Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2012, Vol. 48, No. 3, 884–908
DOI: 10.1214/11-AIHP425

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES

www.imstat.org/aihp

Optimal model selection in density estimation

Matthieu Lerasle[1]

INSA Toulouse, Institut Mathématiques de Toulouse UMR CNRS 5219, IME-USP, São Paulo, Brasil. E-mail: lerasle@gmail.com

Received 7 October 2009; revised 24 January 2011; accepted 15 March 2011

Abstract. In order to calibrate a penalization procedure for model selection, the statistician has to choose a shape for the penalty and a leading constant. In this paper, we study, for the marginal density estimation problem, the resampling penalties as general estimators of the shape of an ideal penalty. We prove that the selected estimator satisfies sharp oracle inequalities without remainder terms under a few assumptions on the marginal density s and the collection of models. We also study the slope heuristic, which yields a data-driven choice of the leading constant in front of the penalty when the complexity of the models is well-chosen.

Résumé. Une procédure de pénalisation en sélection de modèle repose sur la construction d'une forme pour la pénalité ainsi que sur le choix d'une constante de calibration. Dans cet article, nous étudions, pour le problème d'estimation de la densité, les pénalités obtenues par rééchantillonnage de pénalités idéales. Nous montrons l'efficacité de ces procédures pour l'estimation de la forme des pénalités en prouvant, pour les estimateurs sélectionnés, des inégalités d'oracle fines sans termes résiduelles; les résultats sont valides sous des hypothèses faibles à la fois sur la densité inconnue s et sur les collections de modèles. Ces pénalités sont de plus faciles à calibrer puisque la constante asymptotiquement optimale peut être calculée en fonction des poids de rééchantillonnage. En pratique, le nombre de données est toujours fini, nous étudions donc également l'heuristique de pente et justifions l'algorithme de pente qui permet de calibrer la constante de calibration à partir des données.

MSC: 62G07; 62G09

Keywords: Density estimation; Optimal model selection; Resampling methods; Slope heuristic

1. Introduction

Model selection by penalization of an empirical loss is a general approach including famous procedures as AIC [1,2], BIC [24], Mallows C_p [19], cross-validation [23] or hard thresholding [14] as shown by [6]. The objective is to select an estimator satisfying an oracle inequality. In order to achieve this goal, the statistician should evaluate the shape of a good penalty and the constant in front of it.

In this paper, we study theoretically in a density estimation framework the slope heuristic of Birgé and Massart [10]. There is two main reasons for this. First, it provides a general shape of the penalty under a few restrictions on the density s and the collection of models, for example, these models can be of infinite dimension. Then, it gives the precise behavior of the selected model when the leading constant increases. A remarkable fact is that the complexity of the selected model is as large as possible until the leading constant reaches some particular point $K_{\min}$. This complexity decreases then strongly, the selected model becomes more reasonable and the estimator avoids to over fit the data. Another remarkable fact is that the model selected by a leading constant equal to $2K_{\min}$ is a sharp oracle. The heuristic can then be used to justify the slope algorithm, introduced in [5], which allows to evaluate in practice the leading constant in front of a penalty (see Section 2.1 for details). The calibration of this constant is usually an issue

[1] Supported in part by FAPESP Processo 2009/09494-0.

for the statistician. The upper bounds given in theoretical theorems, when computable, are in general too pessimistic and some cross-validation methods have been used to overcome this problem in simulations (see for example [16]). The slope algorithm chooses in general a constant more reasonable than the theoretical one and ensures that the chosen model is of reasonable size. The first main contribution of the paper is a proof of the slope heuristic for density estimation, for general collections of models. Theorems 3.1 and 3.2 extend the results of [10] in Gaussian regression and those of [5] in non-Gaussian heteroscedastic regression over histograms.

The penalty shape obtained in the slope heuristic is not computable in general by the statistician. This is why we also study in this paper the resampling penalties. These penalties were defined by Arlot [4] following Efron's heuristic [15], as natural estimators of the "ideal penalty." We prove that the selected estimators satisfy sharp oracle inequalities, without extra assumptions on s or the collection of models. This extends the results of Arlot [4] in non-Gaussian heteroscedastic regression among histograms. We also prove that they provide sharp estimators of the penalty shape proposed in the slope theorems. Hence, they can be used together with the slope algorithm in the general framework presented in Section 2.

Resampling penalties and the slope heuristic can be defined in a more general statistical learning framework, including the problems of classification and regression (see [4,5]). Our results are therefore contributions to the theoretical understanding of these generic methods. Up to our knowledge, they are the first obtained in density estimation.

The oracle approach is now classical in statistical learning in general and in density estimation in particular. Oracle inequalities can be derived, for example, from ℓ_1 penalization methods [12], aggregation procedures [22], blockwise Stein method [21] or using T-estimators [8]. Up to our knowledge, none of these methods yield oracle inequalities without remainder terms and with a leading constant asymptotically equal to one at the level of generality presented in this paper. For example, our results are valid for data taking value in any metric space and the models can be of infinite dimension. The results of [8] hold for infinite dimensional models but the estimators are not computable in practice.

The paper is organized as follows. Section 2 presents the notations and the main definitions. In Section 3, we state our main results, that is the slope heuristic and the oracle inequality satisfied by the estimator selected by resampling penalties. In Section 4, we compute the rates of convergence in the oracle inequalities using classical collections of models. The proofs of the main results are postponed to Section 5. In the Appendix we prove technical lemmas, in particular all the concentration inequalities required in the main proofs.

2. Notations

Let $X_1, \dots, X_n$, be i.i.d. random variables, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common law P. Let μ be a known measure on $(\mathbb{X}, \mathcal{X})$ and let $L^2(\mu)$ be the space of square integrable real valued functions defined on $\mathbb{X}$. The space $L^2(\mu)$ is endowed with the following scalar product, defined for all t, t' in $L^2(\mu)$ by

$$\langle t, t' \rangle = \int_{\mathbb{X}} t(x) t'(x) \, \mathrm{d}\mu(x)$$

and the associated L^2-norm $\|\cdot\|$, defined for all t in $L^2(\mu)$ by $\|t\| = \sqrt{\langle t, t \rangle}$. We assume that P is absolutely continuous with respect to μ and we want to estimate the density s of P with respect to μ. We assume that s belongs to $L^2(\mu)$ and we measure the risk of an estimator $\hat{s}$ of s by L^2-loss, $\|s - \hat{s}\|^2$. For all functions t in $L_1(P)$, let

$$Pt = \int_{\mathbb{X}} t \, \mathrm{d}P = \int_{\mathbb{X}} ts \, \mathrm{d}\mu = \langle t, s \rangle, \qquad P_n t = \frac{1}{n} \sum_{i=1}^{n} t(X_i), \qquad \nu_n t = (P_n - P)t.$$

We estimate s by minimization of a penalized empirical loss. Let $(S_m, m \in \mathcal{M}_n)$ be a finite collection of linear spaces of measurable, real valued functions. For all m in $\mathcal{M}_n$, let $\mathrm{pen}(m)$ be a positive number and let $\hat{s}_m$ be the least-squares estimator, defined by

$$\hat{s}_m = \arg\min_{t \in S_m} \left\{ \|t\|^2 - 2P_n t \right\}. \tag{1}$$

The final estimator is given by $\tilde{s} = \hat{s}_{\hat{m}}$, where

$$\hat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \|\hat{s}_m\|^2 - 2P_n\hat{s}_m + \mathrm{pen}(m) \right\}. \tag{2}$$

We say that $\tilde{s}$ satisfies an oracle inequality without remainder term when there exists a bounded sequence C_n such that

$$\|\tilde{s} - s\|^2 \leq C_n \inf_{m \in \mathcal{M}_n} \left\{ \|\hat{s}_m - s\|^2 \right\}.$$

We say that the oracle inequality is sharp, or optimal when, moreover, $C_n \to 1$ when $n \to \infty$. An oracle minimizes over $\mathcal{M}_n$ the quantity

$$\|\hat{s}_m - s\|^2 - \|s\|^2 = \|\hat{s}_m\|^2 - 2P\hat{s}_m = \|\hat{s}_m\|^2 - 2P_n\hat{s}_m + \mathrm{pen}_{\mathrm{id}}(m).$$

In the previous inequality, the "ideal" penalty (see [4]), $\mathrm{pen}_{\mathrm{id}}(m)$ is defined by

$$\mathrm{pen}_{\mathrm{id}}(m) = 2\nu_n\hat{s}_m.$$

The ideal penalty is the central object in this paper. We will prove that the slope algorithm can be used with a penalty shape proportional to its expectation and that resampling penalties provide sharp estimators of this expectation.

2.1. *The slope heuristic*

The "slope heuristic" has been introduced by Birgé and Massart [10] in the Gaussian regression framework and developed in a general algorithm by Arlot and Massart [5]. Let $(\Delta_m)_{m \in \mathcal{M}_n}$ be a collection of complexity measures of the models. The heuristic states that there exists a constant $K_{\min}$ satisfying the following properties.

SH1 When $\mathrm{pen}(m) < K_{\min}\Delta_m$, then $\Delta_{\hat{m}}$ is too large, typically $\Delta_{\hat{m}} \geq C \max_{m \in \mathcal{M}_n} \Delta_m$.
SH2 When $\mathrm{pen}(m) \simeq (K_{\min} + \delta)\Delta_m$ for some $\delta > 0$, then $\Delta_{\hat{m}}$ is much smaller.
SH3 When $\mathrm{pen}(m) \simeq 2K_{\min}\Delta_m$, the selected estimator is optimal.

In this paper, we prove SH1–SH3 for $\Delta_m = \mathbb{E}(\mathrm{pen}_{\mathrm{id}}(m))$, $K_{\min} = 1$. Besides the theoretical implications that we discuss later, the heuristic is of particular interest in the following situation. Imagine that there exists, for all m in $\mathcal{M}_n$, a quantity Δ_m computable by the statistician and an unknown constant K_{u} such that, for some $\delta \ll K_{\mathrm{u}}$,

$$(K_{\mathrm{u}} - \delta)\Delta_m \leq \mathbb{E}\big(\mathrm{pen}_{\mathrm{id}}(m)\big) \leq (K_{\mathrm{u}} + \delta)\Delta_m.$$

In that case, it comes from SH3 that a penalty of the form $K\Delta_m$ yields a good procedure if K is large enough. In order to choose K in a data-driven way, we can use the following algorithm (see Arlot and Massart [5]).

Slope algorithm

SA1 For all $K > 0$, compute the selected model $\hat{m}(K)$ given by (2) with the penalty $\mathrm{pen}(m) = K\Delta_m$ and the associated complexity $\Delta_{\hat{m}(K)}$.
SA2 Find a constant $\tilde{K}_{\min}$ such that $\Delta_{\hat{m}(K)}$ is large when $K < \tilde{K}_{\min}$, and "much smaller" when $K > \tilde{K}_{\min}$.
SA3 Take the final $\hat{m} = \hat{m}(2\tilde{K}_{\min})$.

The idea is that $\tilde{K}_{\min} \simeq K_{\min}$ because we observe a jump in the complexity of the selected model at this point (SH1 and SH2). Hence, the model selected by $2\tilde{K}_{\min}\Delta_m \simeq 2K_{\min}\Delta_m$ should be optimal thanks to SH3. By construction, the slope algorithm ensures that the selected model has a reasonable size. It prevents the statistician from choosing a too large model which could have terrible consequences (see Theorem 3.1). The words "much smaller," borrowed from [5,10], are not very clear here. We refer to [5], Section 3.3 for a detailed discussion on what "much smaller" means in this context and for precise suggestions on the implementation of the slope algorithm. We refer also to [7] for a practical implementation of the slope algorithm.

