $X_{N+1}, ..., X_{(k+1)N}$, m features $f_{\theta_1}(.), ..., f_{\theta_m}(.)$ and $\theta(0) = (\theta_1(0), ..., \theta_m(0)) = (0, ..., 0) \in \mathbb{R}^m$. In a first time, compute every α_1^h and $\beta(\varepsilon, h)$ for $h \in \{1, ..., m\}$. Set $n \leftarrow 0$.

Repeat:

- set $n \leftarrow n + 1$;
- set $best\_improvement \leftarrow 0$ and $\theta(n) = \theta(n-1)$;
- for $h \in \{1, ..., m\}$, compute:

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

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

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

  and if $\delta_h > best\_improvement$, set:

$$best\_improvement \leftarrow \delta_h,$$

$$h(n) \leftarrow h;$$

- if $best\_improvement > 0$:

$$\theta_{h(n)}(n) \leftarrow \theta_{h(n)}(n) + sgn(\gamma_h) \Big(|\gamma_h| - \beta(\varepsilon, h) \Big)_+;$$

until $best\_improvement < \kappa$.

Return the estimation:

$$\left[\tilde{Y}_{N+1}, ..., \tilde{Y}_{(k+1)N} \right] = \left[\hat{f}(X_{N+1}), ..., \hat{f}(X_{(k+1)N}) \right]$$

where:

$$\hat{f}(.) = \sum_{h=1}^{m} \theta_h(n) f_{\theta_h}(.).$$

Let us put:

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

We have:

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

that is a Support Vector Machine with different kernels estimate; like in Example 7.3, the kernels K_h can be the same kernel taken at different scales.

Example 9.3 (Kernel PCA Kernel Projection Machine)**.** Let us take Θ as a Hilbert space, with scalar product $\langle ., . \rangle$, let us take a function $\Psi : \mathcal{X} \to \Theta$ and consider the kernel:

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

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

$$\left\{ \Psi(X_1), ..., \Psi(X_{(k+1)N}) \right\}$$

by performing a diagonalization of the matrix:

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

This method is known as Kernel PCA, see for example Schölkopf, Smola and Müller [36]. We obtain eigenvalues:

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

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

$$\Psi_1 = \sum_{i=1}^{(k+1)N} e_i^1 \Psi(X_i), ..., \Psi_{(k+1)N} = \sum_{i=1}^{(k+1)N} e_i^{(k+1)N} \Psi(X_i)$$

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

$$\Theta_0 = \left\{\Psi_1, ..., \Psi_{(k+1)N}\right\}$$

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 [6] in the context of classification.

10. Proof of the theorems in the transductive case

10.1. **Proof of Theorems 9.1 and 9.2.** 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 [10]. We go back to the notations oh theorem 9.1 and 9.2, with test sample of size N.

Definition 10.1. Let $\mathcal{G}$ denote the set of all functions:

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

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

Lemma 10.1. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, ..., Z_{2N})$, 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 $g \in \mathcal{G}$:*

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

and the reverse inequality:

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

where we write:

$$\begin{aligned} \eta &= \eta\left((X_1, Y_1), ..., (X_{2N}, Y_{2N})\right) \\ \lambda &= \lambda\left((X_1, Y_1), ..., (X_{2N}, Y_{2N})\right) \end{aligned}$$

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:

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

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

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

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

We obtain:

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

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

We can now give the proof of the theorems.

Proof of Theorem 9.1. From now we assume that the hypothesis of Theorem 9.1 are satisfied. Let us choose $\varepsilon' > 0$ and apply Lemma 10.1 with $\eta = -\log\varepsilon'$, and g such that $g(u,u') = uu'$. We obtain: for any exchangeable distribution $\mathcal{P}$, 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 $\theta\in\Theta$:

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

and the reverse inequality:

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

Let us denote:

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

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 $\theta\in\Theta$:

$$\mathcal{P}\exp f((Z_1,...,Z_{2N}),\theta,\varepsilon',\lambda) \leq 2\varepsilon'. \tag{10.1}$$

Now, let us introduce a new conditional probability measure:

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

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

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

The measure $\overline{P}$ is exchangeable, so we can apply Equation 10.1. For any values of $Z_1, ..., Z_{2N}$ we have:

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

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

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

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

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

Now let us apply inequality 10.1, we obtain:

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

if we choose:

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

From now, we assume that the event:

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

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

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

Let us divide both inequalities by:

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

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

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

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:

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

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

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

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

Finally, remark that:

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

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

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

This ends the proof. □

Proof of Theorem 9.2. We write:

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

and try to upper bound the second term. We apply Lemma 10.1, but this time with g such that $g(u) = (uu')^2$ that is nonnegative, and obtain, for any ε, for any (exchangeable) θ and λ:

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

We choose:

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

we apply this result to every $\theta \in \Theta_0$, and combine it with Theorem 9.1 by a union bound argument to obtain the result. □

10.2. **Proof of Theorem 9.3.** First of all, we give the following obvious variant of Lemma 10.1:

Lemma 10.2. *For any exchangeable probability distribution $\mathcal{P}$ on $(Z_1, ..., Z_{2N})$, 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 $\theta \in \Theta$:*

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

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 10.1, we just need to remark that $\alpha(\theta)$ is an exchangeable function of $(Z_1, ..., Z_{2N})$, and so we can take in Lemma 10.1:

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

that means that:

$$g\left[f_\theta(X_i), Y_i \right] = f_\theta(X_i) Y_i - \alpha(\theta) f_\theta(X_i)^2.$$

□

Proof of Theorem 9.3. Proceeding exactly in the same way as in the proof of theorem 9.1, we obtain the following inequality with probability at least $1 - \varepsilon$:

(10.2)

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

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:

$$\begin{aligned} &\left[f_{\theta_k}(X_i) Y_i - \alpha_{1,2}^k f_{\theta_k}(X_i)^2 \right]^2 \\ &\quad = \left[f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right]^2 + \left[\alpha_1^k - \alpha_{1,2}^k \right]^2 f_{\theta_k}(X_i)^4 \\ &\qquad + 2 \left[f_{\theta_k}(X_i) Y_i - \alpha_1^k f_{\theta_k}(X_i)^2 \right] \left[\alpha_1^k - \alpha_{1,2}^k \right] f_{\theta_k}(X_i)^2. \end{aligned}$$

Remark that it is obvious that:

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

and so:

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

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:

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

So finally, Equation 10.2 left us with a second degree inequality with respect to $\left|\mathcal{C}^k \alpha_1^k - \alpha_2^k\right|$ 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:

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

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

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

with the following notations:

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

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

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:

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

we have:

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

We can notice that this bound may be written:

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

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

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

with high probability. This can be done very simply, using Lemma 10.1 with this time:

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

We obtain the bound:

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

10.3. **Proof of Theorem 9.4.** The proof is exactly similar, we just use a new variant of Lemma 10.1, that is based on an idea introduced by Catoni [12] in the context of classification.

Definition 10.2. Let us write:

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

for short. We also introduce a conditional probability measure:

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

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

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

Lemma 10.3. *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 $\theta\in\Theta$:

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

and the reverse inequality.

Proof. Let $\mathcal{L}hs$ denote the left-hand side of Lemma 10.3. For short, let us put:

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

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 10.1:

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

□

Proof of Theorem 9.4. We apply both inequalities of Lemma 10.3 to every $\theta_k, k \in \{1,...,m\}$, and we take:

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

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

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

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

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

so:

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

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 10.1 with $g(f_\theta(X_i),Y_i) = f_\theta(X_i)^2Y_i^2 = T_\theta(Z_i)^2$, so with probability at least $1-\varepsilon$, for any k:

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

For the second order term, we use both inequalities of Lemma 10.1 with $g(f_\theta(X_i), Y_i) = f_\theta(X_i)Y_i = T_\theta(Z_i)$, so with probability at least $1-\varepsilon$, for any k::

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

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

10.4. **Proof of Theorem 9.5.**

Proof of Theorem 9.5. We introduce the following conditional probability measures, for any $i \in \{1, ..., N\}$:

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

and:

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

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 $i \in \{1, ..., N\}$:

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

Lemma 10.4. *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:*

$$\mathbb{P}\exp\left\{\lambda\left[\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\chi\Big[f_\theta(X_i)Y_i\Big] - \frac{1}{N}\sum_{i=1}^{N}\chi\Big[f_\theta(X_i)Y_i\Big]\right] - \eta\right\}$$
$$\leq \exp(-\eta)\exp\left\{\frac{\lambda^2(1+k)^2}{2Nk^2}\mathbf{P}\left\{\Big[\chi\Big(f_\theta(X)Y\Big) - \mathbf{P}\chi\Big(f_\theta(X)Y\Big)\Big]^2\right\}\right.$$
$$\left. + \frac{\lambda^3(1+k)^3}{6N^2k^3}\left[\sup_{i\in\{1,\ldots,(k+1)N\}}\chi\left(f_\theta(X_i)Y_i\right) - \inf_{i\in\{1,\ldots,(k+1)N\}}\chi\left(f_\theta(X_i)Y_i\right)\right]^3\right\},$$

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

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

Definition 10.3. We put, for any $\theta \in \Theta$, for any function χ:

$$\chi_i^\theta = \chi\left(Y_i f_\theta(X_i)\right),$$

and:

$$\chi^\theta = \chi\left(Y f_\theta(X)\right)$$

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. 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:

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

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

Now, we have:

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

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

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

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

$$\begin{aligned}
&\frac{1}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\mathbb{P}_i\left[\left(\chi_i^\theta - \frac{\mathbb{P}_i\left\{\chi_i^\theta\exp\left[-\beta\chi_i^\theta\right]\right\}}{\mathbb{P}_i\exp\left[-\beta\chi_i^\theta\right]}\right)^3\exp\left(-\beta\chi_i^\theta\right)\right]\\
&\qquad\qquad \leq \left[\sup_{j\in\{1,\dots,k\}}\chi_{i+(j-1)N}^\theta - \inf_{j\in\{1,\dots,k\}}\chi_{i+(j-1)N}^\theta\right]^3,
\end{aligned}$$

and so:

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

Note that:

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

and so:

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

remark also that:

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

we obtain:

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

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 $h \in \{1,\ldots,m\}$:

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

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

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

the bound becomes:

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

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 $h \in \{1,...,m\}$:

$$r_2\left(\mathcal{C}^h\alpha_1^h\theta_h\right) - r_2\left(\alpha_2^h\theta_h\right) \leq \frac{\left(1+\frac{1}{k}\right)^2}{\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} f_{\theta_h}(X_i)^2}\left[\frac{2\mathbb{V}_{\theta_h}\log\frac{2m}{\varepsilon}}{N}\right.$$
$$\left. + \frac{2\left(\log\frac{2m}{\varepsilon}\right)^{\frac{3}{2}}\mathcal{S}_{id}(\theta_h)^3}{3N^{\frac{3}{2}}\mathbb{V}_{\theta_h}^{\frac{1}{2}}} + \frac{\left(\log\frac{2m}{\varepsilon}\right)^2\mathcal{S}_{id}(\theta_h)^6}{9N^2\mathbb{V}_{\theta_h}^2}\right], \tag{10.3}$$

remember that:

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

We now give a new lemma.

Lemma 10.5. *Let us assume that P is such that, for any $h \in \{1,...,m\}$:*

$$\exists \beta_h > 0, \exists B_h \geq 0, P\exp\left(\beta_h\left|f_{\theta_h}(X)Y\right|\right) \leq B_h.$$

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

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

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

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

Proof of the lemma. We have:

$$P_{(k+1)N}\left(\sup_{1\leq i\leq(k+1)N} f_{\theta_h}(X_i)Y_i \geq s\right)$$
$$= P_{(k+1)N}\left(\exists i \in \{1,...,(k+1)N\}, f_{\theta_h}(X_i)Y_i \geq s\right)$$
$$= \sum_{i=1}^{(k+1)N} P\mathbb{1}_{f_{\theta_h}(X_i)Y_i\geq s}$$
$$\leq (k+1)NP\exp\left(\beta_h\left|f_{\theta_h}(X_i)Y_i - s\right|\right) \leq (k+1)NB_h\exp\left(-\beta_h s\right).$$

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 $h \in \{1,...,m\}$:

$$\mathcal{S}_{id}(\theta_h) = \sup_{i\in\{1,...,(k+1)N\}} f_{\theta_h}(X_i)Y_i - \inf_{i\in\{1,...,(k+1)N\}} f_{\theta_h}(X_i)Y_i$$

$$\leq \frac{2}{\beta_h} \log \frac{2(k+1)mNB_h}{\varepsilon}.$$

By plugging the lemma into Equation 10.3 we obtain the theorem. □

11. Simulations in the transductive case

11.1. **Description of the example.** Here, we assume that we have:

$$Y_i = f(X_i) + \xi_i$$

for $i \in \{1, ..., 2N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0,1] \subset \mathbb{R}$ are i.i.d. from a uniform distribution $\mathcal{U}(0,1)$ (here we DO NOT assume that the statistician knows this point), the η_i are i.i.d. from a Gaussian distribution $\mathcal{N}(0,\sigma)$ and independent from the X_i. The statistician observes $(X_1, Y_1), ..., (X_N, Y_N)$ and $X_{N+1}, ..., X_{2N}$ and wants to estimate $Y_{N+1}, ..., Y_{2N}$.

We will here again use three estimations methods: an inductive method, that does not take advantage of the knowledge of $X_{N+1}, ..., X_{2N}$, and two transductive methods. For the inductive method, we take the thresholded wavelet estimator that we used in the experiments in the inductive case. For the transductive method, we use here again a wavelet estimator and a (multiscale) SVM.

11.2. **The estimators.**

11.2.1. *Thresholded wavelets estimators.* In this case, as we assume that we don't know the distribution $P_{(X)}$, we have to estimate it and use a warped wavelet estimator. We take:

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \leq x),$$

and:

$$\hat{\beta}_{j,k} = \frac{1}{N} \sum_{j=1}^{N} Y_i \psi_{j,k}(F_N(X_i)),$$

$$\tilde{f}_J(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} \hat{\beta}_{j,k} \mathbb{1}(|\hat{\beta}_{j,k}| \geq \kappa t_N) \psi_{j,k}(F_N(.)).$$

Here again, we choose $\kappa = 0.5$ and $J = 7$.

11.2.2. *Wavelet estimators with our algorithm.* Here, we use the same family of functions, and we apply the transductive method described previously. Here again, we use Gaussian approximations for the confidence intervals (but we double their length in order to take into account the variance of both samples).

11.2.3. *SVM estimator.* The transductive SVM estimator is taken with kernel:

$$K_\gamma(x, x') = \exp\left(-2^{2\gamma}(x - x')^2\right)$$

and $\gamma \in \{1, ..., m'(N)\}$ where $m'(N) = 6$. We use the same Gaussian approximation than in the previous example.

11.3. **Experiments and results.** We consider the same functions than in the inductive case. We choose ε=10%. We repeat each experiment 20 times. We give the results in Table 3.

TABLE 3. Results of the experiments. For each experiment, we give the mean risk r_2.