2.2. *Resampling penalties*

Data-driven penalties have already been used in density estimation in particular cross-validation methods as in Stone [25], Rudemo [23] or Celisse [13]. We are interested here in the resampling penalties introduced by Arlot [4]. Let $(W_1, \dots, W_n)$ be a resampling scheme, i.e. a vector of random variables independent of $X, X_1, \dots, X_n$ and exchangeable, that is, for all permutations τ of $(1, \dots, n)$,

$$(W_1, \dots, W_n) \text{ has the same law as } (W_{\tau(1)}, \dots, W_{\tau(n)}).$$

Hereafter, we denote by $\bar{W}_n = \sum_{i=1}^n W_i/n$ and by E^W and $\mathcal{L}^W$ respectively the expectation and the law conditionally on the data $X, X_1, \dots, X_n$. Let $P_n^W = \sum_{i=1}^n W_i \delta_{X_i}/n$, $\nu_n^W = P_n^W - \bar{W}_n P_n$ be the resampled empirical processes. Arlot's procedure is based on the resampling heuristic of Efron (see Efron [15]), which states that the law of a functional $F(P, P_n)$ is close to the conditional law $\mathcal{L}^W(C_W F(\bar{W}_n P_n, P_n^W))$. C_W is a renormalizing constant that depends only on the resampling scheme and on F. Following this heuristic, Arlot defines the resampling penalties as resampling estimates of $\mathbb{E}(\text{pen}_{\text{id}}(m))$, that is

$$\text{pen}(m) = 2C_W \mathbb{E}^W\big(\nu_n^W\big(\hat{s}_m^W\big)\big), \quad \text{where } \hat{s}_m^W = \arg\min_{t \in S_m}\big\{\|t\|^2 - 2P_n^W t\big\}. \tag{3}$$

We prove concentration inequalities for $\text{pen}(m)$ and we compute the value of C_W such that $\text{pen}(m)$ provides a sharp estimator of $\text{pen}_{\text{id}}(m)$, we deduce that $\text{pen}(m)$ provides an optimal model selection procedure (see Theorem 3.3).

2.3. *Main assumptions*

For all m, m' in $\mathcal{M}_n$, let $D_m = n\mathbb{E}(\|s_m - \hat{s}_m\|^2)$,

$$\frac{R_m}{n} = \mathbb{E}\big(\|s - \hat{s}_m\|^2\big) = \|s - s_m\|^2 + \frac{D_m}{n},$$

$$v_{m,m'}^2 = \sup_{t \in S_m + S_{m'}, \|t\| \le 1} \text{Var}\big(t(X)\big),$$

$$e_{m,m'} = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\| \le 1} \|t\|_\infty^2.$$

For all $k \in \mathbb{N}$, let $\mathcal{M}_n^k = \{m \in \mathcal{M}_n, R_m \in [k, k+1)\}$. For all n in $\mathbb{N}$, for all $k > 0$, $k' > 0$ and $\gamma \ge 0$, let $[k]$ be the integer part of k and let

$$l_{n,\gamma}\big(k, k'\big) = \ln\big(1 + \text{Card}\big(\mathcal{M}_n^{[k]}\big)\big) + \ln\big(1 + \text{Card}\big(\mathcal{M}_n^{[k']}\big)\big) + \ln\big((k+1)\big(k'+1\big)\big) + (\ln n)^\gamma. \tag{4}$$

[V] There exist $\gamma > 1$ and a sequence $(\varepsilon_n)_{n \in \mathbb{N}}$, with $\varepsilon_n \to 0$ such that, for all n in $\mathbb{N}$,

$$\sup_{(k,k') \in (\mathbb{N}^*)^2} \sup_{(m,m') \in \mathcal{M}_n^k \times \mathcal{M}_n^{k'}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2\big(k, k'\big) \right\} \le \varepsilon_n^4.$$

Comments.

- [V] *ensures that the fluctuations of the ideal penalty are uniformly small compared to the risk of the estimator* $\hat{s}_m$. *Note that for all* k, k', $l_{n,\gamma}(k, k') \ge (\ln n)^\gamma$, *thus,* [V] *holds only in non-parametric situations where* $R_n = \inf_{m \in \mathcal{M}_n} R_m \to \infty$ *as* $n \to \infty$.

[BR] *There exist two sequences* $(h_n^*)_{n \in \mathbb{N}^*}$ *and* $(h_n^o)_{n \in \mathbb{N}^*}$ *with* $(h_n^o \vee h_n^*) \to 0$ *as* $n \to \infty$ *such that, for all* n *in* $\mathbb{N}^*$, *for all* $m_o \in \arg\min_{m \in \mathcal{M}_n} R_m$ *and all* $m^* \in \arg\max_{m \in \mathcal{M}_n} D_m$,

$$\frac{R_{m_o}}{D_{m^*}} \le h_n^o, \qquad \frac{n\|s - s_{m^*}\|^2}{D_{m^*}} \le h_n^*.$$

Comments.

- *The slope heuristic states that the complexity $\Delta_{\hat{m}}$ of the selected estimator is too large when the penalty term is too small. A minimal assumption for this heuristic to hold with $\Delta_m = D_m$ is that there exists a sequence $(\theta_n)_{n\in\mathbb{N}^*}$ with $\theta_n \to 0$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o \in \arg\min_{m\in\mathcal{M}_n} \mathbb{E}(\|s - \hat{s}_m\|^2)$ and all $m^* \in \arg\max_{m\in\mathcal{M}_n} \mathbb{E}(\|s_m - \hat{s}_m\|^2)$,*

$$D_{m_o} \leq \theta_n D_{m^*}.$$

Assumption [BR] *is slightly stronger but will always hold in the examples (see Section 4).*

In order to have an idea of the rates R_n, ε_n, h_n^*, h_n^0, let us briefly consider the very simple following example:

***Example* HR.** *We assume that s is supported in $[0,1]$ and that $(S_m)_{m\in\mathcal{M}_n}$ is the collection of the regular histograms on $[0,1]$, with $d_m = 1,\dots,n$ pieces. We will see in Section 4.2 that $D_m \sim d_m$ asymptotically, hence $D_{m^*} \simeq n$. Moreover, if s is Hölderian and not constant, there exist positive constants $c_{\mathrm{l}}, c_{\mathrm{u}}, \alpha_{\mathrm{l}}, \alpha_{\mathrm{u}}$ such that, for all m in $\mathcal{M}_n$, see [4],*

$$c_{\mathrm{l}} d_m^{-\alpha_{\mathrm{l}}} \leq \|s - s_m\|^2 \leq c_{\mathrm{u}} d_m^{-\alpha_{\mathrm{u}}}.$$

In Section 4.2, we prove that this assumption implies [V] *with $\varepsilon_n \leq C \ln(n) n^{-1/(8\alpha_{\mathrm{l}}+4)}$.*

Moreover, there exists a constant $C > 0$ such that $R_{m_o} \leq \inf_{m\in\mathcal{M}_n}(c_{\mathrm{u}} n d_m^{-\alpha_{\mathrm{u}}} + d_m) \leq C n^{1/(2\alpha_{\mathrm{u}}+1)}$, thus $R_{m_o}/D_m^* \leq C n^{-2\alpha_{\mathrm{u}}/(2\alpha_{\mathrm{u}}+1)}$. Since there exists $C > 0$ such that $n\|s - s_{m^*}\|^2/D_{m^*} \leq C d_{m^*}^{-\alpha_{\mathrm{u}}} = C n^{-\alpha_{\mathrm{u}}}$, [BR] *holds with $h_n^o = C n^{-2\alpha_{\mathrm{u}}/(2\alpha_{\mathrm{u}}+1)}$ and $h_n^* = C n^{-\alpha_{\mathrm{u}}}$.*

Other examples can be found in Birgé and Massart [9], see also Section 4.

3. Main results

3.1. *The slope heuristic*

The first result deals with the behavior of the selected estimator when $\mathrm{pen}(m)$ is too small.

Theorem 3.1 (Minimal penalty). *Let $\mathcal{M}_n$ be a collection of models satisfying* [V] *and* [BR] *and let $\varepsilon_n^* = \varepsilon_n \vee h_n^*$. Assume that there exists $0 < \delta_n < 1$ such that $0 \leq \mathrm{pen}(m) \leq (1-\delta_n) D_m/n$. Let $\hat{m}, \tilde{s}$ be the random variables defined in (2) and let*

$$c_n = \frac{\delta_n - 28\varepsilon_n^*}{1 + 16\varepsilon_n}.$$

There exists a constant $C > 0$ such that, with probability larger than $1 - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$,

$$D_{\hat{m}} \geq c_n D_{m^*}, \qquad \|s - \tilde{s}\|^2 \geq \frac{c_n}{5h_n^o} \inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2. \tag{5}$$

Comments.

- *Assume that $\mathrm{pen}(m) \leq (1-\delta) D_m/n$, then, for n sufficiently large, $D_{\hat{m}} \geq c D_{m^*}$ is as large as possible. This proves* SH1 *with $\Delta_m = D_m/n$, $K_{\min} = 1$.*
- *The second part of (5) proves also that, in that case, we cannot obtain an oracle inequality.*

Theorem 3.2. *Let $\mathcal{M}_n$ be a collection of models satisfying* [V]. *Assume that there exist $\delta^+ \geq \delta_- > -1$ and $0 \leq p' < 1$ such that, with probability at least $1 - p'$,*

$$2\frac{D_m}{n} + \delta_- \frac{R_m}{n} \leq \mathrm{pen}(m) \leq 2\frac{D_m}{n} + \delta^+ \frac{R_m}{n}. \tag{6}$$

Let $\hat{m}, \tilde{s}$ be the random variables defined in (2) and let

$$C_n\big(\delta_-, \delta^+\big) = \left(\frac{1+\delta_- - 46\varepsilon_n}{1+\delta^+ + 26\varepsilon_n} \vee 0\right)^{-1}.$$

There exists a constant $C > 0$ such that, with probability larger than $1 - p' - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$,

$$D_{\hat{m}} \le C_n\big(\delta_-, \delta^+\big) R_{m_o}, \qquad \|s - \tilde{s}\|^2 \le C_n\big(\delta_-, \delta^+\big) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2. \tag{7}$$

Comments.

- *Assume first that* $\mathrm{pen}(m) = (2+\delta) D_m/n$ *with δ close to -1, then, for n sufficiently large, (7) ensures that* $D_{\hat{m}} = \mathrm{O}(R_{m_o})$. *Hence,* $D_{\hat{m}}$ *jumps from* D_{m^*} *(Theorem 3.1) to* R_{m_o}. *This proves* SH2, *with* $\Delta_m = D_m/n$ *and* $K_{\min} = 1$.
- *Assume then that* $\delta_- = \delta^+ = 0$, *so that the penalty term* $\mathrm{pen}(m) = 2K_{\min}\Delta_m$. *The second part of (7) ensures that $\tilde{s}$ satisfies a sharp oracle inequality. This proves* SH3. *In general, the rate of convergence of the leading constant to 1 is given by the supremum between δ_-, δ^+ and ε_n.*
- *The second part of (7) proves also that, if* $\mathrm{pen}(m) = K\Delta_m$, *with* $K > 2K_{\min}$, *then, we only loose a constant in the oracle inequality. This is why it is better to choose a too large penalty which only yields a loss in the constant in the oracle inequality, than a too small one that may lead to an explosion of the risk.*
- D_m/n *is unknown in general and cannot be used in the slope algorithm. We propose two alternatives to solve this issue. In Section 3.2, we give a resampling estimator of D_m, it can be used for every collection of models satisfying* [V]. *This estimator satisfies (6) with* $-\delta_- = \delta^+ = \mathrm{O}(\varepsilon_n)$. *In Section 4.2, we will also see that, in regular models, we can use d_m instead of D_m and the error is upper bounded by $C R_m/R_{m_o}$, thus Theorem 3.2 holds with* $(\delta_- \vee \delta^+) \le C/R_{m_o} \ll \varepsilon_n$, $p' = 0$. *In both cases, we deduce from Theorem 3.2 that the estimator $\tilde{s}$ given by the slope algorithm achieves an optimal oracle inequality. In Example* HR, *for example, we obtain* $\varepsilon_n = Cn^{-1/(8\alpha_1+4)} \ln n$.
- *Let us notice here that the constant* $K_{\min} = 1$ *is absolute when we choose* $\Delta_m = D_m$. *This is not true in general and* $K_{\min}$ *can even depend on s. In order to see that, let us consider the collection of regular histograms on* $[0,1]$. *In that case, we have (see (9))* $D_m = (F(1) - F(0))d_m - \|s_m\|^2$, *where* $F(x) = \int_{-\infty}^{x} s(t)\,\mathrm{d}\mu(t)$. *Hence, the slope heuristic holds for* $\Delta_m = d_m$ *but the associated constant* $K_{\min} = (F(1) - F(0))$ *is not absolute if s is not supported on* $[0,1]$.

3.2. *Resampling penalties*

Theorem 3.3. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with common density s. Let $\mathcal{M}_n$ be a collection of models satisfying* [V]. *Let $W_1, \ldots, W_n$ be a resampling scheme, let* $\bar{W}_n = \sum_{i=1}^n W_i/n$, $v_W^2 = \mathrm{Var}(W_1 - \bar{W}_n)$ *and* $C_W = 2(v_W^2)^{-1}$. *Let $\tilde{s}$ be the penalized least-squares estimator defined in (2) with* $\mathrm{pen}(m)$ *defined in (3). Then, there exists a constant $C > 0$ such that*

$$\mathbb{P}\Big(\|s - \tilde{s}\|^2 \le (1 + 100\varepsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\Big) \ge 1 - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}. \tag{8}$$

Comments. *The main advantage of this result is that the penalty term is always totally computable. It does not depend on an arbitrary choice of a constant* $K_{\min}$ *made by the observer, that may be hard to detect in practice (see the paper of Alot and Massart* [5] *for an extensive discussion on this important issue). However, C_W is only optimal asymptotically. It is sometimes useful to overpenalize a little in order to improve the non-asymptotic performances of our procedures (see Massart* [20] *and the remarks after Theorem 3.2) and the slope heuristic can be used to do it in an optimal way.*

4. Rates of convergence for classical examples

The aim of this section is to show that [V] can be derived from more classical hypotheses in two classical collections of models: the histograms and Fourier spaces. We obtain the rates ε_n under these new hypotheses.

4.1. *Assumption on the risk of the oracle*

Recall that $R_n = \inf_{m\in\mathcal{M}_n} R_m$. In this section, we make the following assumption.

[BR′] (Bounds on the Risk) There exist constants $C_{\mathrm{u}} > 0$, $\alpha_{\mathrm{u}} > 0$, $\gamma > 1$, and a sequence $(\theta_n)_{n\in\mathbb{N}}$ with $\theta_n \to \infty$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all m in $\mathcal{M}_n$

$$\theta_n^2 (\ln n)^{2\gamma} \le R_n \le R_m \le C_{\mathrm{u}} n^{\alpha_{\mathrm{u}}}.$$

Comments. *Let $(S_m, m \in \mathcal{M}_n)$ be the collection of regular histograms of Example* HR. *Assume that s is an Hölderian, non-constant and compactly supported function, then there exist positive constants c_i, c_{u}, α_i, α_{u} such that (see for example Arlot* [3])

$$c_i d_m^{-\alpha_i} \le \|s - s_m\| \le c_{\mathrm{u}} d_m^{-\alpha_{\mathrm{u}}}.$$

We have also, see (9), $D_m = d_m - \|s_m\|^2$. *Hence,*

$$R_n \ge c_\star \inf_{d_m=1,\dots,n} \left(n d_m^{-2\alpha_i} + d_m\right) \ge c_\star n^{1/(2\alpha_i+1)}.$$

For all m in $\mathcal{M}_n$, we also have $R_m \le (c_{\mathrm{u}} + 1)n$. Hence, [BR′] *holds for all $\gamma > 1$ with $\theta_n = c_\star n^{1/(4\alpha_i+2)} (\ln n)^{-2\gamma}$. It is also a classical result of minimax theory that there exist functions in Sobolev spaces satisfying this kind of assumption when $\mathcal{M}_n$ is the collection of Fourier spaces that we will introduce below.*

We also make the following assumption on the collection $(S_m, m \in \mathcal{M}_n)$.

[PC] (Polynomial collection) There exist constants $c_{\mathcal{M}} \ge 0$, $\alpha_{\mathcal{M}} \ge 0$, such that, for all n in $\mathbb{N}$,

$$\mathrm{Card}(\mathcal{M}_n) \le c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}.$$

Under Assumptions [BR′] and [PC], for all $m \in \mathcal{M}_n$, $R_m \le C_{\mathrm{u}} n^{\alpha_{\mathrm{u}}}$, thus for all $k > C_{\mathrm{u}} n^{\alpha_{\mathrm{u}}}$, $\mathrm{Card}(\mathcal{M}_n^{[k]}) = 0$. In particular, we only have to take into account in [V] the integers k and k' such that $k \le C_{\mathrm{u}} n^{\alpha_{\mathrm{u}}}$ and $k' \le C_{\mathrm{u}} n^{\alpha_{\mathrm{u}}}$ and there exists a constant $\kappa > 0$ such that $\ln[(1+k)(1+k')] \le \kappa \ln n$. Moreover, under [PC], $\ln(1 + \mathrm{Card}(\mathcal{M}_n^{[k]})) \le \kappa \ln n$, hence, there exists a constant $\kappa > 0$ such that, for all $\gamma > 1$ and $n \ge 3$,

$$\sup_{(k,k')\in(\mathbb{N}^*)^2} \sup_{(m,m')\in\mathcal{M}_n^{[k]}\times\mathcal{M}_n^{[k']}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2(k,k') \right\}$$
$$\le \sup_{(m,m')\in(\mathcal{M}_n)^2} \left\{ \left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right\} \kappa (\ln n)^{2\gamma}.$$

4.2. *The histogram case*

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let $(P_m)_{m\in\mathcal{M}_n}$ be a growing collection of measurable partitions $P_m = (I_\lambda)_{\lambda\in m}$ of subsets of $\mathbb{X}$ such that, for all $m \in \mathcal{M}_n$, for all $\lambda \in m$, $0 < \mu(I_\lambda) < \infty$. Let m in $\mathcal{M}_n$, the set S_m of histograms associated to P_m is the set of functions which are constant on each I_λ, $\lambda \in m$. S_m is a linear space. Setting, for all $\lambda \in m$, $\psi_\lambda = (\sqrt{\mu(I_\lambda)})^{-1} 1_{I_\lambda}$, the functions $(\psi_\lambda)_{\lambda\in m}$ form an orthonormal basis of S_m.

Let us recall that, for all m in $\mathcal{M}_n$,

$$D_m = \sum_{\lambda\in m} \mathrm{Var}\left(\psi_\lambda(X)\right) = \sum_{\lambda\in m} P\left(\psi_\lambda^2\right) - (P\psi_\lambda)^2 = \sum_{\lambda\in m} \frac{P(X \in I_\lambda)}{\mu(I_\lambda)} - \|s_m\|^2. \tag{9}$$

Moreover, from Cauchy–Schwarz inequality, for all x in $\mathbb{X}$, for all m, m' in $\mathcal{M}_n$

$$\sup_{t\in B_{m,m'}} t^2(x) \le \sum_{\lambda\in m\cup m'} \psi_\lambda^2(x), \quad \text{thus } e_{m,m'} = \frac{1}{n} \sup_{\lambda\in m\cup m'} \frac{1}{\mu(I_\lambda)}. \tag{10}$$

Finally, it is easy to check that, for all m ,m' in $\mathcal{M}_n$

$$v_{m,m'}^2 = \sup_{\lambda\in m\cup m'} \mathrm{Var}\big(\psi_\lambda(X)\big) = \sup_{\lambda\in m\cup m'} \frac{P(X\in I_\lambda)(1-P(X\in I_\lambda))}{\mu(I_\lambda)}. \tag{11}$$

We will consider two particular types of histograms.

Example 1* ([Reg]: *μ-regular histograms). *For all m in $\mathcal{M}_n$, P_m is a partition of $\mathbb{X}$ and there exist a family $(d_m)_{m\in\mathcal{M}_n}$ bounded by n and two constants c_{rh}, C_{rh} such that, for all m in $\mathcal{M}_n$, for all $\lambda\in\mathcal{M}_n$,*

$$\frac{c_{\mathrm{rh}}}{d_m} \le \mu(I_\lambda) \le \frac{C_{\mathrm{rh}}}{d_m}.$$

The typical example here is the collection described in Example HR, where $c_r = C_r = 1$. Remark that a collection satisfying [Reg] can be of infinite dimension.

Example 2* ([Ada]: *Adapted histograms). *There exist positive constants c_r, C_{ah} such that, for all m in $\mathcal{M}_n$, for all $\lambda\in m$, $\mu(I_\lambda)\ge c_r n^{-1}$ and*

$$\frac{P(X\in I_\lambda)}{\mu(I_\lambda)} \le C_{\mathrm{ah}}.$$

[Ada] is typically satisfied when s is bounded on $\mathbb{X}$. Remark that the models satisfying [Ada] have finite dimension $d_m \le Cn$ since

$$1 \ge \sum_{\lambda\in m} P(X\in I_\lambda) \ge C_{\mathrm{ah}} \sum_{\lambda\in m} \mu(I_\lambda) \ge C_{\mathrm{ah}} c_r d_m n^{-1}.$$

The example [Reg]
It comes from Eqs (9)–(11) and Assumption [Reg] that

$$C_{\mathrm{rh}}^{-1} d_m - \|s_m\|^2 \le D_m \le c_{\mathrm{rh}}^{-1} d_m - \|s_m\|^2,$$
$$e_{m,m'} \le c_{\mathrm{rh}}^{-1} \frac{d_m \vee d_{m'}}{n}, \qquad v_{m,m'}^2 \le \sup_{t\in B_{m,m'}} \|t\|_\infty \|t\| \|s\| \le c_{\mathrm{rh}}^{-1/2} \|s\| \sqrt{d_m \vee d_{m'}}.$$

Thus

$$\frac{e_{m,m'}}{R_m \vee R_{m'}} \le C_{\mathrm{rh}} c_{\mathrm{rh}}^{-1} \frac{(R_m \vee R_{m'}) + \|s\|^2}{n(R_m \vee R_{m'})} \le Cn^{-1}.$$

If $D_m \vee D_{m'} \le \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \le \sqrt{C_{\mathrm{rh}} c_{\mathrm{rh}}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{R_{m_o}} \le \frac{C}{\theta_n (\ln n)^\gamma}.$$

If $D_m \vee D_{m'} \ge \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \le \sqrt{C_{\mathrm{rh}} c_{\mathrm{rh}}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{D_m \vee D_{m'}} \le \frac{C}{\theta_n (\ln n)^\gamma}.$$

There exists $\kappa > 0$ such that $\theta_n^2 (\ln n)^{2\gamma} \le \kappa n$ since for all m in $\mathcal{M}_n$, $R_m \le n\|s - s_m\|^2 + c_{\mathrm{rh}}^{-1} d_m \le (\|s\|^2 + c_{\mathrm{rh}}^{-1}) n$. Hence Assumption [V] holds with γ given in Assumption [BR′] and $\varepsilon_n = C\theta_n^{-1/2}$.