Function $f(.)$	s.d. σ	"inductive" thresholded wavelets	transductive thresh. wav. with our method	transductive multiscale SVM
Doppler	0.3	0.234	0.174	0.165
HeaviSine	0.3	0.134	0.156	0.134
Blocks	0.3	0.187	0.171	0.177
Doppler	1	1.179	1.152	1.120
HeaviSine	1	1.092	1.110	1.065
Blocks	1	1.153	1.144	1.129

12. BOUND ON A MULTIDIMENSIONNAL MODEL

12.1. **Theorem and algorithm.** In this subsection, we try to generalize the algorithm described in section 9 to the case where there are multidimensional models. The idea is that, for example, if $\Theta_0 = \{\theta_1, \theta_2, \theta_3\}$, we could try not only to make projections on:

$$\{\alpha\theta_i, \alpha \in \mathbb{R}\} \text{ for } i \in \{1, 2, 3\}$$

but also on a bidimensional space like:

$$\left\{\alpha\theta_1 + \beta\theta_2, (\alpha, \beta) \in \mathbb{R}^2\right\}.$$

More precisely, let us give the following definitions. First of all, we assume that we are in the case where $k = 1$, so the test sample and the learning sample have size N. We always assume that:

$$\Theta_0(Z_1, ..., Z_{2N}) = \{\theta_1, ..., \theta_m\}$$

is such that every θ_k is an exchangeable function of $(Z_1, ..., Z_{2N})$.

Definition 12.1. For every $d \geq 0$, $0 < j_1 < ... < j_d < m+1$ and $\mathcal{S} = (\theta_{j_1}, ..., \theta_{j_d}) \in \Theta_0^d$ we put:

$$f_{\mathcal{S}}(x) = \left(f_{\theta_{j_1}}(x), ..., f_{\theta_{j_d}}(x)\right).$$

For convenience, let us put, for any $\alpha = (\alpha^1, ..., \alpha^d) \in \mathbb{R}^d$:

$$\alpha\mathcal{S}' = \sum_{k=1}^{d} \alpha^k \theta_{j_k}$$

Remark that we have:

$$\alpha f_{\mathcal{S}}(.)' = f_{\alpha\mathcal{S}'}(.) : \mathcal{X} \to \mathbb{R};$$

let us put:

$$C_{1,2}^{\mathcal{S}} = \frac{1}{N} \sum_{i=1}^{2N} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)$$

$$C_1^{\mathcal{S}} = \frac{1}{N} \sum_{i=1}^{N} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)$$

$$\mathcal{M}^{\mathcal{S}} = \frac{1}{2} C_{1,2}^{\mathcal{S}} \left(C_1^{\mathcal{S}}\right)^{-1},$$

and finally:

$$\alpha_{1,2}^{\mathcal{S}} = \arg\min_{\alpha\in\mathbb{R}^d} r_{1,2}\left(\alpha\mathcal{S}'\right) = \frac{1}{N}\sum_{i=1}^{2N} Y_i f_{\mathcal{S}}(X_i)\left(C_{1,2}^{\mathcal{S}}\right)^{-1}$$

$$\alpha_{1}^{\mathcal{S}} = \arg\min_{\alpha\in\mathbb{R}^d} r_{1}\left(\alpha\mathcal{S}'\right) = \frac{1}{N}\sum_{i=1}^{N} Y_i f_{\mathcal{S}}(X_i)\left(C_{1}^{\mathcal{S}}\right)^{-1}.$$

For any matrix M we will let $\rho(M)$ denote the biggest eigenvalue of M.

Here, $\alpha_{1,2}^{\mathcal{S}}$ is our objective but we can only observe $\alpha_1^{\mathcal{S}}$, and the matrix $\mathcal{M}^{\mathcal{S}}$.

Remark 12.1. Note the change in the objective. In this subsection, we try to minimize $r_{1,2}$ and not r_2.

Theorem 12.1. *Let $d \geq 0$, let $\mathcal{S} \in \Theta^d$. Let us put:*

$$B_{1,2}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{2N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)\mathcal{C}_{1,2}^{-\frac{1}{2}}.$$

For any $\varepsilon > 0$, we have, with P_{2N}-probability at least $1-\varepsilon$:

$$r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right) \leq \frac{4\rho\left(B_{1,2}^{\mathcal{S}}\right)}{N}\left(d\log(2) + 2\log\frac{1}{\varepsilon}\right).$$

Note that $B_{1,2}^{\mathcal{S}}$ is not observable, except in the case of classification where we have $Y_i \in \{-1,+1\}$ and so $Y_i^2 = 1$, which implies that $B_{1,2}^{\mathcal{S}} = I$ and so:

$$\rho\left(B_{1,2}^{\mathcal{S}}\right) = 1.$$

In the general case we have the following corollary.

Corollary 12.2. *Let $d \geq 0$, let $\mathcal{S} \in \Theta^d$. Let us put:*

$$B_{1}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i)\mathcal{C}_{1,2}^{-\frac{1}{2}},$$

that is observable, and:

$$D_{1,2}^{\mathcal{S}} = \frac{1}{N}\sum_{i=1}^{2N} Y_i^4\left(\lambda_{1,2}C_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i) C_{1,2}^{-\frac{1}{2}}\lambda_{1,2}'\right)^2,$$

where:

$$\rho\left(B_{1,2}\right) = \sup_{\|\lambda\|=1} \lambda B_{1,2}\lambda' = \lambda_{1,2}B_{1,2}\lambda_{1,2}'.$$

For any $\varepsilon > 0$, we have with P_{2N}-probability at least $1-\varepsilon$:

$$r_{1,2}\left(\mathcal{M}^{\mathcal{S}}\alpha_1^{\mathcal{S}}\mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}}\mathcal{S}'\right) \leq \frac{8\rho\left(B_1^{\mathcal{S}}\right)}{N}\left(d\log(2) + 2\log\frac{2}{\varepsilon}\right) + \frac{4\left[D_{1,2}^{\mathcal{S}} + \log\frac{2}{\varepsilon}\right]}{N^{\frac{3}{2}}}.$$

We can now give a new algorithm to perform regression estimation, that is a variant of the one given in section 9. Before all, we have to choose k dimensions $d_1, ..., d_k$ and k models

$$\mathcal{S}_1 \in \Theta^{d_1}, ..., \mathcal{S}_k \in \Theta^{d_k}.$$

We apply Theorem 12.1 to all the models simultaneously by a union bound argument and we obtain k confidence regions:

$$\mathcal{CR}_1, ..., \mathcal{CR}_k$$

and the corresponding projections:

$$\Pi_1, ..., \Pi_k.$$

We then use the following algorithm:

- choose $\theta(0) = 0$;
- at step $n \in \mathbb{N}^*$, define:

$$k'(n) = \arg\max_{k'} d_{1,2}\left(\theta(n-1), \Pi_{k'}\theta(n-1)\right)$$

and

$$\theta(n) = \Pi_{k'(n)}\theta(n-1);$$

- stop when $d_{1,2}\left(\theta(n), \theta(n-1)\right) \leq \kappa$.

Example 12.1. By taking $k = m$, $d_1 = ... = d_m = 1$ and $\mathcal{S}_i = \theta_i$ for all i, we obtain exactly the projection algorithm described in section 9.

Example 12.2. Let us take $k = m$, $d_i = i$ for any i and $\mathcal{S}_i = (\theta_1, ..., \theta_i)$: we are in the case of nested submodels, and we obtain a procedure similar to Lepski's method [26], at least in its form proposed by Birgé [5].

12.2. **Proofs.** For convenience, we assume that $\mathcal{S}$ is chosen once and for all, and so we will let $B_{1,2}$ stand for $B_{1,2}^{\mathcal{S}}$, $\mathcal{C}_{1,2}$ for $\mathcal{C}_{1,2}^{\mathcal{S}}$, and $\mathcal{D}_{1,2}$ for $\mathcal{D}_{1,2}^{\mathcal{S}}$. We keep the notation $f_{\mathcal{S}}(.)$ to avoid confusion with the true regression function $f(.)$.

Proof of Theorem 12.1. Let us state the following variant of Lemma 10.1, obtained exactly in the same way. 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 $\gamma : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+$ that is exchangeable with respect to its $2 \times 2N$ arguments, for any $\lambda \in \mathbb{R}^d$:

$$P_{2N} \exp\left(\gamma \left\langle \frac{\mathcal{C}_{1,2}^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} \left\{ f_{\mathcal{S}}(X_i)Y_i - f_{\mathcal{S}}(X_{i+N})Y_{i+N} \right\}, \lambda \right\rangle - \|\lambda\|^2 - \eta \right)$$
$$\leq P_{2N} \exp\left(\frac{\gamma^2}{N} \frac{1}{N} \sum_{i=1}^{2N} \left\langle \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)Y_i, \lambda \right\rangle^2 - \|\lambda\|^2 - \eta \right),$$

that can be written:

$$P_{2N} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) \leq P_{2N} \exp\left(\frac{\gamma^2}{N}\lambda B_{1,2}\lambda' - \lambda I\lambda' - \eta\right)$$

where:

$$A = \frac{\mathcal{C}_{1,2}^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} \left\{ \Psi(X_i)Y_i - \Psi(X_{i+N})Y_{i+N} \right\}.$$

So we have:

$$\int_{\mathbb{R}^d} P_{2N} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) d\lambda \leq \int_{\mathbb{R}^d} P_{2N} \exp\left(\frac{\gamma^2}{N}\lambda B_{1,2}\lambda' - \lambda I\lambda' - \eta\right) d\lambda.$$

Using Fubini's theorem we obtain:

$$P_{2N} \int_{\mathbb{R}^d} \exp\left(\gamma A\lambda' - \lambda I\lambda' - \eta\right) d\lambda \leq P_{2N} \int_{\mathbb{R}^d} \exp\left(\lambda \left(\frac{\gamma^2}{N} B_{1,2} - I\right) \lambda' - \eta\right) d\lambda.$$

Now, let us assume that γ is small enough for the matrix

$$I - \frac{\gamma^2}{N} B_{1,2}$$

to be definite positive. Actually this means that:

$$\frac{N}{\gamma^2} > \rho\left(B_{1,2}\right)$$

or:

$$\gamma < \sqrt{\frac{N}{\rho\left(B_{1,2}\right)}}.$$

Then we get:

$$P_{2N}\left[\pi^{\frac{d}{2}} \exp\left(\frac{\gamma^2}{4} AA' - \eta\right)\right] \leq P_{2N}\left[\frac{\pi^{\frac{d}{2}} \exp(-\eta)}{\sqrt{\det\left(I - \frac{\gamma^2}{N} B_{1,2}\right)}}\right],$$

or:

$$P_{2N}\left[\exp\left(\frac{\gamma^2}{4} AA' - \eta\right)\right] \leq P_{2N}\left[\exp\left(-\eta - \frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N} B_{1,2}\right)\right)\right].$$

Let us put:

$$\eta = -\frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N} B_{1,2}\right) + \log\frac{1}{\varepsilon},$$

we get:

$$P_{2N}\left[\exp\left(\frac{\gamma^2}{4} AA' + \frac{1}{2}\log\det\left(I - \frac{\gamma^2}{N} B_{1,2}\right) - \log\frac{1}{\varepsilon}\right)\right] \leq \varepsilon.$$

This implies that:

$$P_{2N}\left[AA' \leq \frac{4}{\gamma^2}\log\frac{1}{\varepsilon\sqrt{\det\left(I - \frac{\gamma^2}{N} B\right)}}\right] \geq 1 - \varepsilon.$$

Finally, note that:

$$AA' = r_{1,2}\left(\mathcal{M}^{\mathcal{S}} \alpha_1^{\mathcal{S}} \mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}} \mathcal{S}'\right).$$

We obtain the following result. For any $\varepsilon > 0$, for any measurable function $\gamma : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}_+$ that is exchangeable with respect to its $2 \times 2N$ arguments:

$$P_{2N}\left[r_{1,2}\left(\mathcal{M}^{\mathcal{S}} \alpha_1^{\mathcal{S}} \mathcal{S}'\right) - r_{1,2}\left(\alpha_{1,2}^{\mathcal{S}} \mathcal{S}'\right) \leq \frac{4}{\gamma^2}\log\frac{1}{\varepsilon\sqrt{\det\left(I - \frac{\gamma^2}{N} B_{1,2}\right)}}\right] \geq 1 - \varepsilon.$$

In particular if we choose:

$$\gamma = \sqrt{\frac{N}{2\rho\left(B_{1,2}\right)}}$$

then we obtain the theorem. □

Proof of corollary 12.2. In a first time, let us introduce the following obvious notation:

$$B_2^{\mathcal{S}} = B_2 = \frac{1}{N}\sum_{i=N+1}^{2N} Y_i^2 \mathcal{C}_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i) \mathcal{C}_{1,2}^{-\frac{1}{2}}.$$

Now, we state a new variant of Lemma 10.1: 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 $\lambda : (\mathcal{X} \times \mathbb{R})^{2N} \to \mathbb{R}^d$ that is exchangeable with respect to its $2 \times 2N$ arguments:

$$P_{2N}\exp\left(\lambda B_2\lambda' - \lambda B_1\lambda' - \eta\right)$$
$$\leq P_{2N}\exp\left(\frac{1}{N^2}\sum_{i=1}^{2N} Y_i^4\left(\lambda C_{1,2}^{-\frac{1}{2}} f_{\mathcal{S}}(X_i)' f_{\mathcal{S}}(X_i) C_{1,2}^{-\frac{1}{2}}\lambda'\right)^2 - \eta\right).$$

Now, taking:

$$\lambda = N^{\frac{1}{4}}\lambda_{1,2}$$

and:

$$\eta = \log\frac{1}{\varepsilon}$$

we obtain:

$$P_{2N}\left[\lambda_{1,2}B_2\lambda'_{1,2} \leq \lambda_{1,2}B_1\lambda'_{1,2} + \frac{D_{1,2} + \log\frac{1}{\varepsilon}}{\sqrt{N}}\right] \geq 1-\varepsilon.$$

So, with probability at least $1-\varepsilon$:

$$\rho\left(B_{1,2}\right) = \lambda_{1,2}B_{1,2}\lambda'_{1,2} = \lambda_{1,2}B_1\lambda'_{1,2} + \lambda_{1,2}B_2\lambda'_{1,2}$$
$$\leq 2\lambda_{1,2}B_1\lambda'_{1,2} + \frac{D_{1,2} + \log\frac{1}{\varepsilon}}{\sqrt{N}}$$
$$\leq 2\sup_{\|\lambda\|=1}\lambda B_1\lambda' + \frac{D_{1,2} + \log\frac{1}{\varepsilon}}{\sqrt{N}} = 2\rho\left(B_1\right) + \frac{D_{1,2} + \log\frac{1}{\varepsilon}}{\sqrt{N}}.$$

The last step is to combine this inequality with Theorem 12.1 by a union bound argument. □

12.3. **PAC-Bayesian bound for a multidimensionnal model in the inductive case.** A similar result can be derived in the inductive case. However, note that as soon as we can obtain a bound under the form:

$$R(\hat{\theta}) - \arg\min_{\theta\in\mathcal{S}} R(\theta) \leq Cst\frac{d\log\frac{1}{\varepsilon}}{N}$$

for some constant Cst, $d = |\mathcal{S}|$ and some estimator $\hat{\theta}$, we can apply our method.

Note that this is exactly the purpose of the bound obtained in the first part by a PAC-Bayesian technique, bound 5.4.

13. Interpretation of Theorem 7.4 as an oracle inequality

We conclude this second part by going back to the inductive case. We first give a weak variant of Theorem 7.1, in order to obtain an easily observable bound. We then use Theorem 7.4 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 right speed of convergence $N^{\frac{-2\beta}{2\beta+1}}$ up to a $\log N$ factor.

13.1. **A weak version of Theorem 7.1.** 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, and we simply take, for any x and θ:

$$f_\theta(x) = \theta(x),$$

that m is chosen and we still have:

$$\Theta_0 = (\theta_1, ..., \theta_m).$$

Moreover, let us assume that P is such that $Y_i = f(X_i) + \eta_i$ where η_i is independent of X_i and has an unknown distribution, with of course $P\eta = 0$ and $P(\eta^2) =\leq \sigma^2 < \infty$ with a known σ. We do not assume stronger hypothesis about η.