The example [Ada]
It comes from inequalities (10), (11) and Assumption [Ada] that, for all m and m' in $\mathcal{M}_n$

$$e_{m,m'} \le c_r^{-1} \quad \text{and} \quad v_{m,m'}^2 \le C_{\text{ah}}.$$

Thus, there exists a constant $\kappa > 0$ such that, for all m an m' in $\mathcal{M}_n$,

$$\sup_{(m,m')\in(\mathcal{M}_n)^2} \left\{ \left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right\} \le \frac{\kappa}{\theta_n^2 (\ln n)^{2\gamma}}.$$

Therefore Assumption [V] holds also with γ given in Assumption [BR′] and $\varepsilon_n = \kappa \theta_n^{-1/2}$.

4.3. *Fourier spaces*

In this section, we assume that s is supported in $[0, 1]$. We introduce the classical Fourier basis. Let $\psi_0 : [0, 1] \to \mathbb{R}$, $x \mapsto 1$ and, for all $k \in \mathbb{N}^*$, we define the functions

$$\psi_{1,k} : [0, 1] \to \mathbb{R}, \qquad x \mapsto \sqrt{2}\cos(2\pi k x), \qquad \psi_{2,k} : [0, 1] \to \mathbb{R}, \qquad x \mapsto \sqrt{2}\sin(2\pi k x).$$

For all j in $\mathbb{N}^*$, let

$$m_j = \{0\} \cup \big\{(i, k), i = 1, 2, k = 1, \ldots, j\big\} \quad \text{and} \quad \mathcal{M}_n = \{m_j, j = 1, \ldots, n\}.$$

For all m in $\mathcal{M}_n$, let S_m be the space spanned by the family $(\psi_\lambda)_{\lambda\in m}$. $(\psi_\lambda)_{\lambda\in m}$ is an orthonormal basis of S_m and for all j in $1, \ldots, n$, $d_{m_j} = 2j + 1$.

Let j in $1, \ldots, n$, for all x in $[0, 1]$,

$$\sum_{\lambda\in m_j} \psi_\lambda^2(x) = 1 + 2\sum_{k=1}^{j} \cos^2(2\pi k x) + \sin^2(2\pi k x) = 1 + 2j = d_{m_j}.$$

Hence, for all m in $\mathcal{M}_n$,

$$D_m = P\left(\sum_{\lambda\in m_j} \psi_\lambda^2 \right) - \|s_m\|^2 = d_m - \|s_m\|^2. \tag{12}$$

It is also clear that, for all m, m' in $\mathcal{M}_n$,

$$e_{m,m'} = \frac{d_m \vee d_{m'}}{n}, \qquad v_{m,m'}^2 \le \|s\| \sqrt{d_m \vee d_{m'}}. \tag{13}$$

The collection of Fourier spaces of dimension $d_m \le n$ satisfies Assumption [PC], and the quantities D_m $e_{m,m'}$ and $v_{m,m'}^2$ satisfy the same inequalities as in the collection [Reg], therefore, [V] comes also in this collection from [BR′]. We have obtained the following corollary of Theorem 3.3.

Corollary 4.1. *Let $\mathcal{M}_n$ be either a collection of histograms satisfying Assumptions* [PC]–[Reg] *or* [PC]–[Ada] *or the collection of Fourier spaces of dimension $d_m \le n$. Assume that s satisfies Assumption* [BR′] *for some $\gamma > 1$ and $\theta_n \to \infty$. Then, there exist constants $\kappa > 0$ and $C > 0$ such that the estimator $\tilde{s}$ selected by a resampling penalty satisfies*

$$\mathbb{P}\Big(\|s - \tilde{s}\|^2 \le \big(1 + \kappa\theta_n^{-1/2}\big) \inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2 \Big) \ge 1 - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}.$$

Comment. *Assumption* [BR′] *is hard to check in practice. We described in the beginning of this section some examples where it holds. In more general situations, following Arlot* [4], *we use our main theorem only for the models with dimension* $d_m \geq (\ln n)^{4+2\gamma}$, *they satisfy* [BR′] *with* $\theta_n = (\ln n)^2$, *at least when* n *is sufficiently large, because*

$$\|s\|^2 + R_m \geq \|s\|^2 + D_m \geq cd_m \geq c(\ln n)^4 (\ln n)^{2\gamma}.$$

With our concentration inequalities, we can control easily the risk of the models with dimension $d_m \leq (\ln n)^{4+2\gamma}$ *by* $\kappa(\ln n)^{3+5\gamma/2}$ *with probability larger than* $1 - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$.

We can then deduce the following corollary.

Corollary 4.2. *Let* $\mathcal{M}_n$ *be either a collection of histograms satisfying Assumptions* [PC]–[Reg] *or* [PC]–[Ada] *or the collection of Fourier spaces of dimension* $d_m \leq n$. *There exist constants* $\kappa > 0$, $\eta > 3 + 5\gamma/2$ *and* $C > 0$ *such that the estimator* $\tilde{s}$ *selected by a resampling penalty satisfies*

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \leq \bigl(1 + \kappa(\ln n)^{-1}\bigr) \left(\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 + \frac{(\ln n)^\eta}{n} \right) \right) \geq 1 - C\mathrm{e}^{-(1/2)(\ln n)^\gamma}.$$

5. Proofs of the main results

5.1. *Notations*

Let us recall here the main notations. For all m, m' in $\mathcal{M}_n$, let

$$\begin{aligned} &p(m) = \|s_m - \hat{s}_m\|^2, \qquad D_m = n\mathbb{E}\bigl(p(m)\bigr) = n\mathbb{E}\bigl(\|\hat{s}_m - s_m\|^2\bigr), \\ &R_m = n\mathbb{E}\bigl(\|s - \hat{s}_m\|^2\bigr) = n\|s - s_m\|^2 + D_m, \qquad \delta\bigl(m, m'\bigr) = \nu_n(s_m - s_{m'}). \end{aligned}$$

For all $n \in \mathbb{N}^*$, $k > 0$, $k' > 0$, $\gamma > 0$, let $[k]$ be the integer part of k and let

$$l_{n,\gamma}\bigl(k, k'\bigr) = \ln\bigl(\bigl(1 + \mathrm{Card}\bigl(\mathcal{M}_n^{[k]}\bigr)\bigr)\bigl(1 + \mathrm{Card}\bigl(\mathcal{M}_n^{[k']}\bigr)\bigr)\bigr) + \ln\bigl((1+k)\bigl(1+k'\bigr)\bigr) + (\ln n)^\gamma.$$

Recall that Assumption [V] implies that, for all m, m' in $\mathcal{M}_n$,

$$\begin{aligned} &v_{m,m'}^2 l_{n,\gamma}(R_m, R_{m'}) \leq \varepsilon_n^2 (R_m \vee R_{m'}), \\ &e_{m,m'}\bigl(l_{n,\gamma}(R_m, R_{m'})\bigr)^2 \leq \varepsilon_n^4 (R_m \vee R_{m'}). \end{aligned} \tag{14}$$

Let $(\psi_\lambda)_{\lambda \in m}$ be an orthonormal basis of S_m. Easy algebra leads to

$$s_m = \sum_{\lambda \in m} (P\psi_\lambda)\psi_\lambda, \qquad \hat{s}_m = \sum_{\lambda \in m} (P_n\psi_\lambda)\psi_\lambda, \quad \text{thus } \|s_m - \hat{s}_m\|^2 = \sum_{\lambda \in m} \bigl(\nu_n(\psi_\lambda)\bigr)^2.$$

Therefore, $\hat{s}_m$ is an unbiased estimator of s_m and

$$\mathrm{pen}_{\mathrm{id}}(m) = 2\nu_n(\hat{s}_m) = 2\nu_n(\hat{s}_m - s_m) + 2\nu_n(s_m) = 2\|s_m - \hat{s}_m\|^2 + 2\nu_n(s_m).$$

By definition, $\hat{m}$ minimizes $\|s - \hat{s}_m\|_2^2 + \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$. Hence, for all m in $\mathcal{M}_n$,

$$\|s - \tilde{s}\|_2^2 \leq \|s - \hat{s}_m\|_2^2 + \bigl(\mathrm{pen}(m) - 2p(m)\bigr) + \bigl(2p(\hat{m}) - \mathrm{pen}(\hat{m})\bigr) + \delta(\hat{m}, m). \tag{15}$$

5.2. *Proof of Theorem 3.1*

If $c_n < 0$, there is nothing to prove. We can then assume that $c_n \geq 0$, this implies in particular that

$$28\varepsilon_n \leq \delta_n < 1.$$

We use the notations of Lemma A.10. From Lemma A.10, inequalities (5) will be proved if, on Ω_T, $D_{\hat{m}} \geq c_n D_{m^*}$ and

$$\|s - \tilde{s}\|^2 \geq \frac{c_n}{5h_n^o} \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

Let $m_o \in \arg\min_{m \in \mathcal{M}_n} R_m$, $\hat{m}$ minimizes over $\mathcal{M}_n$ the following criterion.

$$\begin{aligned}\mathrm{Crit}(m) &= \|\hat{s}_m\|^2 - 2P_n\hat{s}_m + \mathrm{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m).\end{aligned}$$

Recall that $0 \leq \mathrm{pen}(m) \leq (1 - \delta_n)D_m/n$. On Ω_T, for all m in $\mathcal{M}_n$, since $R_m \geq R_{m_o}$,

$$\begin{aligned}&\mathrm{Crit}(m) \geq \|s - s_m\|^2 - \frac{D_m}{n} - 16\varepsilon_n \frac{R_m}{n} \geq -(1 + 16\varepsilon_n)\frac{D_m}{n}, \\ &\mathrm{Crit}(m) \leq \|s - s_m\|^2 + 26\varepsilon_n \frac{R_m}{n} - \delta_n \frac{D_m}{n} = (1 + 26\varepsilon_n)\|s - s_m\|^2 - (\delta_n - 26\varepsilon_n)\frac{D_m}{n}.\end{aligned}$$

When $D_m \leq c_n D_{m^*}$,

$$(1 + 16\varepsilon_n)D_m \leq D_{m^*}\left((\delta_n - 26\varepsilon_n) - (1 + 26\varepsilon_n)\frac{n\|s - s_{m^*}\|^2}{D_{m^*}}\right).$$

Thus $\mathrm{Crit}(m) \geq \mathrm{Crit}(m^*)$. This implies that $D_{\hat{m}} \geq c_n D_{m^*}$.