Theorem 13.1. *We have, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1-\varepsilon$, for any $k \in \{1, ..., m\}$:*

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

The proof is given at the end of the section. Note that this theorem is more general than Theorem 7.1 in the following way: we do not require the existence of exponential moments for the noise η_i. But, at least for large values of N, the bound is less tight.

13.2. **Rate of convergence of the estimator: the Sobolev space case.** Now, let us put:

$$\overline{\theta}_m = \arg \min_{\theta \in Span(\Theta_0)} R(\theta)$$

(that depends effectively on m by $\Theta_0 = \{\theta_1, ..., \theta_m\}$), 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:

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

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 $k \in \{1, ..., m\}$:

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

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

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

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 7.5 is just:

$$\hat{f}(x) = f_{\hat{\theta}}(x),$$

with:

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

Theorem 13.2. *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:

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

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 $\hat{f}$ (given in definition 7.5 with $n_0 = m$ here, build using the bound $\beta(\varepsilon, k)$ given in Theorem 13.1), with $\varepsilon = N^{-2}$ and $m = N$, is such that, for any $N \geq 2$,

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

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

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

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

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

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

13.3. **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 Härdle, Kerkyacharian, Picard and Tsybakov [21], or Donoho, Johnstone, Kerkyacharian and Picard [19]).

Theorem 13.3. *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* [19]*, 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:*

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

(with obious 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:

$$\{f_{\theta_1}, ..., f_{\theta_m}\} = \{\phi\} \cup \{\psi_{j,k}, j = 1, ..., 2^{\lfloor \frac{\log N}{\log 2} \rfloor}, k = 1, ..., 2^j\}$$

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

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

Let us remark that we obtain nearly the same rate of convergence than in [19], namely the minimax rate of convergence up to a $\log N$ factor.

13.4. **Proof of the theorems: Theorem 7.4 used as an oracle inequality.**

Proof of Theorem 13.1. Actually, the proof is quite straightforward: 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. There are several ways to perform this integration (see for example Catoni [10]), here we choose to apply a result obtained by Panchenko [31] that gives a particularly simple result here.

Lemma 13.4 (Panchenko [31], corollary 1). *Let us assume that we have i.i.d. variables $T_1, ..., T_N$ (with distribution P and values in $\mathbb{R}$) and an independent copy $T' = (T'_1, ..., T'_N)$ of $T = (T_1, ..., T_N)$. Let $\xi_j(T, T')$ for $j \in \{1, 2, 3\}$ be three measurable 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 $u > 0$:*

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

Then, for any $u > 0$:

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

The proof of this lemma can is given in the annex.

Now, a simple application of the first inequality of Lemma 10.1 (given in the transductive section) with $\varepsilon > 0$, any $k \in \{1, ..., m\}$, $g = id$, $\eta = 1 + \log\frac{2m}{\varepsilon}$ and:

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

leads us to the following bound, for any k:

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

or:

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

We now apply Panchenko lemma with:

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

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

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\left[f_{\theta_k}(X_i)Y_i - P\left[f_{\theta_k}(X)Y\right]\right]\right.$$

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

Remark finally that:

$$P\left[f_{\theta_k}(X)^2Y^2\right] \leq P\left[f_{\theta_k}(X)^2\right](B^2+\sigma^2),$$

and by the orthonormality property of the basis $(\theta_k)_{k\geq 1}$:

$$P\left[f_{\theta_k}(X)^2\right] = 1.$$

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

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

that ends the proof. □

Proof of Theorem 13.2. 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 13.1. We have:

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

First of all, it is obvious that:

$$\begin{aligned}
&P^{\otimes N}\left[\left(1-\mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\left\|\Pi_P^{\mathcal{F},m}\hat{f} - f\right\|_P^2\right] \\
&\qquad \leq 2P^{\otimes N}\left[\left(1-\mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\left(\left\|\Pi_P^{\mathcal{F},m}\hat{f}\right\|_P^2 + \|f\|_P^2\right)\right] \\
&\qquad\qquad \leq 2\varepsilon\left(B^2m + B^2\right) = 2\varepsilon(m+1)B^2.
\end{aligned}$$

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

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

This is where Theorem 7.4 has been used as an oracle inequality: the estimator that we have, with $m \geq m'$, is better than the one with the "good choice" m'. We have too:

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

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

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

The choice of:

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

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

Proof of Theorem 13.3. Here again let us write $\mathcal{E}(\varepsilon)$ the event satisfied with probability at least $1-\varepsilon$ in theorem 13.1. We have:

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

For the first term we still have:

$$\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2 \leq 2(m+1)B^2.$$

For the second term, let us write the development 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:

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

the estimator $\hat{f}$. Let us put:

$$J = 2^{\lfloor\frac{\log N}{\log 2}\rfloor}.$$

$$\left\|\Pi_P^{\mathcal{F},m}\hat{f}-f\right\|_P^2 \leq \left\|\hat{f}-f\right\|_P^2 = \left\|\Pi_P^{m,\varepsilon}...\Pi_P^{1,\varepsilon}0-f\right\|_P^2$$

$$= (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 + \sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2$$

$$\leq (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 \mathbb{1}(|\beta_{j,k}| \geq \kappa) + \sum_{j=0}^{J} \sum_{k=1}^{2^j} \beta_{j,k}^2 \mathbb{1}(|\beta_{j,k}| < \kappa)$$

$$+ \sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2$$

for any $\kappa \geq 0$, as soon as $\mathcal{E}(\varepsilon)$ is satisfied (here again we used theorem 7.4 as an oracle inequality). Now, we follow the technique used in [19] and [21] (see also the end of the third chapter in [9]). As soon as $\mathcal{E}(\varepsilon)$ is satisfied we have:

$$\sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 \mathbb{1}(|\beta_{j,k}| \geq \kappa) \leq \frac{8(B^2 + \sigma^2) \log \frac{2m}{\varepsilon}}{N} \sum_{j=0}^{J} \sum_{k=1}^{2^j} \mathbb{1}(|\beta_{j,k}| \geq \kappa)$$

$$\leq \frac{8(B^2 + \sigma^2) \log \frac{2m}{\varepsilon}}{N} \sum_{j=0}^{J} \sum_{k=1}^{2^j} \left(\frac{|\beta_{j,k}|}{\kappa} \right)^{\frac{2}{2s+1}}$$

$$= \frac{8(B^2 + \sigma^2) \log \frac{2m}{\varepsilon}}{N} \kappa^{-\frac{2}{2s+1}} \sum_{j=0}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^{\frac{2}{2s+1}}.$$

In the same way, we have:

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

So we have to upper bound:

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

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

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

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

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

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

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

for some C'' and so:

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

for some C'''. In the case where $p < 2$ we use (see [21], for $s > \frac{1}{p} - \frac{1}{2}$):

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

to obtain:

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

So we have:

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

Let us remember that:

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

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

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

to obtain the desired rate of convergence. □

Annex: proof of Panchenko's lemma

For the sake of completeness we give the proof of Panchenko's symmetrization result (Lemma 13.4) used in the last section.

The proof of this lemma uses another result.

Definition 13.1. Let us put, for any $a \in \mathbb{R}$:

$$\phi_a : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto (x-a)_+ = (x-a) \vee 0.$$

Note that for any $a \in \mathbb{R}$, Φ_a is nondecreasing and convex.

Lemma 13.5 (Panchenko, lemma 1). *Let $(\Omega, \mathcal{A})$ be a measurable space and P a probability measure on Ω. Let $X, X' : \Omega \to \mathbb{R}$ be real-valued random variables such that, for any $a \in \mathbb{R}$:*

$$P\left[\phi_a\left(X'\right)\right] \leq P\left[\phi_a\left(X\right)\right].$$

Let us assume moreover that there is some $\Gamma \leq 1$ and $\gamma > 0$ such that for any $t \geq 0$:

$$P\left(X \geq t\right) \leq \Gamma \exp\left(-\gamma t\right).$$

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

$$P\left(X' \geq t\right) \leq \Gamma \exp\left(1 - \gamma t\right).$$

Proof of Lemma 13.5. Note that there is nothing to prove for $t < 1/\gamma$ (the bound one the probability is then bigger than 1). For any $t \geq 1/\gamma$, for any $a \in \mathbb{R}$ such that $\Phi_a(t) > 0$ we have:

$$P\left(X' \geq t\right) \leq \frac{P\left[\Phi_a\left(X'\right)\right]}{\Phi_a(t)} \leq \frac{P\left[\Phi_a\left(X\right)\right]}{\Phi_a(t)}$$
$$= \frac{1}{\Phi_a(t)} \left[\Phi_a(0) + \int_0^\infty \Phi_a'(x) P\left(X \geq x\right) dx\right]$$
$$\leq \frac{1}{\Phi_a(t)} \left[\Phi_a(0) + \Gamma \int_0^\infty \Phi_a'(x) \exp\left(-\gamma x\right) dx\right].$$

Let us take:

$$a = t - \frac{1}{\gamma},$$

note that this choice is valid since:

$$\Phi_a(t) = \frac{1}{\gamma} > 0.$$

We also have $\Phi_a(0) = 0$, and so:

$$P\left(X' \geq t\right) \leq \frac{1}{\Phi_a(t)} \left[\Phi_a(0) + \Gamma \int_0^\infty \Phi_a'(x) \exp\left(-\gamma x\right) dx\right]$$
$$= \gamma\Gamma \int_{t-\frac{1}{\gamma}}^\infty \exp\left(-\gamma x\right) dx = \Gamma \exp\left(1 - \gamma t\right).$$

This ends the proof. □

Proof of Lemma 13.4. Let us remark that:

$$\left\{\xi_1(T,T') \geq \xi_2(T,T') + \sqrt{t\xi_3(T,T')}\right\}$$
$$= \left\{\xi_1(T,T') \geq \xi_2(T,T'), \quad \frac{\left(\xi_1(T,T') - \xi_2(T,T')\right)^2}{\xi_3(T,T')} \geq t\right\}$$
$$= \left\{\sup_{\delta>0} 4\delta\Big[\xi_1(T,T') - \xi_2(T,T') - \delta\xi_3(T,T')\Big] \geq t\right\}.$$

In the same way we obtain:

$$\left\{P\left[\xi_1(T,T')|T\right] \geq P\left[\xi_2(T,T')|T\right] + \sqrt{tP\left[\xi_2(T,T')|T\right]}\right\}$$
$$= \left\{\sup_{\delta>0} 4\delta\Big[P\left[\xi_1(T,T')|T\right] - P\left[\xi_2(T,T')|T\right] - \delta P\left[\xi_3(T,T')|T\right]\Big] \geq t\right\}.$$

Now, we put:

$$X = \sup_{\delta>0} 4\delta\Big[\xi_1(T,T') - \xi_2(T,T') - \delta\xi_3(T,T')\Big]$$

and:

$$X' = \sup_{\delta>0} 4\delta\Big[P\left[\xi_1(T,T')|T\right] - P\left[\xi_2(T,T')|T\right] - \delta P\left[\xi_3(T,T')|T\right]\Big].$$

We have:

$$X' = \sup_{\delta>0} P\Big[4\delta\left[\xi_1(T,T') - \xi_2(T,T') - \delta\xi_3(T,T')\right]\Big|T\Big]$$
$$\leq P\Big[\sup_{\delta>0} 4\delta\left[\xi_1(T,T') - \xi_2(T,T') - \delta\xi_3(T,T')\right]\Big|T\Big] = P\left(X|T\right).$$

So we have (remember that Φ_a is nondecreasing and convex):

$$P\left[\Phi_a(X')\right] = P\left\{\Phi_a\Big[P\left(X|T\right)\Big]\right\} \leq P\left\{P\Big[\Phi_a\left(X\right)|T\Big]\right\} = P\left[\Phi_a(X)\right]$$

for any $a \in \mathbb{R}$. So we can apply Lemma 13.5, we obtain that, under our hypothesis that:

$$P\left[\xi_1(T,T') \geq \xi_2(T,T') + \sqrt{t\xi_3(T,T')}\right]$$
$$= P\left(X \geq t\right) \leq \Gamma \exp\left(-\gamma t\right)$$

we have:

$$P\left(X' \geq t\right) \leq \Gamma \exp\left(1 - \gamma t\right).$$

This ends the proof. □

Part 3. Density estimation with quadratic loss: a confidence intervals method

We propose a feature selection method for density estimation with quadratic loss. This method relies on the study of unidimensional approximation models and on the definition of confidence regions for the density thanks to these models. It is quite general and includes cases of interest like detection of relevant wavelets coefficients or selection of support vectors in SVM. In the case of wavelets, we prove that this method is equivalent to a soft thresholding estimator, that is adaptative in the sense that it reaches the minimax rate of convergence (up to a log factor) under the assumption that the density has a given (but unknown to the statistician) regularity. In the case of SVM, we focus more particularly on the algorithmic aspect, as our method provides a theoretical justification to any reasonable heuristic for the choice of the set of support vectors, as well as the possibility to use several kernels simultaneously.

14. Introduction: the density estimation setting

14.1. **Notations.** Let us assume that we are given a measure space $(\mathcal{X}, \mathcal{B}, \lambda)$ where λ is positive and σ-finite, and a probability measure P on $(\mathcal{X}, \mathcal{B})$ such that P has a density with respect to λ:

$$P(dx) = f(x)\lambda(dx).$$

We assume that we observe a realization of the canonical process $(X_1, ..., X_N)$ on $(\mathcal{X}^N, \mathcal{B}^{\otimes N}, P^{\otimes N})$. Our objective here is to estimate f on the basis of the observations $X_1, ..., X_N$.

More precisely, let $\mathcal{L}^2(\mathcal{X}, \lambda)$ denote the Hilbert space of all measurable functions from $(\mathcal{X}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B}_R)$ where $\mathcal{B}_R$ is the Borel σ-algebra on $\mathbb{R}$. We will write $\mathcal{L}^2(\mathcal{X}, \lambda) = \mathcal{L}^2$ for short. Remark that $f \in \mathcal{L}^2$. Let us put, for any $(g, h) \in \left(\mathcal{L}^2\right)^2$:

$$d^2(g, h) = \int_{\mathcal{X}} \Big(g(x) - h(x)\Big)^2 \lambda(dx),$$

and let $\|.\|$ and $\langle ., . \rangle$ denote the corresponding norm and scalar product. We are here looking for an estimator $\hat{f}$ that tries to minimize our objective:

$$d^2(\hat{f}, f).$$

Let us choose an integer $m \in \mathbb{N}$ and a family of functions $(f_1, ..., f_m) \in \left(\mathcal{L}^2\right)^m$. There is no particular assumptions about this family: it is not necessarily linearly independent for example. In a first time, we assume that these functions are not data dependent, but we will see later in this last part how to include the case where $f_k(.) = K(X_k, .)$ for some kernel K for example.

14.2. **Objective.** Density estimation under quadratic loss is a classical problem in statistics and a lot of work has been done, we refer the reader to the general introduction by Tsybakov [39] and the references within for example. There is a wide range of applications, among them let us mention multiclass pattern recognition (by the estimation of the density of every class and then classification of a pattern by likelihood maximization) and image segmentation, see Zhang and al. [47] for example.

The objective here is to provide a practical algorithm to select and aggregate the functions f_k that are relevant to perform density estimation.

In the case of a wavelet basis, such algorithms are known and are based on coefficient thresholding, see Härdle, Kerkyacharian, Picard and Tsybakov [21] and the references within for an introduction. Under suitable hypotheses, these estimators are able to reach the minimax rate of convergence on spaces of function with a given

regularity β: $N^{-2\beta/(2\beta+1)}$, up to a $\log N$ term if β is unknown to the statistician in Donoho, Johnstone, Kerkyacharian and Picard [19]. We show that in this case our algorithm produces a soft-thresholded estimator that reaches the same rate of convergence.

We focus also particularly on the case of kernel methods and support vector machines (SVM). SVM are a class of learning algorithm introduced by Boser, Guyon and Vapnik in the case of classification, [7]. They were later generalized by Vapnik [41] to regression, and density estimation of a real-valued random variables with the Kolmogorov-Smirnov distance as a loss function:

$$d_{KS}^2(g,h) = \sup_{x\in\mathcal{X}} \left| \int_{-\infty}^{x} g(t)dt - \int_{-\infty}^{x} h(t)dt \right|.$$

Note that a lot of variants of SVM were introduced in order to modify the set of support vectors (of basis kernel functions used in the estimation of the function). For example Tipping [38] introduced Relevance Vector Machine: in this variant of SVM, the support vectors are meant to be close to the center of clusters of data. Blanchard, Bousquet, Massart and Zwald [6] proposed to perform a principal component analysis on the space induced by the kernel. Here, we generalize the definition of SVM for density estimation to the quadratic loss and propose a method that justifies the use of a wide range of heuristics to select the set of support vectors. The choice of a kernel is also of interest for practitioners. For example, the Gaussian kernel is very often used, but the choice of its parameter remains a problem. Algorithms using several kernels (for example the Gaussian kernel with different values for the parameter) were proposed, see for example Ratsch, Schafer, Scholkopf and Sorensen [33], often without theoretical justifications. Our method allows the use of multiple kernels.

The guarantee obtained here is that every selected feature actually improves the performance of the estimator: the quadratic distance to f decreases. Moreover the estimator is sparse, that means that often only a few of the functions f_k are actually selected. From this point of view the method can be seen as an implementation of Rissanen's MDL [4].

14.3. **Organization of the part.** The method is an adaptation to the case of density estimation of the method we proposed in part 2 for regression estimation. In a first time, we are going to study estimators of f in every unidimensional approximation model $\{\alpha f_k(.), \alpha \in \mathbb{R}\}$. Note that these models are too small and the obtained estimators do not have good properties in general. But they are used to obtained, by a PAC bound, confidence regions on f that have a very simple geometry. We then propose an iterative method that selects and aggregate such estimators in order to build a suitable estimator of f (section 15). For the sake of simplicity, we describe the method in this section for a family (f_k) that is not allowed to be data-dependent.

In section 16 we focus more particularly on the statistical point of view: we study the rate of convergence of the obtained estimator in the case of a basis of wavelets.

Section 17 is devoted to technical improvements and generalizations of the method. Improvements consists in more accurate PAC bounds leading to tighter confidence regions. Generalizations consists in including the case where the basis functions (f_k) are allowed to be data-dependent.

In section 18 we provide some simulations in order to compare the practical performances of our estimator with the density estimators described in [19].

Finally, section 19 is dedicated to the proofs of the theorem.

15. Estimation method

15.1. **Main hypothesis.** Until the end of the thesis we will refer to the following hypothesis about f and/or the basis functions $f_k, k \in \{1, ..., m\}$.

Definition 15.1. We will say that f and $(f_1, ..., f_m)$ satisfies the conditions $\mathcal{H}(p)$ for $1 < p < +\infty$ if, for:

$$\frac{1}{p} + \frac{1}{q} = 1,$$

there exists some $(c, c_1, ..., c_m) \in \left(\mathbb{R}_+^*\right)^{m+1}$ (known to the statistician) such that:

$$\forall k \in \{1, ..., m\}, \quad \left(\int_{\mathcal{X}} |f_k|^{2p} \lambda(dx)\right)^{\frac{1}{p}} \leq c_k \int_{\mathcal{X}} |f_k|^2 \lambda(dx)$$

$$\text{and} \quad \left(\int_{\mathcal{X}} |f|^q \lambda(dx)\right)^{\frac{1}{p}} \leq c \int_{\mathcal{X}} |f| \lambda(dx) \quad (= c).$$

For $p = 1$ the condition $\mathcal{H}(1)$ is: f is bounded by a (known) constant c and we put $c_1 = ... = c_k = 1$. For $p = +\infty$ the condition $\mathcal{H}(+\infty)$ is just that every $|f_k|$ is bounded by

$$\sqrt{c_k \int_{\mathcal{X}} f_k(x)^2 \lambda(dx)}$$

where c_k is known, and we put $c = 1$. In any case, we put, for any k:

$$C_k = c_k c.$$

Note that this condition is not very restrictive, actually $\mathcal{H}(+\infty)$ does not require any information about f and just imposes condition on the family $(f_k)_{k=1...m}$ to be chosen by the statistician.

Definition 15.2. We put, for any $k \in \{1, ..., m\}$:

$$D_k = \int_{\mathcal{X}} |f_k|^2 \lambda(dx) = d^2(f_k, 0) = \|f_k\|^2.$$

15.2. **Unidimensional models.** Let us choose $k \in \{1, ..., m\}$ and consider the unidimensional model $\mathcal{M}_k = \{\alpha f_k(.), \alpha \in \mathbb{R}\}$. Remark that the orthogonal projection (denoted by $\Pi_{\mathcal{M}_k}$) of f on $\mathcal{M}_k$ is known, it is namely:

$$\Pi_{\mathcal{M}_k} f(.) = \overline{\alpha}_k f_k(.)$$

where:

$$\overline{\alpha}_k = \arg\min_{\alpha \in \mathbb{R}} d^2(\alpha f_k, f) = \frac{\int_{\mathcal{X}} f_k(x) f(x) \lambda(dx)}{\int_{\mathcal{X}} f_k(x)^2 \lambda(dx)} = \frac{\int_{\mathcal{X}} f_k(x) f(x) \lambda(dx)}{D_k}.$$

A natural estimator of this coefficient is:

$$\hat{\alpha}_k = \frac{\frac{1}{N} \sum_{i=1}^N f_k(X_i)}{\int_{\mathcal{X}} f_k(x)^2 \lambda(dx)},$$

because we expect to have, by the law of large numbers:

$$\frac{1}{N} \sum_{i=1}^N f_k(X_i) \xrightarrow[N \to \infty]{a.s.} P\left[f_k(X)\right] = \int_{\mathcal{X}} f_k(x) f(x) \lambda(dx).$$

Actually, we can formulate a more precise result.

Theorem 15.1. *Let us assume that condition $\mathcal{H}(p)$ holds for some $p \in [1, +\infty]$. Then for any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\Bigg\{\forall k \in \{1,...,m\}, d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k) \leq \left\{\frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\right\}\left[\frac{\frac{1}{N}\sum_{i=1}^N f_k(X_i)^2}{D_k} + C_k\right]\Bigg\} \geq 1-\varepsilon.$$

The proof is given in section 19, more precisely in subsection 19.1 page 143.

15.3. **The selection algorithm.** Until the end of this section we assume that $\mathcal{H}(p)$ is satisfied for some $1 \leq p \leq +\infty$.

Let $\beta(\varepsilon, k)$ denote the upper bound for the model k in Theorem 15.1:

$$\forall \varepsilon > 0, \forall k \in \{1,...,m\}: \quad \beta(\varepsilon,k) = \left\{\frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\right\}\left[\frac{\frac{1}{N}\sum_{i=1}^N f_k(X_i)^2}{D_k} + C_k\right].$$

Let us put:

$$\mathcal{CR}_{k,\varepsilon} = \left\{g \in \mathcal{L}^2, d^2(\hat{\alpha}_k f_k, \Pi_{\mathcal{M}_k} g) \leq \beta(\varepsilon, k)\right\}.$$

Then Theorem 15.1 implies the following result.

Corollary 15.2. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, f \in \mathcal{CR}_{k,\varepsilon}\right\} \geq 1-\varepsilon.$$

So for any k, $\mathcal{CR}_{k,\varepsilon}$ is a confidence region at level k for f. Moreover, $\mathcal{CR}_{k,\varepsilon}$ being convex we have the following corollary.

Corollary 15.3. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, \forall g \in \mathcal{L}^2, d^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, f) \leq d^2(g,f)\right\} \geq 1-\varepsilon.$$

It just means that for any g, $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$ is a better estimator than g. Note that g and k being given, it is easy to compute explicitly $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$, this is done in the remark concluding this subsection, remark 15.1 page 131.

So we propose the following algorithm (generic form):

- we choose ε and start with $g_0 = 0$;
- at each step n, we choose an indice $k(n)$ using any heuristic we want, it is of course allowed to be data-dependent, then we take:

$$g_{n+1} = \Pi_{\mathcal{CR}_{k(n),\varepsilon}} g_n;$$

- we choose a stopping time n_s in any convenient way and take:

$$\hat{f} = g_{n_s}.$$

So corollary 15.3 implies that:

$$P^{\otimes N}\left\{d^2(\hat{f}, f) = d^2(g_{n_s}, f) \leq ... \leq d^2(g_0, f) = d^2(0, f)\right\} \geq 1-\varepsilon.$$

Actually, a more accurate version of corollary 15.3 can give an idea of the way to choose $k(n)$ in the algorithm. Let us use corollary 15.2 and remember the fact that each $\mathcal{CR}_{k,\varepsilon}$ is convex.

Corollary 15.4. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, \forall g \in \mathcal{L}^2, d^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, f) \leq d^2(g,f) - d^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, g)\right\} \geq 1-\varepsilon.$$

This suggests to choose as $k(n)$ the direction k such that $d^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g_n, g_n)$ is maximal. This is very close to the greedy algorithms already used in the context of regression estimation, see Barron, Cohen, Wahmen and DeVore [3] for example. Note however that this is not necessarily the optimal choice.

This leads us to the following version of our previous algorithm:

- we choose ε and $0 < \kappa \leq 1/N$ and start with $g_0 = 0$;
- at each step n, we take:

$$k(n) = \arg \max_{k \in \{1,\dots,m\}} d^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g_n, g_n)$$

and:

$$g_{n+1} = \Pi_{\mathcal{CR}_{k(n),\varepsilon}} g_n;$$

- we take:

$$n_s = \inf\left\{n \in \mathbb{N} : \quad d^2(g_n, g_{n-1}) \leq \kappa\right\}$$

and:

$$\hat{f} = g_{n_s}.$$

Corollary 15.4 implies that:

$$P^{\otimes N}\left\{d^2(\hat{f},f) \leq d^2(0,f) - \sum_{n=0}^{n_s-1} d^2(g_n, g_{n+1})\right\} \geq 1-\varepsilon.$$

Remark 15.1. Given g and k, the computation of $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$, is quite easy. First, note that $\Pi_{\mathcal{CR}_{k,\varepsilon}} g = g + b f_k$ so we just have to compute the coefficient b. Moreover, the conditions $\Pi_{\mathcal{CR}_{k,\varepsilon}} g \in \mathcal{CR}_{k,\varepsilon}$ gives:

$$\left\| \frac{\langle g + b f_k, f_k \rangle}{\|f_k\|^2} f_k - \hat{\alpha}_k f_k \right\|^2 \leq \beta(\varepsilon, k)$$

or:

$$\left(\frac{< g, f_k >}{\|f_k\|^2} + b - \hat{\alpha}_k \right)^2 \leq \frac{\beta(\varepsilon,k)}{\|f_k\|^2}.$$

There are two possibilities. If this condition is satisfied for $b = 0$, this means that $g \in \mathcal{CR}_{k,\varepsilon}$ and so $\Pi_{\mathcal{CR}_{k,\varepsilon}} g = g$. Otherwise, $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$ will lie on the boundary of $\mathcal{CR}_{k,\varepsilon}$, this means that b will satisfy:

$$\frac{< g, f_k >}{\|f_k\|^2} + b - \hat{\alpha}_k \in \left\{ \pm \frac{\sqrt{\beta(\varepsilon,k)}}{\|f_k\|} \right\}.$$

Finally, note that $|b|$ should be minimal. This leads to the following formula:

$$\begin{aligned} b &= \hat{\alpha}_k - \frac{< g, f_k >}{\|f_k\|^2} - sgn\left(\hat{\alpha}_k - \frac{< g, f_k >}{\|f_k\|^2}\right) \frac{\sqrt{\beta(\varepsilon,k)}}{\|f_k\|} \\ &= \hat{\alpha}_k - \frac{< g, f_k >}{D_k} - sgn\left(\hat{\alpha}_k - \frac{< g, f_k >}{D_k}\right) \sqrt{\frac{\beta(\varepsilon,k)}{D_k}} \end{aligned}$$

where sgn is the sign function given by $sgn(x) = \mathbb{1}_{\mathbb{R}_+}(x) - \mathbb{1}_{\mathbb{R}_-}(x)$.

15.4. **An example: the histogram.** Here, we just make things more explicit in a classical example in statistics, the histogram, so the reader more interested in kernel methods can skip this subsection. Let us assume that λ is a finite measure and let $A_1, ..., A_m$ be a partition of $\mathcal{X}$. We put, for any $k \in \{1, ..., m\}$:

$$f_k(.) = \mathbb{1}_{A_k}(.).$$

Remark that:

$$D_k = \int_{\mathcal{X}} f_k(x)^2 \lambda(dx) = \lambda(A_k),$$

and that condition $\mathcal{H}(+\infty)$ is satisfied with constants:

$$c_k = \frac{1}{\lambda(A_k)}$$

and (as we have the convention $c = 1$ in this case) $C_k = c_k c = c_k$.

In this context we have:

$$\overline{\alpha}_k = \frac{P(X \in A_k)}{\lambda(A_k)},$$

$$\hat{\alpha}_k = \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{A_k}(X_i)}{\lambda(A_k)},$$

$$\beta(\varepsilon, k) = \left\{ \frac{4\left[1 + \log \frac{2m}{\varepsilon}\right]}{N\lambda(A_k)} \right\} \left[\frac{1}{N} \sum_{i=1}^N f_k(X_i)^2 + 1 \right].$$

Finally, note that all the confidence regions $\mathcal{CR}_{k,\varepsilon}$ are orthogonal in this case. So the order of projection does not affect the obtained estimator here, and we can take:

$$\hat{f} = \Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0.$$

We have:

$$\hat{f}(x) = \sum_{k=1}^m \left(\hat{\alpha}_k - \sqrt{\frac{\beta(\varepsilon, k)}{\lambda(A_k)}} \right)_+ f_k(x)$$

where, for any $y \in \mathbb{R}$ we have: $(y)_+ = \max(y, 0) = y \vee 0$.