Moreover, on Ω_T, we also have, for all m in $\mathcal{M}_n$

$$\|s - \tilde{s}\|^2 = \frac{R_{\hat{m}}}{n} + \left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) \geq (1 - 20\varepsilon_n)\frac{R_{\hat{m}}}{n},$$

and

$$\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 \leq \inf_{m \in \mathcal{M}_n} \frac{R_m}{n}(1 + 10\varepsilon_n) \leq \frac{R_{m_o}}{n}(1 + 10\varepsilon_n).$$

Thus

$$\begin{aligned}\|s - \tilde{s}\|^2 &\geq (1 - 20\varepsilon_n)\frac{R_{\hat{m}}}{n} \geq (1 - 20\varepsilon_n)\frac{D_{\hat{m}}}{n} \geq (1 - 20\varepsilon_n)c_n\frac{D_{m^*}}{n} \\ &\geq c_n \frac{1 - 20\varepsilon_n}{h_n^o}\frac{R_{m_o}}{n} \geq \frac{c_n}{h_n^o}\frac{1 - 20\varepsilon_n}{1 + 10\varepsilon_n} \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.\end{aligned}$$

We conclude the proof, saying that $\varepsilon_n \leq 1/28$ implies that $(1 - 20\varepsilon_n)(1 + 10\varepsilon_n)^{-1} \geq 8/38 \geq 1/5$.

5.3. *Proof of Theorem 3.2*

If $\delta_- - 46\varepsilon_n < -1$, there is nothing to prove, hence, we can assume in the following that $\delta_- - 46\varepsilon_n > -1$. We keep the notation Ω_T introduced in Lemma A.10. Let

$$\Omega_{\mathrm{pen}} = \bigcap_{m \in \mathcal{M}_n} \left\{\frac{2D_m}{n} + \delta_- \frac{R_m}{n} \leq \mathrm{pen}(m) \leq \frac{2D_m}{n} + \delta^+ \frac{R_m}{n}\right\},$$

$\Omega = \Omega_T \cap \Omega_{\mathrm{pen}}$ and $m_o \in \arg\min_{m\in\mathcal{M}_n} R_m$. Recall that $\mathbb{P}(\Omega_{\mathrm{pen}}) \geq 1 - p'$ and that, $\hat{m}$ minimizes over m the following criterion.

$$\begin{aligned}\mathrm{Crit}(m) &= \|\hat{s}_m\|^2 - 2P_n(\hat{s}_m) + \mathrm{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m).\end{aligned}$$

Therefore, on Ω, for all m in $\mathcal{M}_n$, since $R_m \geq R_{m_o}$,

$$\begin{aligned}\mathrm{Crit}(m) &\geq (1+\delta_-)\frac{R_m}{n} + \left(\frac{D_m}{n} - p(m)\right) - 6\varepsilon_n\frac{R_m}{n} \\ &\geq (1+\delta_- - 16\varepsilon_n)\|s - s_m\|^2 + (1+\delta_- - 16\varepsilon_n)\frac{D_m}{n} \geq (1+\delta_- - 16\varepsilon_n)\frac{D_m}{n}, \\ \mathrm{Crit}(m) &\leq \bigl(1+\delta^+ + 26\varepsilon_n\bigr)\frac{R_m}{n}.\end{aligned}$$

If $D_m > C_n(\delta_-, \delta^+)R_{m_o}$,

$$\begin{aligned}\mathrm{Crit}(m) &\geq (1+\delta_- - 16\varepsilon_n)\frac{D_m}{n} > (1+\delta_- - 46\varepsilon_n)\frac{D_m}{n} \\ &> \bigl(1+\delta^+ + 26\varepsilon_n\bigr)\frac{R_{m_o}}{n} \geq \mathrm{Crit}(m_o).\end{aligned}$$

Hence $D_{\hat{m}} \leq C_n(\delta_-, \delta^+)R_{m_o}$. Moreover, from (15), for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s - \tilde{s}\|^2 &\leq \|s - \hat{s}_m\|^2 + \bigl(\mathrm{pen}(m) - 2p(m)\bigr) + \bigl(2p(\hat{m}) - \mathrm{pen}(\hat{m})\bigr) + \delta(\hat{m}, m) \\ &\leq \|s - \hat{s}_m\|^2 + 2\left(\frac{D_m}{n} - p(m)\right) + \bigl(\delta^+ + 6\varepsilon_n\bigr)\frac{R_m}{n} + 2\left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) + (-\delta_- + 6\varepsilon_n)\frac{R_{\hat{m}}}{n} \\ &\leq \|s - \hat{s}_m\|^2 + \bigl(46\varepsilon_n + \delta^+\bigr)\frac{R_m}{n} + (26\varepsilon_n - \delta_-)\frac{R_{\hat{m}}}{n}.\end{aligned}$$

For all m in $\mathcal{M}_n$, on Ω_T,

$$\|s - \hat{s}_m\|^2 = \frac{R_m}{n} + \left(p(m) - \frac{D_m}{n}\right) \geq (1 - 20\varepsilon_n)\frac{R_m}{n}.$$

Hence, for all $m \in \mathcal{M}_n$,

$$\|s - \tilde{s}\|^2 \leq \|s - \hat{s}_m\|^2\left(1 + \frac{46\varepsilon_n + \delta^+}{1 - 20\varepsilon_n}\right) + \frac{26\varepsilon_n - \delta_-}{1 - 20\varepsilon_n}\|s - \tilde{s}\|^2,$$

i.e.,

$$\frac{1 - 46\varepsilon_n - \delta_-}{1 - 20\varepsilon_n}\|s - \tilde{s}\|^2 \leq \frac{1 + 26\varepsilon_n + \delta^+}{1 - 20\varepsilon_n}\|s - \hat{s}_m\|^2.$$

This concludes the proof of Theorem 3.2.

5.4. *Proof of Theorem 3.3*

We keep the notation Ω_T introduced in Lemma A.10. Recall that $\mathbb{P}(\Omega_T^c) \leq C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$, and that, on Ω_T,

$$\begin{aligned}&\forall m \in \mathcal{M}_n, \quad (1 - 20\varepsilon_n)\frac{R_m}{n} \leq \|s - \hat{s}_m\|^2, \\ &\forall m, m' \in \mathcal{M}_n^2, \quad \delta\bigl(m, m'\bigr) \leq 6\varepsilon_n\frac{R_m \vee R_{m'}}{n}.\end{aligned}$$

Let $\tilde{\Omega}_p$ be the event defined in Lemma A.13 and let $\Omega = \tilde{\Omega}_p \cap \Omega_T$, from Lemma A.10, $\mathbb{P}(\Omega^c) \leq C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$. Recall that $\mathrm{pen}(m) = 2D_m^W/n$ (the notation D_m^W is introduced in Proposition A.12). On Ω, from (15), for all n such that $20\varepsilon_n < 1$, for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s - \tilde{s}\|^2 &\leq \|s - \hat{s}_m\|^2 + 26\varepsilon_n \frac{R_m}{n} + 16\varepsilon_n \frac{R_{\hat{m}}}{n} \\ &\leq \|s - \hat{s}_m\|^2 + \frac{26\varepsilon_n}{1 - 20\varepsilon_n}\|s - \hat{s}_m\|^2 + \frac{16\varepsilon_n}{1 - 20\varepsilon_n}\|s - \tilde{s}\|^2.\end{aligned}$$

Hence, for all n such that $20\varepsilon_n < 1$, on Ω,

$$(1 - 36\varepsilon_n)\|s - \tilde{s}\|^2 \leq (1 + 6\varepsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

For all n such that $1 - 36\varepsilon_n > 0$ and $42/(1 - 36\varepsilon_n) < 100$,

$$\|s - \tilde{s}\|^2 \leq \left(1 + \frac{42\varepsilon_n}{1 - 36\varepsilon_n}\right) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 \leq (1 + 100\varepsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

Hence (8) holds for sufficiently large n, it holds in general provided that we enlarge the constant C if necessary.

Appendix

This appendix is devoted to the proof of the concentration inequalities that we used in the main proofs.

A.1. *Probabilistic tools*

The main tool is Lemma A.5 based on Bousquet's version of Talagrand's inequality. It is a concentration inequality for the square of the supremum of the empirical process over a uniformly bounded class of functions. Recall first Bousquet's [11] and Klein and Rio [18] versions of Talagrand's inequality.

Theorem A.1 (Bousquet's bound). *Let $X_1, \ldots, X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t \in S} \mathrm{Var}(t(X))$ and let $Z = \sup_{t \in S} \nu_n t$. Then*

$$\forall x > 0, \quad \mathbb{P}\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}\big(v^2 + 2b\mathbb{E}(Z)\big)x} + \frac{bx}{3n}\right) \leq \mathrm{e}^{-x}.$$

Theorem A.2 (Klein and Rio's bound). *Let $X_1, \ldots, X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t \in S} \mathrm{Var}(t(X))$ and let $Z = \sup_{t \in S} \nu_n t$. Then*

$$\forall x > 0, \quad \mathbb{P}\left(Z < \mathbb{E}(Z) - \sqrt{\frac{2}{n}\big(v^2 + 2b\mathbb{E}(Z)\big)x} - \frac{8bx}{3n}\right) \leq \mathrm{e}^{-x}.$$

Let us now also recall Bernstein's inequality.

Proposition A.3 (Bernstein's inequality). *Let $X_1, \ldots, X_n$ be i.i.d. random variables valued in a measurable space $(X, \mathcal{X})$ and let t be a measurable real valued and bounded function. Then, for all $x > 0$,*

$$\mathbb{P}\left(\nu_n(t) > \sqrt{\frac{2\,\mathrm{Var}(t(X_1))x}{n}} + \frac{\|t\|_\infty x}{3n}\right) \leq \mathrm{e}^{-x}.$$

We derive from these bounds the following useful corollary.