In this case corollary 15.4 (page 131) becomes:

$$P^{\otimes N}\left\{ d^2(\hat{f}, f) \leq d^2(0, f) - \sum_{k=1}^m \left(\hat{\alpha}_k - \sqrt{\frac{\beta(\varepsilon, k)}{\lambda(A_k)}} \right)_+^2 \lambda(A_k) \right\} \geq 1 - \varepsilon.$$

15.5. **Remarks on the intersection of the confidence regions.** Actually, corollary 15.2 (page 130) could motivate another method. Note that:

$$\forall k \in \{1, ..., m\}, f \in \mathcal{CR}_{k,\varepsilon} \Leftrightarrow f \in \bigcap_{k=1}^m \mathcal{CR}_{k,\varepsilon}.$$

Definition 15.3. Let us put, for any $I \subset \{1, ..., m\}$:

$$\mathcal{CR}_{I,\varepsilon} = \bigcap_{k \in I} \mathcal{CR}_{k,\varepsilon},$$

and:

$$\hat{f}_I = \Pi_{\mathcal{CR}_{I,\varepsilon}} 0.$$

The estimator $\hat{f}_{\{1,...,m\}}$ can be reached by solving the following optimization problem:

$$\min_{g \in \mathcal{L}^2} \|g\|^2,$$

$$s.t. \quad \forall k \in \{1, ..., m\} :$$
$$\begin{cases} \langle g - \hat{\alpha}_k f_k, f_k \rangle - \sqrt{D_k \beta(\varepsilon, k)} \leq 0, \\ -\langle g - \hat{\alpha}_k f_k, f_k \rangle - \sqrt{D_k \beta(\varepsilon, k)} \leq 0. \end{cases}$$

The problem can be solved in dual form:

$$\max_{\gamma \in \mathbb{R}^m} \left[-\sum_{i=1}^{m} \sum_{k=1}^{m} \gamma_i \gamma_k \langle f_i, f_k \rangle + 2 \sum_{k=1}^{m} \gamma_k \hat{\alpha}_k \|f_k\|^2 - 2 \sum_{k=1}^{m} |\gamma_k| \sqrt{D_k \beta(\varepsilon, k)} \right].$$

with solution $\gamma^* = (\gamma_1^*, ..., \gamma_m^*)$ and:

$$\hat{f}_{\{1,...,m\}} = \sum_{k=1}^{m} \gamma_k^* f_k.$$

Note the similarity with usual SVM algorithms (at least in the case where we consider data-dependent kernel functions, as we will do in section 17), but with one major difference: we replaced the kernel induced scalar product by the scalar product associated to the distance $d(.,.)$ that we try to minimize.

From a statistical point of view, as:

$$-\sum_{i=1}^{m} \sum_{k=1}^{m} \gamma_i^* \gamma_k^* \langle f_i, f_k \rangle = \|f^*\|^2$$

and:

$$2 \sum_{k=1}^{m} \gamma_k^* \hat{\alpha}_k \|f_k\|^2 = 2 \sum_{k=1}^{m} \gamma_k^* \frac{\frac{1}{N} \sum_{i=1}^{N} f_k(X_i)}{\|f_k\|^2} \|f_k\|^2 = \frac{2}{N} \sum_{i=1}^{N} f^*(X_i)$$

we can see this as a penalized maximization of the likelihood.

16. Some classical examples in statistics with rates of convergence

16.1. **General remarks when $(f_k)_k$ is an orthonormal family and condition $\mathcal{H}(1)$ is satisfied.** In subsections 16.1, 16.2 and 16.3, we study the rate of convergence of our estimator in the special case where $(f_k)_{k \in \mathbb{N}^*}$ is an orthonormal basis of $\mathcal{L}^2$, so we have:

$$D_k = \int_{\mathcal{X}} f_k(x)^2 \lambda(dx) = 1$$

and:

$$\int_{\mathcal{X}} f_k(x) f_{k'}(x) \lambda(dx) = 0$$

if $k \neq k'$.

We also assume that condition $\mathcal{H}(1)$ is satisfied: $\forall x \in \mathcal{X}, f(x) \leq c$, remember that in this case we have taken $c_k = 1$ and so $C_k = c$, so:

$$\beta(\varepsilon, k) = \left\{ \frac{4 \left[1 + \log \frac{2m}{\varepsilon} \right]}{N} \right\} \left[\frac{1}{N} \sum_{i=1}^{N} f_k(X_i)^2 + c \right].$$

Note that in this case all the order of application of the projections $\Pi_{\mathcal{CR}_{k,\varepsilon}}$ does not matter because these projections works on orthogonal directions. So we can define, once m is chosen:

$$\hat{f} = \Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0 = \Pi_{\mathcal{CR}_{\{1,...,m\},\varepsilon}} 0.$$

Note that:

$$\hat{f}(x) = \sum_{k=1}^{m} sgn(\hat{\alpha}_k) \left(|\hat{\alpha}_k| - \sqrt{\beta(\varepsilon, k)} \right)_+ f_k(x),$$

and so $\hat{f}$ is a soft-thresholded estimator. Let us also make the following remark. As for any x, $f(x) \leq c$, we have:

$$d^2(f, 0) \leq c.$$

So the region:

$$\mathcal{B} = \left\{ g \in \mathcal{L}^2 : \forall k \in \mathbb{N}^*, \int_{\mathcal{X}} g(x) f_k(x) \lambda(dx) \leq \sqrt{c} \right\}$$

is convex, and contains f. So the projection on $\mathcal{B}$, $\Pi_{\mathcal{B}}$ can only improve $\hat{f}$. We put:

$$\tilde{f} = \Pi_{\mathcal{B}} \hat{f}. \tag{16.1}$$

Note that this transformation is needed to obtain the following theorem, but does not have practical incidence in general. Actually:

$$\tilde{f}(x) = \sum_{k=1}^{m} sgn(\hat{\alpha}_k) \left\{ \left(|\hat{\alpha}_k| - \sqrt{\beta(\varepsilon, k)} \right)_+ \wedge \sqrt{c} \right\} f_k(x),$$

where we let $a \wedge b$ denote $\min(a, b)$ for any $(a, b) \in \mathbb{R}^2$.

16.2. **Rate of convergence in Sobolev spaces.** It is well known that if f has regularity β (known by the statistician) then we have the choice

$$m = N^{\frac{1}{2\beta+1}}$$

and a standard estimation of coefficients leads to the optimal rate of convergence:

$$N^{\frac{-2\beta}{2\beta+1}}.$$

Here, we assume that we don't know β, and we show that taking $m = N$ leads to the rate of convergence:

$$N^{\frac{-2\beta}{2\beta+1}} \log N$$

namely the optimal rate of convergence up to a $\log N$ factor.

Theorem 16.1. *Let us assume that $(f_k)_{k \in \mathbb{N}^*}$ is an orthonormal basis of $\mathcal{L}^2$. Let us put:*

$$\overline{f}_m = \arg \min_{g \in Span(f_1, \ldots, f_m)} d^2(g, f),$$

and let us assume that $f \in \mathcal{L}^2$ satisfies condition $\mathcal{H}(1)$ and is such that there are unknown constants $D > 0$ and $\beta \geq 1$ such that:

$$d^2(\overline{f}_m, f) \leq D m^{-2\beta}.$$

Let us choose $m = N$ and $\varepsilon = N^{-2}$ in the definition of $\tilde{f}$. Then we have, for any $N \geq 2$:

$$P^{\otimes N} d^2(\tilde{f}, f) \leq D'(c, D) \left(\frac{\log N}{N} \right)^{\frac{2\beta}{2\beta+1}},$$

where $\tilde{f}$ is the estimator defined by Equation 16.1.

The proof is given in subsection 19.2 page 148. Let us just remark that, in the case where $\mathcal{X} = [0, 1]$, λ is the Lebesgue measure, and $(f_k)_{k \in \mathbb{N}^*}$ is the trigonometric basis, the condition:

$$d^2(\overline{f}_m, f) \leq D m^{-2\beta}$$

is satisfied for $D = D(\beta, L) = L^2 \pi^{-2\beta}$ as soon as $f \in W(\beta, L)$ where $W(\beta, L)$ is the Sobolev class:

$$\left\{ f \in \mathcal{L}^2 : f^{(\beta-1)} \text{ is absolutely continuous and } \int_0^1 f^{(\beta)}(x)^2 \lambda(dx) \leq L^2 \right\},$$

see Tsybakov [39] for example. The minimax rate of convergence in $W(\beta, L)$ is $N^{-\frac{2\beta}{2\beta+1}}$, so we can see that our estimator reaches the best rate of convergence up to a $\log N$ factor with an unknown β.

16.3. **Rate of convergence in Besov spaces.** We here extend the previous result to the case of a Besov space $B_{s,p,q}$. Note that we have, for any $L \geq 0$ and $\beta \geq 0$:

$$W(\beta, L) \subset B_{\beta,2,2}$$

so this result is really an extension of the previous one (see Härdle, Kerkyacharian, Picard and Tsybakov [21], or Donoho, Johnstone, Kerkyacharian and Picard [19]). We define the Besov space:

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

with obvious changes for $p = +\infty$ or $q = +\infty$. We also define the weak Besov space:

$$W_{\rho,\pi} = \left\{ g : [0,1] \to \mathbb{R}, \quad g(.) = \alpha\phi(.) + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}\psi_{j,k}(.), \right.$$
$$\left. \sup_{\lambda>0} \lambda^{\rho} \sum_{j=0}^{\infty} 2^{j\left(\frac{\pi}{2}-1\right)} \sum_{k=1}^{2^j} \mathbb{1}_{\{|\beta_{j,k}|>\lambda\}} < +\infty \right\}$$
$$= \left\{ g : [0,1] \to \mathbb{R}, \quad g(.) = \alpha\phi(.) + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}\psi_{j,k}(.), \right.$$
$$\left. \sup_{\lambda>0} \lambda^{\pi-\rho} \sum_{j=0}^{\infty} 2^{j\left(\frac{\pi}{2}-1\right)} \sum_{k=1}^{2^j} |\beta_{j,k}|^{\pi} \mathbb{1}_{\{|\beta_{j,k}|\leq\lambda\}} < +\infty \right\},$$

see Cohen [15] for the equivalence of both definitions. Let us remark that $B_{s,p,q}$ is a set of functions with regularity s while $W_{\rho,\pi}$ is a set of functions with regularity:

$$s' = \frac{1}{2}\left(\frac{\pi}{\rho} - 1\right).$$

Theorem 16.2. *Let us assume that $\mathcal{X} = [0,1]$, 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* [19] *and having regularity R (for example Daubechies' families), with ϕ and $\psi_{0,1}$ supported by $[-A, A]$. Let us assume that $f \in B_{s,p,q}$ with $R+1 \geq s > \frac{1}{p}$, $1 \leq q \leq \infty$, $2 \leq p \leq +\infty$, or that $f \in B_{s,p,q} \cap W_{\frac{2}{2s+1},2}$ with $R+1 \geq s > \frac{1}{p}$, $1 \leq p \leq +\infty$, with unknown constants s, p and q and that f satisfies condition $\mathcal{H}(1)$ with a known constant c. Let us choose:*

$$\{f_1, ..., f_m\} = \{\phi\} \cup \{\psi_{j,k}, j = 1, ..., 2^{\lfloor \frac{\log N}{\log 2} \rfloor}, k = 1, ..., 2^j\}$$

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

$$P^{\otimes N} d^2(\tilde{f}, f) = \mathcal{O}\left(\left(\frac{\log N}{N} \right)^{\frac{2s}{2s+1}} \right),$$

where $\tilde{f}$ is the estimator defined by Equation 16.1 (page 134).

The proof of this theorem is also given in subsection 19.2 page 148. Let us remark that we obtain nearly the same rate of convergence as in [19], namely the minimax rate of convergence up to a $\log N$ factor.

16.4. **Kernel estimators.** Here, we assume that $\mathcal{X} = \mathbb{R}$ and that f is compactly supported, say by $[0,1]$. We put, for any $m \in \mathbb{N}$ and $k \in \{1, ..., m\}$:

$$f_k(x) = K\left(\frac{k}{m}, x\right)$$

where K is some function $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and we obtain some estimator that has the form of a kernel estimator:

$$\hat{f}_{\{1,...,m\}}(x) = \sum_{k=1}^{m} \tilde{\alpha}_k K\left(\frac{k}{m}, x\right).$$

Moreover, is is possible to use a multiple kernel estimator. Let us choose $n \in \mathbb{N}$, $h \in \mathbb{N}$, h kernels $K_1, ..., K_h$ and put, for any $k = i + n * j \in \{1, ..., m = hn\}$:

$$f_k(x) = K_j\left(\frac{i}{n}, x\right).$$

We obtain a multiple kernel estimator:

$$\hat{f}_{\{1,...,m\}}(x) = \sum_{i=1}^{n} \sum_{j=1}^{h} \tilde{\alpha}_{i+nj} K_j\left(\frac{i}{n}, x\right).$$

However, note that the use of kernel functions is more justified in large dimension where we will take as basis functions: $K_j(X_i, .)$, namely data-dependent functions. We show in the next section that our algorithm can be extended to this case.

17. Improvements and generalization of Theorem 15.1

It appears in simulations that the bound on $d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k)$, as given by Theorem 15.1, has to be very sharp if we want to obtain a good estimator. Actually, as pointed out by Catoni [10], the symmetrization technique used in the proof of Theorem 15.1 causes the loss of a factor 2 in the bound because we upper bound the variance of two samples instead of 1. So it is possible to obtain sharper bounds. In this section, we try to use this remark to improve our bound, using techniques already used by Catoni [9].

We then remark that a technique due to Seeger [37] allows to include the case of data-dependent basis functions (f_k) and to deal with SVM in particular.

First, remark that the estimation technique described in section 15 does not necessarily require a bound on $d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k)$. Actually, a simple confidence interval on $\overline{\alpha}_k$ is sufficient.

17.1. **An improvement of Theorem 15.1 under condition $\mathcal{H}(+\infty)$.** Let us remember that $\mathcal{H}(+\infty)$ just means that every f_k is bounded by $\sqrt{C_k D_k}$.

Theorem 17.1. *Under condition $\mathcal{H}(+\infty)$, for any $\varepsilon > 0$, for any $\beta_{k,1}, \beta_{k,2}$ such that:*

$$0 < \beta_{k,j} < \frac{N}{\sqrt{C_k D_k}}, \quad j \in \{1, 2\},$$

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

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) \leq \overline{\alpha}_k \leq \alpha_k^{\sup}(\varepsilon, \beta_{k,2})$$

with:

$$\alpha_k^{\sup}(\varepsilon, \beta_{k,2}) = \frac{N - N \exp\left[\frac{1}{N}\sum_{i=1}^{N} \log\left(1 - \frac{\beta_{k,2}}{N} f_k(X_i)\right) - \frac{\log \frac{2m}{\varepsilon}}{N}\right]}{D_k \beta_{k,2}}$$

and:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) = \frac{N \exp\left[\frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \frac{\beta_{k,1}}{N} f_k(X_i)\right) - \frac{\log \frac{2m}{\varepsilon}}{N}\right] - N}{D_k \beta_{k,1}}.$$

The proof is given in subsection 19.3 page 150.

First, let us see why this theorem really improves Theorem 15.1. Let us define:

$$V_k = P\left\{[f_k(X) - P(f_k(X))]^2\right\}$$

and let us choose:

$$\beta_{k,1} = \beta_{k,2} = \sqrt{\frac{N \log \frac{2m}{\varepsilon}}{V_k}}.$$

Then we obtain:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) = \hat{\alpha}_k - \frac{1}{D_k}\sqrt{\frac{2V_k \log \frac{2m}{\varepsilon}}{N}} + \mathcal{O}_P\left(\frac{\log \frac{2m}{\varepsilon}}{N}\right)$$

and:

$$\alpha_k^{\sup}(\varepsilon, \beta_{k,2}) = \hat{\alpha}_k + \frac{1}{D_k}\sqrt{\frac{2V_k \log \frac{2m}{\varepsilon}}{N}} + \mathcal{O}_P\left(\frac{\log \frac{2m}{\varepsilon}}{N}\right).$$

So, the first order term for $d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k)$ is:

$$\frac{2V_k \log \frac{2m}{\varepsilon}}{D_k N},$$

there is an improvement by a factor 4 when we compare this bound to Theorem 15.1.

Remark that this particular choice for $\beta_{k,1}$ and $\beta_{k,2}$ is valid as soon as:

$$\sqrt{\frac{N \log \frac{2m}{\varepsilon}}{V_k}} < \frac{N}{\sqrt{C_k D_k}}$$

or equivalently as soon as N is greater than

$$\frac{C_k D_k \log \frac{2m}{\varepsilon}}{V_k}.$$

In practice, however, this particular $\beta_{k,1}$ and $\beta_{k,2}$ are unknown. We can use the following procedure (see Catoni [10]). We choose a value $a > 1$ and:

$$B = \left\{a^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{\sqrt{C_k D_k}}}{\log a} \right\rfloor - 1\right\}.$$

By taking a union bound over all possibles values of B, with:

$$|B| \leq \frac{\log \frac{N}{\sqrt{C_k D_k}}}{\log a}$$

we obtain the following corollary.

Corollary 17.2. *Under condition $\mathcal{H}(+\infty)$, for any $a > 1$, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$ we have:*

$$\sup_{\beta \in B} \alpha_k^{\inf}\left(\frac{\varepsilon \log a}{\log N - \frac{1}{2}\log C_k D_k}, \beta\right) \leq \overline{\alpha}_k \leq \inf_{\beta \in B} \alpha_k^{\sup}\left(\frac{\varepsilon \log a}{\log N - \frac{1}{2}\log C_k D_k}, \beta\right),$$

with:

$$B = \left\{a^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{\sqrt{C_k D_k}}}{\log a} \right\rfloor - 1\right\}.$$

Note that the price to pay for the optimization with respect to $\beta_{k,1}$ and $\beta_{k,2}$ was just a $\log\log N$ factor.

17.2. **A generalization to data-dependent basis functions.** We now extend the previous method to the case where the family $(f_1, ..., f_m)$ is allowed to be data-dependent, in a particular sense. This subsection requires some modifications of the notations of section 15.

Definition 17.1. For any $m' \in \mathbb{N}^*$ we define a function $\Theta_{m'} : \mathcal{X} \to \left(\mathcal{L}^2\right)^{m'}$. For any $i \in \{1, ..., N\}$ we put:

$$\Theta_{m'}(X_i) = (f_{i,1}, ..., f_{i,m'}) .$$

Finally, consider the family of functions:

$$(f_1, ..., f_m) = (f_{1,1}, ..., f_{1,m'}, ..., f_{N,1}, ..., f_{N,m'}) .$$

So we have $m = m'N$ (of course, m' is allowed to depend on N). Let us take, for any $i \in \{1, ..., N\}$:

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

We put, for any $(i,k) \in \{1, ..., N\} \times \{1, ..., m'\}$:

$$D_{i,k} = \int_{\mathcal{X}} f_{i,k}(x)^2 \lambda(dx),$$

and we still assume that condition $\mathcal{H}(\infty)$ is satisfied, that means here that we have known constants $C_{i,k} = c_{i,k}$ such that:

$$\forall x \in \mathcal{X}, \quad |f_{i,k}(x)| \leq \sqrt{C_{i,k} D_{i,k}}.$$

Finally, we put:

$$\overline{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} d^2(\alpha f_{i,k}, f).$$

Theorem 17.3. *For any $\varepsilon > 0$, for any $\beta_{i,k,1}, \beta_{i,k,2}$ such that:*

$$0 < \beta_{i,k,j} < \frac{N-1}{\sqrt{C_{i,k} D_{i,k}}}, \quad j \in \{1,2\},$$

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

$$\tilde{\alpha}_k^{\inf}(\varepsilon, \beta_{i,k,1}) \leq \overline{\alpha}_k \leq \tilde{\alpha}_k^{\sup}(\varepsilon, \beta_{i,k,2})$$

with:

$$\tilde{\alpha}_k^{\sup}(\varepsilon, \beta_{i,k,2}) = \frac{N - 1 - (N-1)\exp\left[\frac{1}{N-1}\sum_{j \neq i} \log\left(1 - \frac{\beta_{i,k,2}}{N-1} f_{i,k}(X_j)\right) - \frac{\log \frac{2m'N}{\varepsilon}}{N-1}\right]}{D_{i,k}\beta_{i,k,2}}$$

and:

$$\tilde{\alpha}_k^{\inf}(\varepsilon, \beta_{i,k,1}) = \frac{(N-1)\exp\left[\frac{1}{N-1}\sum_{j \neq i} \log\left(1 + \frac{\beta_{i,k,1}}{N-1} f_{i,k}(X_j)\right) - \frac{\log \frac{2m'N}{\varepsilon}}{N-1}\right] - N + 1}{D_{i,k}\beta_{i,k,1}}.$$

The proof of this theorem is also given in section 19 (subsection 19.3 page 150).

Example 17.1 (Multiple kernel SVM). We propose the following choice:

$$\Theta_{m'}(X_i) = \Big\{K_1(X_i,.),...,K_{m'}(X_i,.)\Big\}$$

for some family of functions $\mathcal{X}^2 \to \mathbb{R}$: $(K_1,...,K_{m'})$. Note that we have $m = m'N$. In this case, the estimator is under the form:

$$\forall x \in \mathcal{X}, \quad \hat{f}(x) = \sum_{j=1}^{m'} \sum_{i=1}^{N} \tilde{\alpha}_{i,j} K_j(X_i, x),$$

and the number of $\tilde{\alpha}_{i,j} \neq 0$ is expected to be small. This estimator has the form of a SVM (if $h = 1$ and K_1 is a Mercer's kernel). However, if we take the general form of the algorithm, we can see that it is possible to use whatever heuristic to choose the next pair (i,j), and so we can use a wide range of methods to choose the set of support vectors.

For example, if N is large, we can use only the following method:

- use a clustering algorithms on the data to obtain c clusters;
- at each step, try to use only one vector from each cluster, for example the one that is the closest to the mean point of the cluster.

In this case have only to try c projection instead of N. This proposition is suggested by Tipping's Relevance Vector Machine [38].

One of the most used kernels is the Gaussian kernel. If $\mathcal{X}$ is a metric space, with a distance $\delta(.,.)$, we choose $\gamma \in \mathbb{R}_+^*$ and we put:

$$K(x,x') = \exp\left[-\gamma\delta^2(x,x')\right].$$

In practice, the choice of γ is problematic. Here, we can choose a grid of values $(\gamma_1,...,\gamma_h) \in (\mathbb{R}_+^*)^h$ and take:

$$K_j(x,x') = \exp\left[-\gamma_j\delta^2(x,x')\right]$$

and let the algorithm selects the relevant values of γ_j.

Note that in this case, hypothesis $\mathcal{H}(\infty)$ is obviously satisfied. If $\mathcal{X} = \mathbb{R}$, $\delta(x,x') = |x - x'|$ and λ is the Lebesgue measure we have:

$$C_{i,j} = c_{i,j} = \sqrt{\frac{\gamma_j}{\pi}} = \frac{1}{D_{i,j}},$$

with obvious adaptations of the notations.

17.3. **The histogram example continued.** We apply here the improved bounds in the case of the histogram introduced in subsection 15.4 page 132.

In the case of the histogram, $f_k(.) = \mathbb{1}_{A_k}(.)$ can take only two values: 0 and 1. Remember that $D_k = \lambda(A_k)$. So:

$$\alpha_k^{\inf}(\varepsilon,\beta_{k,1}) = \frac{N}{\lambda(A_k)\beta_{k,1}} \left\{ \left[\left(1 + \frac{\beta_{k,1}}{N}\right)^{|\{i:X_i\in A_k\}|} \frac{\varepsilon}{2m} \right]^{\frac{1}{N}} - 1 \right\}.$$

Remember that, for any $x \geq 0$:

$$(1+x)^\gamma \geq 1 + \gamma x + \frac{\gamma(\gamma-1)}{2}x^2$$

and so:

$$\alpha_k^{\inf}(\varepsilon,\beta_{k,1}) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} \left[1 - \frac{\beta_{k,1}(1-\hat{\alpha}_k D_k)}{2N}\right] - \frac{N}{D_k\beta_{k,1}} \left[1 - \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}}\right].$$

Now, we take the grid:

$$B = \left\{ 2^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{\sqrt{D_k}}}{\log 2} \right\rfloor - 1 \right\}.$$

Remark that, for any β in:

$$\left[1, \frac{N}{2\sqrt{D_k}}\right]$$

there is some $b \in B$ such that $\beta \leq b \leq 2\beta$, and so:

$$\alpha_k^{\inf}(\varepsilon, b) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} \left[1 - \frac{\beta_{k,1}(1 - \hat{\alpha}_k D_k)}{2N}\right] - \frac{N}{D_k 2\beta_{k,1}} \left[1 - \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}}\right].$$

This allows us to choose whatever value for $\beta_{k,1}$ in

$$\left[1, \frac{N}{2\sqrt{D_k}}\right].$$

Let us choose:

$$\beta_{k,1} = \sqrt{\frac{N^2 \left[\left(\frac{\varepsilon}{2m}\right)^{\frac{-1}{N}} - 1\right]}{\hat{\alpha}_k D_k (1 - \hat{\alpha}_k D_k)}}$$

that is allowed for N large enough. So we have:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} - \sqrt{\hat{\alpha}_k D_k (1 - \hat{\alpha}_k D_k) \left[\left(\frac{\varepsilon}{2m}\right)^{\frac{-1}{N}} - 1\right]}.$$

With the union bound term (over the grid B) we obtain:

$$\begin{aligned}
&\alpha_k^{\inf}\left(\frac{\varepsilon \log 2}{\log \frac{N}{\sqrt{D_k}}}, \beta_{k,1}\right) \\
&\quad \geq \hat{\alpha}_k \left(\frac{\varepsilon \log 2}{2m \log \frac{N}{\sqrt{D_k}}}\right)^{\frac{1}{N}} - \sqrt{\hat{\alpha}_k D_k (1 - \hat{\alpha}_k D_k) \left[\left(\frac{\varepsilon \log 2}{2m \log \frac{N}{\sqrt{D_k}}}\right)^{\frac{-1}{N}} - 1\right]} \\
&\quad = \hat{\alpha}_k - \sqrt{\frac{\hat{\alpha}_k D_k (1 - \hat{\alpha}_k D_k) \log \frac{2m \log \frac{N}{\sqrt{D_k}}}{\varepsilon \log 2}}{N}} + \mathcal{O}\left(\frac{\log \frac{m \log N}{\varepsilon}}{N}\right),
\end{aligned}$$

remark that we have this time the "real" variance term of $\mathbb{1}_{A_k}(X)$:

$$\hat{\alpha}_k D_k (1 - \hat{\alpha}_k D_k) = \frac{|\{i : X_i \in A_k\}|}{N} \left(1 - \frac{|\{i : X_i \in A_k\}|}{N}\right).$$

17.4. **Another simple example: the Haar basis.** Let us assume that $\mathcal{X} = [0, 1]$. Let (φ, ψ) be a father wavelet and the associated mother wavelet, and:

$$\psi_{j,k}(x) = \psi(2^j x + k)$$

for $k \in \{0, ..., 2^j - 1\} = S_j$ (note that the wavelet basis is non-normalized here). Here, we use the Haar wavelets, with:

$$\begin{aligned}
\varphi(x) &= \mathbb{1}_{[0,1]}(x) \\
\psi(x) &= \mathbb{1}_{\left[0, \frac{1}{2}\right]}(x) - \mathbb{1}_{\left[\frac{1}{2}, 1\right]}(x).
\end{aligned}$$

For the sake of simplicity, let us write:

$$\psi_{-1,k}(x) = \varphi(x)$$

for $k \in \{0\} = S_{-1}$. By an obvious adaptation of our notations, let us put $\overline{\alpha}_{j,k}$ the coefficient associated to $\psi_{j,k}$:

$$\overline{\alpha}_{j,k} = \frac{P\psi_{j,k}(X)}{\int \psi_{j,k}^2} = P\psi_{j,k}(X),$$

remark that condition $\mathcal{H}(\infty)$ is satisfied with $D_{j,k} = 2^{-j}$ and $C_{j,k} = 1$. In this particular setting, note that $\overline{\alpha}_{-1,0} = 1$ is known, so the associated confidence interval is just $\{1\}$. Moreover, here $\psi_{j,k}(X)$ can take only three values: -1, 0 and 1. Let us put:

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}.$$

Remark that in this case we have:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \log\left(1 - \frac{\beta}{N}\psi_{j,k}(X_i)\right) \\
&\quad = \overline{P}(\psi_{j,k}(X) = 1)\log\left(1 - \frac{\beta}{N}\right) + \overline{P}(\psi_{j,k}(X) = -1)\log\left(1 + \frac{\beta}{N}\right) \\
&\quad = \frac{1}{2}\overline{P}\left[\psi_{j,k}(X)^2\right]\log\left(1 - \frac{\beta^2}{N^2}\right) + \frac{1}{2}\overline{P}\left[\psi_{j,k}(X)\right]\log\left(\frac{1 - \frac{\beta}{N}}{1 + \frac{\beta}{N}}\right).
\end{aligned}$$

So we have:

$$\begin{aligned}
\alpha_{j,k}^{\sup}(\varepsilon, \beta) = \frac{1}{D_k \beta_{k,2}}\Bigg\{ N - N\exp\Bigg[&\frac{1}{2}\overline{P}\left[\psi_{j,k}(X)^2\right]\log\left(1 - \frac{\beta^2}{N^2}\right) \\
&- \frac{1}{2}\overline{P}\left[\psi_{j,k}(X)\right]\log\left(\frac{1 + \frac{\beta}{N}}{1 - \frac{\beta}{N}}\right) - \frac{\log\frac{2m}{\varepsilon}}{N}\Bigg]\Bigg\}
\end{aligned}$$

and:

$$\begin{aligned}
\alpha_{j,k}^{\inf}(\varepsilon, \beta) = \frac{1}{D_k \beta_{k,1}}\Bigg\{ N\exp\Bigg\{ &\frac{1}{2}\overline{P}\left[\psi_{j,k}(X)^2\right]\log\left(1 - \frac{\beta^2}{N^2}\right) \\
&+ \frac{1}{2}\overline{P}\left[\psi_{j,k}(X)\right]\log\left(\frac{1 + \frac{\beta}{N}}{1 - \frac{\beta}{N}}\right) - \frac{\log\frac{2m}{\varepsilon}}{N}\Bigg] - N\Bigg\}.
\end{aligned}$$

18. Simulations

18.1. **Description of the example.** We assume that we observe X_i for $i \in \{1, ..., N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0, 1] \subset \mathbb{R}$ are i.i.d. from a distribution with an unknown density f with respect to the Lebesgue measure. The goal is to estimate f.

Here, we will use three methods. The first estimation method will be a multiple kernel estimator obtained by our method, the second one a thresholded wavelets estimate also obtained by this algorithm, and we will compare both estimators to a thresholded wavelet estimate as given by Donoho, Johnstone, Kerkyacharian and Picard [18].

18.2. **The estimators.**

18.2.1. *Hard-thresholded wavelet estimator.* We first use a classical hard-thresholded wavelet estimator.

In the case of the Haar basis (see subsection 17.4), we take:

$$\hat{\alpha}_{j,k} = 2^j \frac{1}{N} \sum_{j=1}^{N} \psi_{j,k}(X_i).$$

For a given $\kappa \geq 0$ and $J \in \mathbb{N}$, we take:

$$\tilde{f}_J(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} \hat{\alpha}_{j,k} \mathbb{1}(|\hat{\alpha}_{j,k}| \geq \kappa t_{j,N}) \psi_{j,k}(.)$$

where:

$$t_{j,N} = \sqrt{\frac{j}{N}}.$$

Actually, we must choose J in such a way that:

$$2^J \sim t_N^{-1}.$$

Here, we choose $\kappa = 0.7$ and $J = 7$.