Corollary A.4. *Let S be a symetric class of real valued functions upper bounded by b, $v^2 = \sup_{t\in S} \operatorname{Var}(t(X))$, $Z = \sup_{t\in S} \nu_n t$, $n\mathbb{E}(Z^2) = D$, $e_b = b^2/n$ and*

$$nE_m = 225e_b + \bigl(2.1 + \sqrt{2\pi}\bigr)\sqrt{v^2 D} + \sqrt{15}D^{3/4}e_b^{1/4},$$

then

$$\mathbb{E}\bigl(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}\bigr) \leq \bigl(\mathbb{E}(Z)\bigr)^2\mathbb{P}\bigl(Z \geq \mathbb{E}(Z)\bigr) + E_m. \tag{16}$$

In particular,

$$\bigl(\mathbb{E}(Z)\bigr)^2 \leq \mathbb{E}\bigl(Z^2\bigr) \leq \bigl(\mathbb{E}(Z)\bigr)^2 + E_m. \tag{17}$$

Proof. Since S is symetric, we always have $Z \geq 0$. We have

$$\begin{aligned}\mathbb{E}\bigl(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}\bigr) &= \int_0^\infty \mathbb{P}\bigl(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)} > x\bigr)\,\mathrm{d}x = \int_0^\infty \mathbb{P}\bigl(Z\mathbf{1}_{Z\geq\mathbb{E}(Z)} > \sqrt{x}\bigr)\,\mathrm{d}x \\ &= \bigl(\mathbb{E}(Z)\bigr)^2\mathbb{P}\bigl(Z \geq \mathbb{E}(Z)\bigr) + \int_{(\mathbb{E}(Z))^2}^\infty \mathbb{P}\bigl(Z > \sqrt{x}\bigr)\,\mathrm{d}x.\end{aligned}$$

Take $x = (\mathbb{E}(Z) + \sqrt{2(v^2 + 2b\mathbb{E}(Z))y/n} + by/(3n))^2$ in the previous integral, from Bousquet's version of Talagrand's inequality,

$$\begin{aligned}\int_{(\mathbb{E}[Z])^2}^\infty \mathbb{P}\bigl(Z > \sqrt{x}\bigr)\,\mathrm{d}x \leq{}& \mathbb{E}(Z)\sqrt{\frac{2}{n}\bigl(v^2 + 2b\mathbb{E}(Z)\bigr)}\int_0^\infty \frac{\mathrm{e}^{-y}}{\sqrt{y}}\,\mathrm{d}y + \frac{2v^2 + 14b\mathbb{E}(Z)/3}{n}\int_0^\infty \mathrm{e}^{-y}\,\mathrm{d}y \\ &+ \frac{b}{n}\sqrt{\frac{2}{n}\bigl(v^2 + 2b\mathbb{E}(Z)\bigr)}\int_0^\infty \mathrm{e}^{-y}\sqrt{y}\,\mathrm{d}y + \frac{2b^2}{9n^2}\int_0^\infty y\mathrm{e}^{-y}\,\mathrm{d}y.\end{aligned}$$

Classical computations lead to

$$\int_0^\infty \frac{\mathrm{e}^{-y}}{\sqrt{y}}\,\mathrm{d}y = 2\int_0^\infty \mathrm{e}^{-y}\sqrt{y}\,\mathrm{d}y = \sqrt{\pi}, \qquad \int_0^\infty \mathrm{e}^{-y}\,\mathrm{d}y = \int_0^\infty y\mathrm{e}^{-y}\,\mathrm{d}y = 1.$$

Therefore, using repeatedly the inequalities

$$a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha)b \tag{18}$$

and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain, for all $\eta > 0$,

$$\sqrt{ne_b}\mathbb{E}(Z) \leq \frac{e_b}{3\eta^2} + \frac{2\eta}{3}e_b^{1/4}\bigl(\sqrt{n}\mathbb{E}(Z)\bigr)^{3/2}, \qquad \bigl(\sqrt{n}\mathbb{E}(Z)\bigr)^{1/2}e_b^{3/4} \leq \frac{\eta}{3}e_b^{1/4}\bigl(\sqrt{n}\mathbb{E}(Z)\bigr)^{3/2} + \frac{2e_b}{3\sqrt{\eta}}.$$

Thus

$$\begin{aligned}\int_{(\mathbb{E}[Z])^2}^\infty \mathbb{P}\bigl(Z > \sqrt{x}\bigr)\,\mathrm{d}x \leq{}& \left(2v^2 + \frac{2}{9}e_b + v\frac{\sqrt{2\pi e_b}}{2}\right)\frac{1}{n} + \sqrt{\pi}\frac{\sqrt{\sqrt{n}\mathbb{E}(Z)}(e_b)^{3/4}}{n} \\ &+ \left(\frac{14}{3}\sqrt{e_b} + v\sqrt{2\pi}\right)\frac{\sqrt{n}\mathbb{E}(Z)}{n} + 2\sqrt{\pi}\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}(e_b)^{1/4}}{n} \\ \leq{}& \left(2 + \eta\frac{\sqrt{2\pi}}{4}\right)\frac{v^2}{n} + \sqrt{\frac{2\pi}{n}}v\mathbb{E}(Z) + \left(\frac{2}{9} + \frac{\sqrt{2\pi}}{4\eta} + \frac{2\sqrt{\pi}}{3\sqrt{\eta}} + \frac{14}{9\eta^2}\right)\frac{e_b}{n} \\ &+ \left(\eta\left(\frac{\sqrt{\pi}}{3} + \frac{28}{9}\right) + 2\sqrt{\pi}\right)\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}(e_b)^{1/4}}{n}.\end{aligned}$$

Therefore, taking $\eta = 0.088$, we obtain

$$\int_{(\mathbb{E}[Z])^2}^{\infty} \mathbb{P}\big(Z > \sqrt{x}\big)\,\mathrm{d}x \le 2.1\frac{v^2}{n} + 15^2\frac{e_b}{n} + \sqrt{2\pi}v\frac{\sqrt{n}\mathbb{E}(Z)}{n} + \sqrt{15}\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}(e_b)^{1/4}}{n}.$$

Finally, we use Cauchy–Schwarz inequality to obtain that $\sqrt{n}\mathbb{E}(Z) \le (n\mathbb{E}(Z^2))^{1/2} = (D)^{1/2}$. Since $v^2 \le D$, we get (16). □

We deduce from this result the following concentration inequalities for Z^2.

Corollary A.5. *Let $e_b = b^2/n$. We have, for all $x > 0$,*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{Dv^2x} + 3v^2x + e_b(19x)^2}{n}\right) \le \mathrm{e}^{-x},$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_bx^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.25x)^2}{n}\right) \le \mathrm{e}^{-x+1}.$$

Proof. From Bousquet's version of Talagrand's inequality and from $(\mathbb{E}(Z))^2 \le \mathbb{E}(Z^2)$, we obtain that, for all $x > 0$, with probability larger than $1 - \mathrm{e}^{-x}$, $Z^2 - D/n$ is not larger than

$$\frac{4D^{3/4}(e_bx^2)^{1/4} + \sqrt{D}(14\sqrt{e_bx^2}/3 + 2\sqrt{2v^2x}) + 4D^{1/4}(e_bx^2)^{3/4}/3 + 3v^2x + e_bx^2/3}{n}.$$

We use repeatedly the inequality $a^\alpha b^{1-\alpha} \le \alpha a + (1-\alpha)b$ to obtain that, with probability at least $1 - \mathrm{e}^{-x}$, $Z^2 - D/n$ is not larger than

$$\frac{(4 + 32\eta/9)D^{3/4}(e_bx^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2x} + 3v^2x + (3 + 14/\eta^2 + 8/\sqrt{\eta})e_bx^2/9}{n}.$$

For $\eta = 0.07$, this gives

$$Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2x} + 3v^2x + e_b(19x)^2}{n}.$$

For the second one we use Klein's version of Talagrand's inequality to obtain, for all $x > 0$ such that $r(x) = \sqrt{2(v^2 + 2b\mathbb{E}(Z))x/n} + 8bx/3n < \mathbb{E}(Z)$,

$$\mathbb{P}\big(Z^2 < \big(\mathbb{E}(Z) - r(x)\big)^2\big) \le \mathrm{e}^{-x}.$$

We have $(\mathbb{E}(Z) - r(x))^2 = (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x) + r(x)^2 \ge (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x)$, thus

$$\mathbb{P}\big(Z^2 < \big(\mathbb{E}(Z)\big)^2 - 2\mathbb{E}(Z)r(x)\big) \le \mathrm{e}^{-x}.$$

From the previous corollary, $(\mathbb{E}(Z))^2 \ge \mathbb{E}(Z^2) - E_m$, thus

$$\mathbb{P}\big(Z^2 < \mathbb{E}\big(Z^2\big) - E_m - 2\mathbb{E}(Z)r(x)\big) \le \mathrm{e}^{-x}.$$

Remark that

$$\begin{aligned} 2\mathbb{E}(Z)r(x) &\le \frac{4D^{3/4}(e_bx^2)^{1/4} + 3\sqrt{Dv^2x} + 16\sqrt{De_bx^2}/3}{n} \\ &\le \frac{(4 + 32\eta/9)D^{3/4}(e_bx^2)^{1/4} + 3\sqrt{Dv^2x} + 16/(9\eta^2)e_bx^2}{n}. \end{aligned}$$

For $\eta = 0.0357$, we obtain

$$\frac{D}{n} - Z^2 \leq \frac{D^{3/4} e_b^{1/4} (\sqrt{15} + 4.127\sqrt{x}) + \sqrt{v^2 D}(4.61 + 3\sqrt{x}) + 225 e_b (6.2x^2 + 1)}{n}. \tag{19}$$

In order to conclude the proof, we remark that the inequality is trivial when $x \leq 1$, thus we only have to use (19) for $x > 1$ and then $\sqrt{x} > 1$ and $x^2 > 1$. □

We will use this lemma to obtain a concentration inequality for totally degenerate U-statistics of order 2. The following result generalizes a previous inequality due to Houdré and Reynaud-Bouret [17] to random variables taking values in a measurable space.

Lemma A.6. *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(t_\lambda)_{\lambda \in \Lambda}$ be a set of functions in $L^2(\mu)$. Let*

$$B = \left\{ t = \sum_{\lambda \in \Lambda} a_\lambda t_\lambda, \sum_{\lambda \in \Lambda} a_\lambda^2 \leq 1 \right\}, \qquad D = \mathbb{E}\Big(\sup_{t \in B} \big(t(X) - Pt \big)^2 \Big),$$

$$v^2 = \sup_{t \in B} \operatorname{Var}\big(t(X)\big), \qquad b = \sup_{t \in B} \|t\|_\infty \quad \text{and} \quad e_b = \frac{b^2}{n}.$$

Let

$$U = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in \Lambda} \big(t_\lambda(X_i) - Pt_\lambda \big) \big(t_\lambda(X_j) - Pt_\lambda \big).$$

Then the following inequality holds

$$\forall x > 0, \quad \mathbb{P}\left(U > \frac{5.31 D^{3/4} (e_b x^2)^{1/4} + 3\sqrt{v^2 D x} + 3v^2 x + e_b (19.1x)^2}{n-1} \right) \leq 2\mathrm{e}^{-x}, \tag{20}$$

$$\forall x > 0, \quad \mathbb{P}\left(U < -\frac{9 D^{3/4} (e_b x^2)^{1/4} + 7.61\sqrt{v^2 D x} + e_b (40.3x)^2}{n-1} \right) \leq 3.8\mathrm{e}^{-x}. \tag{21}$$

Proof. Remark that, from Cauchy–Schwarz inequality,

$$\sup_{t \in B} \big(\nu_n(t) \big)^2 = \left(\sup_{\sum a_\lambda^2 \leq 1} \sum_{\lambda \in \Lambda} a_\lambda \nu_n(t_\lambda) \right)^2 = \sum_{\lambda \in \Lambda} \big(\nu_n(t_\lambda) \big)^2.$$

For all x in $\mathbb{X}$, from Cauchy–Schwarz inequality,

$$\sup_{t \in B} \big(t(x) - Pt \big)^2 = \sum_{\lambda} \big(t_\lambda(x) - Pt_\lambda \big)^2,$$

in particular, $D = \sum_{\lambda \in \Lambda} \operatorname{Var}(\psi_\lambda(X))$. Moreover, easy algebra leads to

$$\begin{aligned} \sum_{\lambda \in \Lambda} \big(\nu_n(t_\lambda) \big)^2 &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\lambda \in \Lambda} \big(t_\lambda(X_i) - Pt_\lambda \big)^2 + \frac{1}{n^2} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in \Lambda} \big(t_\lambda(X_i) - Pt_\lambda \big) \big(t_\lambda(X_j) - Pt_\lambda \big) \\ &= \frac{1}{n} P_n \left(\sum_{\lambda \in \Lambda} (t_\lambda - Pt_\lambda)^2 \right) + \frac{n-1}{n} U. \end{aligned}$$

Let $Z^2 = \sup_{t\in B}(\nu_n(t))^2$, $T_\Lambda = \sum_{\lambda\in\Lambda}(t_\lambda - Pt_\lambda)^2$,

$$\mathbb{E}\big(Z^2\big) = \mathbb{E}\bigg(\frac{1}{n}P_nT_\Lambda\bigg) = \frac{D}{n}.$$

Hence

$$U = \frac{n}{n-1}\bigg(Z^2 - \mathbb{E}\big(Z^2\big) - \frac{1}{n}\nu_n(T_\Lambda)\bigg).$$

From Corollary A.5, for all $x > 0$,

$$\mathbb{P}\bigg(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b(19x)^2}{n}\bigg) \le \mathrm{e}^{-x},$$

$$\mathbb{P}\bigg(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_b(x)^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.25x)^2}{n}\bigg) \le 2.8\mathrm{e}^{-x}.$$

Moreover, from Bernstein inequality, for all $x > 0$,

$$\mathbb{P}\bigg(-\nu_nT_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\bigg) \le \mathrm{e}^{-x},$$

$$\mathbb{P}\bigg(\nu_nT_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\bigg) \le \mathrm{e}^{-x}.$$

We apply inequality (18) with $a = D^{3/4}(e_bx^2)^{1/4}$, $b = e_b\sqrt{x}$, $\alpha = 2/3$ and we obtain

$$\mathbb{P}\bigg(-\nu_nT_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}\big(e_bx^2\big)^{1/4} + e_b\bigg(\frac{x+\sqrt{2x}}{3}\bigg)\bigg) \le \mathrm{e}^{-x},$$

$$\mathbb{P}\bigg(\nu_nT_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}\big(e_bx^2\big)^{1/4} + e_b\bigg(\frac{x+\sqrt{2x}}{3}\bigg)\bigg) \le \mathrm{e}^{-x}.$$

Therefore, for all $x > 0$,

$$\mathbb{P}\bigg(U > \frac{5.31D^{3/4}(e_bx^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b((19x)^2 + (x+\sqrt{2x})/3)}{n-1}\bigg) \le 2\mathrm{e}^{-x},$$

$$\mathbb{P}\bigg(U < -\frac{9D^{3/4}(e_bx^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b((40.25x)^2 + (x+\sqrt{2x})/3)}{n-1}\bigg) \le 3.8\mathrm{e}^{-x}.$$

These inequalities are trivial when $x < 1$. We only use them when $x > 1$ and we obtain (20) and (21) since $x < x^2$ and $\sqrt{x} < x^2$ when $x > 1$. □

Let us now state the following corollary of Bernstein's inequality.

Lemma A.7. *Let $X, X_1, \dots, X_n$ be i.i.d. random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let L be a linear functional in $L^2(\mu)$ and let $B = \{t = \sum_{\lambda\in\Lambda} a_\lambda L(\psi_\lambda),\ \sum_{\lambda\in\Lambda} a_\lambda^2 \le 1\}$, $v^2 = \sup_{t\in B}\operatorname{Var}(t(X))$, $b = \sup_{t\in B}\|t\|_\infty$ and $e_b = b^2/n$. Let u be a function in S, the linear space spanned by the functions $(\psi_\lambda)_{\lambda\in\Lambda}$ and let $\eta > 0$. Then the following inequality holds*

$$\forall x > 0, \quad \mathbb{P}\bigg(\nu_n\big(L(u)\big) > \frac{\eta}{2}\|u\|^2 + \frac{2v^2x + e_bx^2/9}{\eta n}\bigg) \le \mathrm{e}^{-x}. \tag{22}$$

Proof. From Bernstein's inequality,

$$\forall x>0,\quad \mathbb{P}\left(\nu_n\big(L(u)\big)>\sqrt{\frac{2\operatorname{Var}(L(u)(X))x}{n}}+\frac{\|L(u)\|_\infty x}{3n}\right)\le \mathrm{e}^{-x}.$$

Since $t=L(u/\|u\|)$ belongs to B,

$$\sqrt{\frac{2\operatorname{Var}(L(u)(X))x}{n}}+\frac{\|L(u)\|_\infty x}{3n}=\|u\|\left(\sqrt{\frac{2\operatorname{Var}(t(X))x}{n}}+\frac{\|t\|_\infty x}{3n}\right)\le\frac{\eta}{2}\|u\|^2+\frac{1}{2\eta}\left(\sqrt{\frac{2v^2x}{n}}+\frac{bx}{3n}\right)^2.$$

We conclude the proof using the inequality $(a+b)^2\le 2a^2+2b^2$. □

A.2. Concentration of the ideal penalty

Let us remark that, for all m in $\mathcal{M}_n$, $p(m)$ is the supremum of the centered empirical process over the ellipsoid $B_m=\{t\in S_m,\ \|t\|\le 1\}$. From Cauchy–Schwarz inequality, for all real numbers $(b_\lambda)_{\lambda\in m}$,

$$\sum_{\lambda\in m}b_\lambda^2=\left(\sup_{\sum a_\lambda^2\le 1}\sum_{\lambda\in m}a_\lambda b_\lambda\right)^2. \tag{23}$$

We apply this equality with $b_\lambda=\nu_n(\psi_\lambda)$. We obtain, since the system $(\psi_\lambda)_{\lambda\in m}$ is orthonormal,

$$\sum_{\lambda\in m}\big(\nu_n(\psi_\lambda)\big)^2=\sup_{\sum a_\lambda^2\le 1}\left(\sum_{\lambda\in m}a_\lambda\nu_n(\psi_\lambda)\right)^2=\sup_{\sum a_\lambda^2\le 1}\left(\nu_n\left(\sum_{\lambda\in m}a_\lambda\psi_\lambda\right)\right)^2=\sup_{t\in B_m}\big(\nu_n(t)\big)^2.$$

Hence, $p(m)$ is bounded by a Talagrand's concentration inequality (see Talagrand [26]). This inequality involves $D_m=n\mathbb{E}(\|\hat{s}_m-s_m\|^2)$ and the constants

$$e_m=\frac{1}{n}\sup_{t\in B_m}\|t\|_\infty^2\quad\text{and}\quad v_m^2=\sup_{t\in B_m}\operatorname{Var}\big(t(X)\big). \tag{24}$$

More precisely, the following proposition is a straightforward application of Corollary A.5.

Proposition A.8. *Let $X,X_1,\dots,X_n$ be i.i.d. random variables with common density s with respect to a probability measure μ. Assume that s belongs to $L^2(\mu)$ and let S_m be a linear subspace in $L^2(\mu)$. Let s_m and $\hat{s}_m$ be respectively the orthogonal projection and the projection estimator of s onto S_m. Let $p(m)=\|s_m-\hat{s}_m\|^2$, $D_m=n\mathbb{E}(p(m))$ and let v_m, e_m be the constants defined in* (24)*. Then, for all $x>0$,*

$$\mathbb{P}\left(p(m)-\frac{D_m}{n}>\frac{D_m^{3/4}(e_mx^2)^{1/4}+0.7\sqrt{D_mv_m^2x}+0.15v_m^2x+e_mx^2}{n}\right)\le\mathrm{e}^{-x/20}, \tag{25}$$

$$\mathbb{P}\left(\frac{D_m}{n}-p(m)>\frac{1.8D_m^{3/4}(e_mx^2)^{1/4}+1.71\sqrt{D_mv_m^2x}+4.06e_mx^2}{n}\right)\le 2.8\mathrm{e}^{-x/20}. \tag{26}$$

A.3. Computation of the union bounds

Let us prove a simple result

Lemma A.9. *For all $K>1$,*

$$\Sigma(K)=\sum_{k\in\mathbb{N}}\sum_{m\in\mathcal{M}_n^k}\mathrm{e}^{-K[\ln(1+\operatorname{Card}(\mathcal{M}_n^{[k]}))+\ln(1+k)]}<\infty. \tag{27}$$

For all m in $\mathcal{M}_n$, let $l_m = l_{n,\gamma}(R_m, R_m)$, then, for all $K > 1/\sqrt{2}$,

$$\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-K^2 l_m} = \Sigma\big(2K^2\big)\mathrm{e}^{-K^2(\ln n)^\gamma}. \tag{28}$$

For all m, m' in $\mathcal{M}_n$, let $l_{m,m'} = l_{n,\gamma}(R_m, R_{m'})$, then, for all $K > 1$,

$$\sum_{(m,m')\in(\mathcal{M}_n)^2} \mathrm{e}^{-K^2 l_{m,m'}} = \big(\Sigma\big(K^2\big)\big)^2\mathrm{e}^{-K^2(\ln n)^\gamma}. \tag{29}$$

Proof. Inequality (27) comes from the fact that, when $K > 1$,

$$\forall k \in \mathbb{N}, \quad \sum_{m\in\mathcal{M}_n^{[k]}} \mathrm{e}^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_n^{[k]}))]} \le 1 \quad \text{and} \quad \sum_{k\in\mathbb{N}^*} \mathrm{e}^{-K\ln k} < \infty.$$

For all integers k such that $\mathcal{M}_n^{[k]} \neq \varnothing$, for all m in $\mathcal{M}_n^{[k]}$, $l_m \ge 2[\ln(1+\mathrm{Card}(\mathcal{M}_n^{[k]})) + \ln(1+k)] + (\ln n)^\gamma$, thus, for all $K > 1/\sqrt{2}$, it comes from (27) that

$$\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-K^2 l_m} \le \mathrm{e}^{-K^2(\ln n)^\gamma} \sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^{[k]}} \mathrm{e}^{-2K^2[\ln(1+\mathrm{Card}(\mathcal{M}_n^{[k]}))+\ln(1+k)]} \le \Sigma\big(2K^2\big)\mathrm{e}^{-K^2(\ln n)^\gamma}.$$

Finally, for all integers (k, k') such that $\mathcal{M}_n^{[k]} \times \mathcal{M}_n^{[k']} \neq \varnothing$,

$$l_{m,m'} \ge \ln\big(1+\mathrm{Card}\big(\mathcal{M}_n^{[k]}\big)\big) + \ln(1+k) + \ln\big(1+\mathrm{Card}\big(\mathcal{M}_n^{[k']}\big)\big) + \ln\big(1+k'\big) + (\ln n)^\gamma.$$

Thus, from (27),

$$\sum_{(m,m')\in(\mathcal{M}_n^2)} \mathrm{e}^{-K^2 l_{m,m'}} = \Bigg(\sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^k} \mathrm{e}^{-K^2[\ln(1+\mathrm{Card}(\mathcal{M}_n^{[k]}))+\ln(1+k)]}\Bigg)^2 \mathrm{e}^{-K^2(\ln n)^\gamma}.$$

□

Lemma A.10. *Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* [V]. *We consider the following events.*

$$\Omega_\delta = \Big\{\forall\big(m,m'\big) \in \mathcal{M}_n^2, \delta\big(m,m'\big) \le 6\varepsilon_n \frac{R_m \vee R_{m'}}{n}\Big\},$$

$$\Omega_p = \bigcap_{m\in\mathcal{M}_n} \left\{\left\{p(m) - \frac{D_m}{n} \le 10\varepsilon_n \frac{R_m}{n}\right\} \cap \left\{p(m) - \frac{D_m}{n} \ge -20\varepsilon_n \frac{R_m}{n}\right\}\right\}$$

and $\Omega_T = \Omega_\delta \cap \Omega_p$. Then there exists a constant $C > 0$ such that

$$\mathbb{P}\big(\Omega_\delta^c\big) \le C\mathrm{e}^{-(\ln n)^\gamma}, \qquad \mathbb{P}\big(\Omega_p^c\big) \le C\mathrm{e}^{-(1/2)(\ln n)^\gamma}, \qquad \mathbb{P}\big(\Omega_T^c\big) \le C\mathrm{e}^{-(1/2)(\ln n)^\gamma}.$$