18.2.2. *Wavelet estimators with our algorithm.* We also use the same family of functions, and we apply our thresholding method, with bounds given in subsection 17.4 page 140. So we take:

$$m = 2^J = 128.$$

We use an asymptotic version of our confidence intervals inspired by our theoretical confidence intervals:

$$\overline{\alpha}_{j,k} \in \left[\hat{\alpha}_{j,k} \pm \sqrt{2 \frac{\log \frac{2m}{\varepsilon} V_{j,k}}{N}} \right]$$

where $V_{j,k}$ is the estimated variance of $\psi_{j,k}(X)$:

$$V_{j,k} = \frac{1}{N} \sum_{i=1}^{N} \left[\psi_{j,k}(X_i) - \frac{1}{N} \sum_{h=1}^{N} \psi_{j,k}(X_h) \right]^2.$$

Let us remark that the union bound are always "pessimistic", and that we use a union bound argument over all the m models despite only a few of them are effectively used in the estimator. So, we propose to actually use the individual confidence interval for each model, replacing: the $\log \frac{2m}{\varepsilon}$ by $\log \frac{2}{\varepsilon}$.

18.2.3. *Multiple kernel estimator.* Finally, we use the kernel estimator described in section 16 page 133, with function K:

$$K_j(u,v) = \exp\left[-2^{2j}(u-v)^2\right]$$

with $n = N$ and $j \in \{1, ..., h = 6\}$. We add the constant function 1 to the family.

Here again we use the individuals confidence intervals, and the asymptotic version of this intervals.

TABLE 4. Values of t_i and c_i in the fonction $Blocks(.)$.

i	1	2	3	4	5	6	7	8	9	10	11
c_i	4	-5	3	-4	5	4.2	-2.1	4.3	-3.1	2.1	-4.2
t_i	0.10	0.13	0.15	0.23	0.25	0.40	0.44	0.65	0.76	0.78	0.81

TABLE 5. Results of the experiments. For each experiment, we give the mean of the distance of the estimator the density ($d^2(.,f)$) on the experiences (and the standard deviations).

Function $f(.)$	standard thresholded wavelets		thresh. wav. with our method		multiple kernel	
Doppler	0.101	(0.0172)	0.125	(0.0085)	0.081	(0.0079)
HeaviSine	0.065	(0.0115)	0.061	(0.0071)	0.039	(0.0106)
Blocks	0.110	(0.0216)	0.142	(0.0097)	0.121	(0.0206)

18.3. **Experiments and results.** The simulations were realized with the R software [32].

For the experiments, we use the following functions f that are some variations of the functions used by Donoho and Johnstone for experiments on wavelets, for example in [18] (actually, these functions were used as regression functions, so the modification was to add them a constant in order to ensure they take nonnegative values):

$$Doppler(t) = 1 + 2\sqrt{t(1-t)}\sin\frac{2\pi(1+v)}{t+v} \quad \text{where } v = 0.05$$

$$HeaviSine(t) = 1.5 + \frac{1}{4}\Big[4\sin 4\pi t - sgn(t-0.3) - sgn(0.72-t)\Big]$$

$$Blocks(t) = 1.05 + \frac{1}{4}\sum_{i=1}^{11} c_i \mathbb{1}_{(t_i,+\infty)}(t)$$

where the values of the c_i and t_i are given in Table 4.

We consider 3 experiments (for the three density functions), we choose ε=10%, repeat each experiment 20 times.

18.4. **Results and comments.** The results are reported in Table 5. We also give some illustrations (Figures 6, 7 and 8). The experiments are very simple, note however the following facts.

First, the wavelet estimator obtained by our method give results that are comparable to the thresholded wavelets estimator. Moreover, note that the choice the threshold κ in the tresholded wavelets estimator done in this experiment cannot be done in practice as we choose the value that gave the better results in the experiments. In practice, κ is arbitrary and this leads to lower performances.

Then, multiple kernels SVM performs far better than the other estimators, except on the last function (*Blocks*) but note how the Haar basis seems particularly well adapted for the approximation of this function.

Finally, looking at the standard deviation values, we note the interesting fact that estimators obtained with our method are more "stable". This can be explained by the fact that we did not try to minimize the expectation of the distance of our estimator to f, but rather to control this quantity with high probability.

19. PROOFS

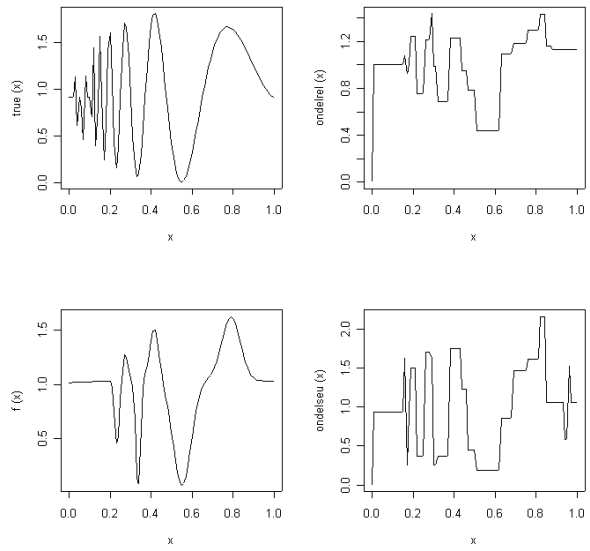

FIGURE 6. Experiment 1, $f = Doppler$. Up-left: true regression function (*true*). Down-left: SVM (f). Up-right: wavelet estimate with our algorithm (*ondelrel*). Down-right: "classical" wavelet estimate (*ondelseu*).

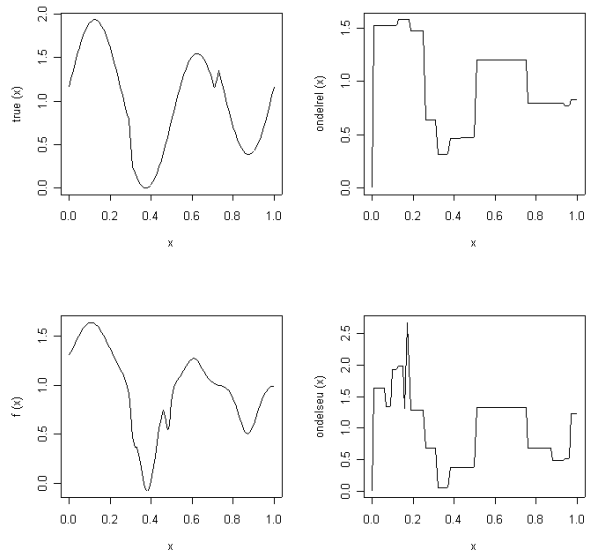

FIGURE 7. Experiment 2, $f = HeaviSine$.

19.1. **Proof of Theorem 15.1 of section 15.** Before we give the proof, let us state two lemmas. The first one is a variant of a lemma by Catoni [10], the second one is due to Panchenko [31].

Lemma 19.1. *Let $(T_1, ..., T_{2N})$ be a random vector taking values in $\mathbb{R}^{2N}$ distributed according to a distribution $\mathcal{P}^{\otimes 2N}$. For any $\eta \in \mathbb{R}$, for any measurable function*

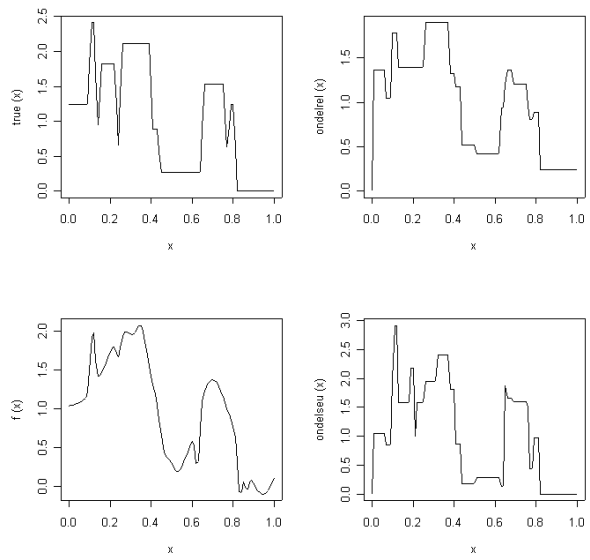

FIGURE 8. Experiment 3, $f = Blocks$.

$\lambda : \mathbb{R}^{2N} \to \mathbb{R}_+^*$ *that is exchangeable with respect to its* $2 \times 2N$ *arguments:*

$$\mathcal{P}^{\otimes 2N} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \Big\{ T_{i+N} - T_i \Big\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right) \leq \exp\left(-\eta\right)$$

and the reverse inequality:

$$\mathcal{P}^{\otimes 2N} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \Big\{ T_i - T_{i+N} \Big\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right) \leq \exp\left(-\eta\right),$$

where we write:

$$\eta = \eta\left(T_1, ..., T_{2N}\right)$$
$$\lambda = \lambda\left(T_1, ..., T_{2N}\right)$$

for short.

Proof of Lemma 19.1. In order to prove the first inequality, we write:

$$\mathcal{P}^{\otimes 2N} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \Big\{ T_{i+N} - T_i \Big\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right)$$
$$= \mathcal{P}^{\otimes 2N} \exp\left(\sum_{i=1}^{N} \log\cosh\left\{ \frac{\lambda}{N} \left(T_{i+N} - T_i\right) \right\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right).$$

We now use the inequality:

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

We obtain:

$$\log\cosh\left\{ \frac{\lambda}{N} \left(T_{i+N} - T_i\right) \right\} \leq \frac{\lambda^2}{2N^2} \left(T_{i+N} - T_i\right)^2 \leq \frac{\lambda^2}{N^2} \left(T_{i+N}^2 + T_i^2\right).$$

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

Lemma 19.2 (Panchenko [31], corollary 1). *Let us assume that we have i.i.d. variables $T_1, ..., T_N$ (with distribution $\mathcal{P}$ and values in $\mathbb{R}$) and an independent copy $T' = (T_{N+1}, ..., T_{2N})$ of $T = (T_1, ..., T_N)$. Let $\xi_j(T, T')$ for $j \in \{1, 2, 3\}$ be three measurable 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 $u > 0$:*

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

Then, for any $u > 0$:

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

The proof of this lemma can be found in Panchenko's paper, [31]. We can now give the proof of Theorem 15.1.

Proof of Theorem 15.1. Let $(X_{N+1}, ..., X_{2N})$ be an independent copy of our sample $(X_1, ..., X_N)$. Let us choose $k \in \{1, ..., m\}$. Let us apply Lemma 19.1 with $\mathcal{P} = P$ and, for any $i \in \{1, ..., 2N\}$:

$$T_i = f_k(X_i).$$

We obtain, for any measurable function $\eta_k \in \mathbb{R}$, for any measurable function $\lambda_k : \mathbb{R}^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments:

$$P^{\otimes 2N}\exp\left(\frac{\lambda_k}{N}\sum_{i=1}^{N}\Big\{f_k(X_{i+N}) - f_k(X_i)\Big\} - \frac{\lambda_k^2}{N^2}\sum_{i=1}^{2N} f_k(X_i)^2 - \eta_k\right) \leq \exp\left(-\eta_k\right)$$

and the reverse inequality:

$$P^{\otimes 2N}\exp\left(\frac{\lambda_k}{N}\sum_{i=1}^{N}\Big\{f_k(X_i) - f_k(X_{i+N})\Big\} - \frac{\lambda_k^2}{N^2}\sum_{i=1}^{2N} f_k(X_i)^2 - \eta_k\right) \leq \exp\left(-\eta_k\right)$$

as well. This implies that:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i) - f_k(X_{i+N})\Big\} \leq \frac{\lambda_k}{N^2}\sum_{i=1}^{2N} f_k(X_i)^2 + \frac{\eta_k}{\lambda_k}\right] \leq \exp\left(-\eta_k\right)$$

and:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_{i+N}) - f_k(X_i)\Big\} \leq \frac{\lambda_k}{N^2}\sum_{i=1}^{2N} f_k(X_i)^2 + \frac{\eta_k}{\lambda_k}\right] \leq \exp\left(-\eta_k\right).$$

Let us choose:

$$\lambda_k = \sqrt{\frac{N^2\eta_k}{\sum_{i=1}^{2N} f_k(X_i)^2}}$$

in both inequalities, we obtain for the first one:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i) - f_k(X_{i+N})\Big\} \geq 2\sqrt{\frac{\eta_k\sum_{i=1}^{2N} f_k(X_i)^2}{N^2}}\right] \leq \exp\left(-\eta_k\right).$$

We now apply Lemma 19.2 with the same $T_i = f_k(X_i)$, $\eta_k = u$, $A = 1$, $a = 1$, $\xi_2 = 0$,

$$\xi_1 = \frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i) - f_k(X_{i+N})\Big\} \quad \text{and}$$

$$\xi_3 = \frac{4\sum_{i=1}^{2N} f_k(X_i)^2}{N^2}.$$

We obtain:

$$P^{\otimes N}\left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i) - P\left[f_k(X)\right] \geq 2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + P\left[f_k(X)^2\right]\right\}}{N}}\right]$$

$$= P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i) - P\left[f_k(X)\right] \geq 2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + P\left[f_k(X)^2\right]\right\}}{N}}\right]$$

$$\leq \exp\left(1-\eta_k\right).$$

Remark that:

$$P\left[f_k(X)^2\right] = \int_{\mathcal{X}} f_k(x)^2 f(x)\lambda(dx).$$

So, using condition $\mathcal{H}(p)$ and Hölder's inequality we have:

$$P\left[f_k(X)^2\right] \leq \left(\int_{\mathcal{X}} |f_k(x)|^{2p}\,\lambda(dx)\right)^{\frac{1}{p}}\left(\int_{\mathcal{X}} f(x)^q\lambda(dx)\right)^{\frac{1}{q}}$$

$$\leq \left(c_k\int_{\mathcal{X}} f_k(x)^2\lambda(dx)\right)\left(c\int_{\mathcal{X}} f(x)\lambda(dx)\right)$$

$$= (c_k c)\int_{\mathcal{X}} f_k(x)^2\lambda(dx) = C_k D_k.$$

Now, let us combine this inequality with the reverse one by a union bound argument, we have:

$$P^{\otimes N}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f_k(X_i) - P\left[f_k(X)\right]\right|\right.$$

$$\left.\geq 2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + C_k D_k\right\}}{N}}\right] \leq 2\exp\left(1-\eta_k\right).$$

We now make a union bound on $k \in \{1,...,m\}$ and put:

$$\eta_k = 1 + \log\frac{2m}{\varepsilon}.$$

We obtain:

$$P^{\otimes N}\left[\forall k \in \{1,...,m\},\quad \left|\frac{1}{N}\sum_{i=1}^{N} f_k(X_i) - P\left[f_k(X)\right]\right|\right.$$

$$\left.\leq 2\sqrt{\frac{\left(1+\log\frac{2m}{\varepsilon}\right)\left\{\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + C_k D_k\right\}}{N}}\right] \geq 1-\varepsilon.$$

We end the proof by noting that:

$$d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k) = \frac{\left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i) - P\left[f_k(X)\right]\right]^2}{\int_{\mathcal{X}} f_k(x)^2\lambda(dx)}.$$

□

19.2. **Proof of the theorems of section 16.**

Proof of Theorem 16.1. 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 write $\mathcal{E}(\varepsilon)$ the event:

$$\mathcal{E}(\varepsilon) = \left\{ \forall k \in \{1, ..., m\}, d^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k) \right. \\ \left. \leq \left\{ \frac{4\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} \right\} \left[\frac{\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2}{D_k} + C_k \right] \right\} \tag{19.1}$$

satisfied with probability at least $1 - \varepsilon$ according to Theorem 15.1. We have:

$$P^{\otimes N} d^2(\tilde{f}, f) = P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)} d^2(\tilde{f}, f) \right] + P^{\otimes N}\left[\left(1 - \mathbb{1}_{\mathcal{E}(\varepsilon)}\right) d^2(\tilde{f}, f) \right].$$

For the first term we have:

$$d^2(\tilde{f}, f) \leq 2\int_{\mathcal{X}} f(x)^2 \lambda(dx) + 2\int_{\mathcal{X}} \tilde{f}(x)^2 \lambda(dx) \leq 2c + 2mc = 2(m+1)c$$

and so:

$$P^{\otimes N}\left[\left(1 - \mathbb{1}_{\mathcal{E}(\varepsilon)}\right) d^2(\hat{f}, f) \right] \leq 2\varepsilon(m+1)c.$$

For the other term, just remark that under $\mathcal{E}(\varepsilon)$:

$$d^2(\tilde{f}, f) = d^2(\Pi_{\mathcal{B}} \Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0, f) \\ \leq d^2(\Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0, f) \leq d^2(\Pi_{\mathcal{CR}_{m',\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0, f)$$

for any $m' \leq m$, because of Theorem 15.1, more precisely of corollary 15.3 page 130. And we have:

$$d^2(\Pi_{\mathcal{M}_{m'}} ... \Pi_{\mathcal{M}_1} 0, f) \\ \leq \sum_{k=1}^{m'} \left\{ \frac{4\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} \right\} \left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + c \right] + d^2(\overline{f}_m, f).$$

So we have:

$$P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)} d^2(\tilde{f}, f) \right] \leq P^{\otimes N}\left[d^2(\tilde{f}, f) \right] \\ \leq P^{\otimes N} \sum_{k=1}^{m'} \left\{ \frac{4\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} \right\} \left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + c \right] + (m')^{-2\beta} D \\ \leq \frac{8m'c\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} + (m')^{-2\beta} D.$$

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

$$P^{\otimes N} d^2(\tilde{f}, f) \leq \frac{8m'c\left[1 + \log \frac{2m}{\varepsilon}\right]}{N} + (m')^{-2\beta} D + 2\varepsilon(m+1)c.$$

The choice of:

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

leads to a first term of order $N^{\frac{-2\beta}{2\beta+1}} \log \frac{m}{\varepsilon} (\log N)^{-\frac{1}{2\beta+1}}$ and a second term of order $N^{\frac{-2\beta}{2\beta+1}} (\log N)^{\frac{2\beta}{2\beta+1}}$. The choice of $m = N$ and $\varepsilon = N^{-2}$ gives a first and second term at order:

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

while keeping the third term at order N^{-1}. This proves the theorem. □

Proof of Theorem 16.2. Let C be a generic constant in the whole proof. We keep the notation $\mathcal{E}(\varepsilon)$ of the preceding proof (Equation 19.1 page 148). We have:

$$P^{\otimes N} d^2(\tilde{f}, f) = P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)} d^2(\tilde{f}, f)\right] + P^{\otimes N}\left[\left(1 - \mathbb{1}_{\mathcal{E}(\varepsilon)}\right) d^2(\tilde{f}, f)\right].$$

For the first term we still have:

$$d^2(\tilde{f}, f) \leq 2(m+1)c.$$

For the second term, let us write the development 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:

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

the estimator $\hat{f}$. Let us put:

$$J = \left\lfloor \frac{\log N}{\log 2} \right\rfloor.$$

For any $J' \leq J$ we have:

$$\begin{aligned} d^2(\tilde{f}, f) = d^2(\Pi_{\mathcal{B}} \Pi_{\mathcal{CR}_{m,\varepsilon}} \ldots \Pi_{\mathcal{CR}_{1,\varepsilon}} 0, f) &\leq d^2(\Pi_{\mathcal{CR}_{m,\varepsilon}} \ldots \Pi_{\mathcal{CR}_{1,\varepsilon}} 0, f) \\ &= (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 + \sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2 \\ \leq (\tilde{\alpha} - \alpha)^2 + \sum_{j=0}^{J'} \sum_{k=1}^{2^j} (\tilde{\beta}_{j,k} - \beta_{j,k})^2 \mathbb{1}(|\beta_{j,k}| \geq \kappa) &+ \sum_{j=0}^{J'} \sum_{k=1}^{2^j} \beta_{j,k}^2 \mathbb{1}(|\beta_{j,k}| < \kappa) \\ &+ \sum_{j=J'+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2 \end{aligned}$$

for any $\kappa \geq 0$, as soon as $\mathcal{E}(\varepsilon)$ is satisfied (here again we applied Theorem 15.1). In the case where $p \geq 2$ we can take:

$$J' = \left\lfloor \frac{\log N^{\frac{1}{1+2s}}}{\log 2} \right\rfloor$$

and $\kappa = 0$ to obtain:

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

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

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

and so:

$$\sum_{j=J+1}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k}^2 \leq C 2^{-2J's} \leq C N^{\frac{-2s}{1+2s}},$$

and:

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

So we obtain the desired rate of convergence. In the case where $p < 2$ we let $J' = J$ and proceed as follows.

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

because f is also assumed to be in the weak Besov space. We also have:

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

For the remainder term we use (see [21, 19]):

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

to obtain:

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

as $s > \frac{1}{p}$. Let us remember that:

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

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

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

to obtain the desired rate of convergence. □

19.3. **Proof of the theorems of section 17.**

Proof of Theorem 17.1. The technique used in the proof is due to Catoni [11]. Let us choose $k \in \{1, ..., m\}$, and:

$$\beta \in \left(0, \frac{N}{\sqrt{C_k D_k}}\right).$$

We have, for any $\eta \in \mathbb{R}$:

$$P^{\otimes N} \exp\left\{\sum_{i=1}^{N} \log\left(1 - \frac{\beta}{N} f_k(X_i)\right) - \eta\right\} \leq \exp\left\{N \log\Big(1 - \frac{\beta}{N} P\left[f_k(X)\right]\Big) - \eta\right\}.$$

Let us choose:

$$\eta = \log \frac{2m}{\varepsilon} + N \log\Big(1 - \frac{\beta}{N} P\left[f_k(X)\right]\Big).$$

We obtain:

$$P^{\otimes N} \exp\left\{\sum_{i=1}^{N} \log\left(1 - \frac{\beta}{N} f_k(X_i)\right) - \log \frac{2m}{\varepsilon} - N \log\Big(1 - \frac{\beta}{N} P\left[f_k(X)\right]\Big)\right\} \leq \frac{\varepsilon}{2m},$$

and so:

$$P^{\otimes N}\left\{\sum_{i=1}^{N} \log\left(1 - \frac{\beta}{N} f_k(X_i)\right) \geq \log \frac{2m}{\varepsilon} + N \log\Big(1 - \frac{\beta}{N} P\left[f_k(X)\right]\Big)\right\} \leq \frac{\varepsilon}{2m},$$

that becomes:

$$P^{\otimes N}\left\{P\left[f_k(X)\right] \geq \frac{N}{\beta}\left[1 - \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log\left(1 - \frac{\beta}{N} f_k(X_i)\right) - \frac{\log \frac{2m}{\varepsilon}}{N}\right]\right)\right\} \leq \frac{\varepsilon}{2m}.$$

We apply the same technique to:

$$P^{\otimes N} \exp\left\{\sum_{i=1}^{N} \log\left(1 + \frac{\beta'}{N} f_k(X_i)\right) - \eta\right\} \leq \exp\left\{N \log\Big(1 + \frac{\beta'}{N} P\left[f_k(X)\right]\Big) - \eta\right\}$$

to obtain the upper bound. We combine both result by a union bound argument. □

Proof of Theorem 17.3. Let us choose $(i,k) \in \{1,...,N\} \times \{1,...,m'\}$. Using Seeger's idea, we follow the preceding proof, replacing $P^{\otimes N}$ by P_i, and using the $N-1$ random variables:

$$\Big(f_{i,k}(X_j)\Big)_{\substack{j \in \{1,\ldots,N\} \\ j \neq i}}$$

with

$$\eta = \log \frac{2m'N}{\varepsilon} + (N-1) \log\left(1 - \frac{\beta}{N-1} P\left[f_{i,k}(X)\right]\right)$$

and we obtain:

$$P_i \exp\left\{\sum_{j \neq i} \log\left(1 - \frac{\beta}{N-1} f_{i,k}(X_j)\right) - \log \frac{2m'N}{\varepsilon} - (N-1) \log\Big(1 - \frac{\beta}{N-1} P\left[f_{i,k}(X)\right]\Big)\right\} \leq \frac{\varepsilon}{2m'N}.$$

Note that for any random variable H that is a function of the X_i:

$$P^{\otimes N} P_i H = P^{\otimes N} H.$$

So we conclude exactly in the same way as in the proof of the previous theorem and we obtain the claimed result. □

References

[1] J.-Y. Audibert, Aggregated Estimators and Emprical Complexity for Least Square Regression, *Annales de l'Institut Henri Poincaré (B): Probability and Statistics*, vol. 40, issue 6, pp. 685-736 (2004).

[2] J.-Y. Audibert, *PAC-Bayesian Statistical Learning Theory*, PhD Thesis, University Paris VI, 2004.

[3] A. Barron, A. Cohen, W. Dahmen and R. DeVore, Adaptative Approximation and Learning by Greedy Algorithms, *preprint, 2006.*

[4] A. Barron, J. Rissanen and B. Yu, The minimum description length principle in coding and modeling, *IEEE Trans. Information Theory*, vol. 44 (1998), no. 6, pp. 2743-2760.

[5] L. Birgé, An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, 113-133, Leiden, 1999.

[6] G. Blanchard, P. Massart, R. Vert and L. Zwald, Kernel Projection Machine: a New Tool for Pattern Recognition, *Proceedings of NIPS* 2004.

[7] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144-152. ACM Press, 1992.

[8] S. Boucheron, O. Bousquet and G. Lugosi, Theory of classification: some recent advances, *ESAIM Probability & Statistics*, 9:323-375, 2005.

[9] O. Catoni, Statistical learning theory and stochastic optimization, *Lecture notes, Saint-Flour summer school on Probability Theory, 2001*, Springer.

[10] O. Catoni, A PAC-Bayesian approach to adaptative classification, *preprint Laboratoire de Probabilités et Modèles Aléatoires 2003.*

[11] O. Catoni, PAC-Bayesian Inductive and Transductive Learning, manuscript, 2006.

[12] O. Catoni, Improved Vapnik Cervonenkis bounds, *preprint Laboratoire de Probabilités et Modèles Aléatoires 2005.*

[13] A. J. Cervonenkis and V. N. Vapnik, On the uniform convergence of relative frequencies of events to their probabilities, *Doklady Akademii Nauk USSR*, vol. 181(4) (1968).

[14] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning*, 20:273-297, 1995.

[15] A. Cohen, Wavelet methods in numerical analysis, in *Handbook of numerical analysis*, vol. VII, pages 417-711, North-Holland, Amsterdam, 2000.

[16] N. Cristianini and J. Shawe Taylor, *An introduction to Support Vector Machines and other kernel based learning methods,* Cambridge University Press, 2000.

[17] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New-York, 1996.

[18] D. L. Donoho and I. M. Johnstone, Ideal Spatial Adaptation by Wavelets, *Biometrika*, Vol. 81, No. 3 (Aug., 1994), 425-455.

[19] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, Density Estimation by Wavelet Thresholding, *Annals of Statistics, 1996*, 24: 508-539.

[20] G. M. Fung, O. L. Mangasarian and A. J. Smola, Minimal Kernel Classifiers, *Journal of Machine Learning Research*, No. 3 (2002).

[21] W. Härdle, G. Kerkyacharian, D. Picard and A. B. Tsybakov, *Wavelets, Approximations and Statistical Applications, 1998*, Lectude Notes in Statistics, Springer.

[22] T. Hastie, R. Tibshirani and J. Friedman, *The Elements Of Statistical Learning*, Springer, New York, 2001.

[23] W. Hoeffding, Probability Inequalities for Sums of Bounded Random Variables, *Journal of the American Statistical Association*, 58 (1963), 13-30.

[24] G. Kerkyacharian and D. Picard, Regression in random desgin and warped wavelets, *preprint Laboratoire de Probabilités et Modèles aléatoires 2003*

[25] J. Langford and J. Shawe-Taylor, PAC-Bayes and Margins, *Proceedinds of Neural Information Processing Systems (NIPS) 2002*, Vancouver (tutorials and conference) and Whistler (workshops).

[26] O. V. Lepski, E. Mammen and V. G. Spokoiny, Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernels Estimates with Variable Bandwidth Selectors, *The Annals Of Statistics*, Vol. 25, No. 31 (1997).

[27] N. Littlestone and M. Warmuth, Relating data compression and learnability. *Technical report,* University of California, Santa Cruz, 1986.

[28] D. A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230–234 (electronic), ACM, New York, 1998.

[29] D. A. McAllester, PAC-Bayesian Model Averaging, *Proceedings of the Twelth Annual Conference On Computational Learning Theory (Santa Cruz, CA, 1999)*, 164-170 (electronic), ACM, New York, 1999.

[30] C. McDiarmid, Concentration, *Probabilistic Methods for Algorithmic Discrete Mathematics*, Habib C., McDiarmid C. and Reed B. Eds., Springer, 1998.

[31] D. Panchenko, Symmetrization Approach to Concentration Inequalities for Empirical Processes, *The Annals Of Probability*, Vol. 31, No. 4 (2003), 2068-2081.

[32] R Development Core Team, R: A Language And Environment For Statistical Computing, *R Foundation For Statistical Computing*, Vienna, Austria, 2004. URL http://www.R-project.org.

[33] G. Ratsch, C. Schafer, B. Scholkopf and S. Sonnenburg, Large Scale Multiple Kernel Learning, *Journal of Machine Learning Research*, 7 (2006), 1531-1565.

[34] R. Schapire, Y. Freund, P. Bartlett and W. S. Lee, Boosting the margin: A new explanation for the effectiveneness of voting methods, *Annals of Statistics*, 26:1651-1686, 1998.

[35] B. Schölkopf, P. Bartlett, A. Smola and R. Williamson, Support vector regression with automatic accuracy control, in L. Niklasson, M. Bodén and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 147-152, Springer-Verlag, 1998.

[36] B. Schölkopf, A. J. Smola and K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10:1299:1319, 1998.

[37] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, *Journal of Machine Learning Research* **3** (2002), 233–269.

[38] M. Tipping, The Relevance Vector Machine, in *Advances in Neural Information Processing Systems*, San Mateo, CA, 2000. Morgan Kaufmann.

[39] A. B. Tsybakov, *Introduction à l'estimation non-paramétrique, 2004*, Mathématiques et Applications, Springer.

[40] L. G. Valiant, A Theory of the Learnable,Communications of the ACM, 27(11), November 1984.

[41] V. N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 1998.

[42] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[43] E. T. Whittaker and G. N. Watson, *A course in modern analysis*, Cambridge Univsersity Press, 1927.

[44] B. Widrow and M. Hoff, *Adaptative switching circuits*, IRE WESCON Convention Record, 4:96-104, 1960.

[45] J. Rissanen, Modeling by shortest data description, *Automatica*, vol. 14 (1978), pp. 465-471.

[46] T. Zhang, Statistical Behavior And Consistency of Classification Methods based on Convex Risk Minimization, *The Annals of Statistics*, 32:56-85, 2004.

[47] Zhao Zhang, Su Zhang, Chen-xi Zhang and Ya-zhu Chen, SVM for density estimation and application to medical image segmentation, *Journal of Zhejiang University Science B*, vol. 7(5), may 2006.