Proof. Let $K > 1$ be a constant to be chosen later. We apply Lemma A.7 to $u = s_m - s_{m'}$, $S = S_m + S_{m'}$, $L = id$, $x = K^2 l_{n,\gamma}(R_m, R_{m'})$. For all $\eta > 0$, for all m, m' in $\mathcal{M}_n$, on an event of probability larger than $1 - \mathrm{e}^{-K^2 l_{n,\gamma}(R_m, R_{m'})}$,

$$\delta\big(m,m'\big) \le \frac{\eta}{2}\|s_m - s_{m'}\|^2 + \frac{2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m, R_{m'}) + e_{m,m'}(K^2 l_{n,\gamma}(R_m, R_{m'}))^2/9}{\eta n}. \tag{30}$$

From [V], for all m, m' in $\mathcal{M}_n$,

$$2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m, R_{m'}) + \frac{e_{m,m'}(K^2 l_{n,\gamma}(R_m, R_{m'}))^2}{9} \le \left(2(K\varepsilon_n)^2 + \frac{(K\varepsilon_n)^4}{9}\right)\frac{R_m \vee R_{m'}}{n}.$$

Moreover, for all m, m' in $\mathcal{M}_n$,

$$\|s_m - s_{m'}\|^2 \le 2\big(\|s - s_m\|^2 + \|s - s_{m'}\|^2\big) \le 2(R_m + R_{m'}) \le 4(R_m \vee R_{m'}).$$

Let $e_n(K) = \sqrt{(K\varepsilon_n)^2 + (K\varepsilon_n)^4/18}$. In (30) we take $\eta = e_n(K)$ and we obtain

$$\mathbb{P}\left(\delta\big(m, m'\big) > 4e_n(K)\frac{R_m \vee R_{m'}}{n}\right) \le \mathrm{e}^{-Kl_{n,\gamma}(R_m, R_{m'})}. \tag{31}$$

From (29), for all $K > 1$,

$$\mathbb{P}\left(\forall\big(m, m'\big) \in \mathcal{M}_n^2, \delta\big(m, m'\big) > 4e_n(K)\frac{R_m \vee R_{m'}}{n}\right) \le \big(\Sigma(K)\big)^2\mathrm{e}^{-K(\ln n)^2}.$$

Let $K = 1.1$ and take n sufficiently large so that $K^4\varepsilon_n^2/18 \le 1$, then $4e_n(K) \le 6\varepsilon_n$. Hence, the first conclusion of Lemma A.10 holds for sufficiently large n, it holds in general, provided that we increase the constant C if necessary.

We apply Assumption [V] (see (14)) with $m = m'$, let $l_m = l_{n,\gamma}(R_m, R_m)$, for all $K > 0$, for all n such that $4.06(K\varepsilon_n)^3 \le 2$,

$$\begin{aligned}
&\frac{D_m^{3/4}(e_m(K^2l_m)^2)^{1/4} + 0.7\sqrt{D_m v_m^2 K^2 l_m} + 0.15v_m^2K^2l_m + e_m(K^2l_m)^2}{n} \\
&\quad\le \big(1.7K\varepsilon_n + 0.15(K\varepsilon_n)^2 + (K\varepsilon_n)^4\big)\frac{R_m}{n} \le 3K\varepsilon_n\frac{R_m}{n}, \\
&\frac{1.8D_m^{3/4}(e_m(K^2l_m)^2)^{1/4} + 1.71\sqrt{D_m v_m^2(K^2l_m)} + 4.06e_m(K^2l_m)^2}{n} \\
&\quad\le \big(3.51K\varepsilon_n + 4.06(K\varepsilon_n)^4\big)\frac{R_m}{n} \le 6K\varepsilon_n\frac{R_m}{n}.
\end{aligned}$$

It comes then from Proposition A.8 applied with $x = K^2l_m$ that, for all m in $\mathcal{M}_n$

$$\mathbb{P}\left(p(m) - \frac{D_m}{n} > 3K\varepsilon_n\frac{R_m}{n}\right) \le \mathrm{e}^{-(K^2/20)l_m}.$$

Thus, from (28), for all $K > \sqrt{10}$, and for all n sufficiently large,

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ p(m) - \frac{D_m}{n} > 3K\varepsilon_n\frac{R_m}{n}\right) \le \Sigma\big(K^2/10\big)\mathrm{e}^{-(K^2/20)(\ln n)^\gamma}.$$

We use the same arguments to prove that

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n, p(m) - \frac{D_m}{n} < 6K\varepsilon_n\frac{R_m}{n}\right) \le \Sigma\big(K^2/10\big)\mathrm{e}^{-(K^2/20)(\ln n)^\gamma}.$$

Fix $K = \sqrt{10.5}$, then for all n sufficiently large, the conclusion of Lemma A.10 holds. It holds in general provided that we increase the constant C if necessary. □

A.4. *Concentration of the resampling penalty*

Lemma A.11. *Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$ and let L be a linear functional defined on $L^2(\mu)$. Let $p(\Lambda) = \sum_{\lambda\in\Lambda}(v_n(L(\psi_\lambda)))^2$. Let $(W_1, \dots, W_n)$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$ and let $v_W^2 = \mathrm{Var}(W_1 - \bar{W}_n)$. Let*

$$D_\Lambda^W = n\big(v_W^2\big)^{-1}\sum_{\lambda\in\Lambda}\mathbb{E}^W\big(\big(v_n^W\big(L(\psi_\lambda)\big)\big)^2\big),$$

$T = \sum_{\lambda\in\Lambda}(L(\psi_\lambda) - PL(\psi_\lambda))^2$, $D = PT$ *and*

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} \big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big)\big(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)\big)$$

then

$$p(\Lambda) = \frac{1}{n} P_n T + \frac{n-1}{n} U, \qquad D_\Lambda^W = P_n T - U, \qquad p(\Lambda) - \frac{D_\Lambda^W}{n} = U,$$
$$\mathbb{E}\big(D_\Lambda^W\big) = D, \qquad D_\Lambda^W - D = \nu_n T - U.$$

Proof. It is easy to check that

$$\begin{aligned} p(\Lambda) &= \sum_{\lambda\in\Lambda} \left(\frac{1}{n} \sum_{i=1}^{n} L(\psi_\lambda)(X_i) - PL(\psi_\lambda) \right)^2 \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big)^2 + \frac{1}{n^2} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} \big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big)\big(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)\big) \\ &= \frac{1}{n} P_n T + \frac{n-1}{n} U. \end{aligned}$$

Recall that $\nu_n^W = P_n^W - \bar{W}_n P_n$. For all λ in Λ, since $\sum_{i=1}^n (W_i - \bar{W}_n) = 0$,

$$\nu_n^W\big(L(\psi_\lambda)\big) = \frac{1}{n} \sum_{i=1}^{n} (W_i - \bar{W}_n) L(\psi_\lambda)(X_i) = \frac{1}{n} \sum_{i=1}^{n} (W_i - \bar{W}_n)\big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big).$$

Thus, if $E_{i,j} = \mathbb{E}((W_i - \bar{W}_n)(W_j - \bar{W}_n))/v_W^2$,

$$\begin{aligned} D_\Lambda^W &= n\big(v_W^2\big)^{-1} \sum_{\lambda\in\Lambda} \mathbb{E}^W \left(\left(\frac{1}{n} \sum_{i=1}^{n} (W_i - \bar{W}_n)\big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big) \right)^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{E}((W_i - \bar{W}_n)^2)}{v_W^2} \big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big)^2 \\ &\quad + \frac{1}{n} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} E_{i,j}\big(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)\big)\big(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)\big). \end{aligned}$$

Since the weights are exchangeable, for all $i = 1, \ldots, n$, $\mathbb{E}((W_i - \bar{W}_n)^2) = \mathrm{Var}(W_1 - \bar{W}_n) = v_W^2$ and for all $i \neq j = 1, \ldots, n$,

$$v_W^2 E_{i,j} = \mathbb{E}\big((W_i - \bar{W}_n)(W_j - \bar{W}_n)\big) = \mathbb{E}\big((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\big).$$

Moreover, since $\sum_{i=1}^n (W_i - \bar{W}_n) = 0$,

$$\begin{aligned} 0 &= E\left[\left(\sum_{i=1}^{n} (W_i - \bar{W}_n) \right)^2 \right] = \sum_{i=1}^{n} \mathbb{E}\big((W_i - \bar{W}_n)^2\big) + \sum_{i\neq j=1}^{n} v_W^2 E_{i,j} \\ &= n\mathbb{E}\big((W_1 - \bar{W}_n)^2\big) + n(n-1)\mathbb{E}\big((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\big). \end{aligned}$$

Hence, for all $i \neq j = 1, \dots, n$, $E_{i,j} = -1/(n-1)$, thus

$$D_\Lambda^W = P_n T - U.$$

The last inequalities of Lemma A.11 follow from the fact that $\mathbb{E}(U) = 0$. Finally,

$$p(\Lambda) - \frac{D_\Lambda^W}{n} = \frac{1}{n} P_n T + \frac{n-1}{n} U - \left(\frac{1}{n} P_n T - \frac{1}{n} U \right) = U. \qquad \square$$

Proposition A.12. *Let $(W_1, \dots, W_n)$ be a resampling scheme, let S_m be a linear space, $B_m = \{t \in S_m, \|t\| \leq 1\}$, $p(m) = \sup_{t \in B_m} (\nu_n(t))^2$, $D_m = n\mathbb{E}(p(m))$ and let D_m^W be the resampling estimator of D_m based on $(W_1, \dots, W_n)$, that is $D_m^W = nC_W^2 \mathbb{E}^W(\nu_n^W(\hat{s}_m^W))$, where $v_W^2 = \mathrm{Var}(W_1 - \bar{W}_n)$ and $C_W^2 = (v_W^2)^{-1}$.*

Then, for all m in $\mathcal{M}_n$, $\mathbb{E}(D_m^W) = D_m$. Moreover, let e_m, v_m be the quantities defined in (24). For all $x > 0$, on an event of probability larger than $1 - 7.8\mathrm{e}^{-x}$,

$$D_m^W - D_m \leq \sqrt{8 e_m D_m x} + e_m \left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1} \right) + \frac{9 D_m^{3/4} (e_m x^2)^{1/4} + 7.61 \sqrt{v_m^2 D_m x}}{n-1}, \tag{32}$$

$$D_m^W - D_m \geq -\sqrt{8 e_m D_m x} - e_m \left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1} \right) - \frac{5.31 D_m^{3/4} (e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3 v_m^2 x}{n-1}. \tag{33}$$

For all $x > 0$,

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31 D_m^{3/4} (e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3 v_m^2 x + e_m (19.1x)^2}{n-1} \right) \leq 2\mathrm{e}^{-x}, \tag{34}$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) \leq \frac{9 D_m^{3/4} (e_m x^2)^{1/4} + 7.61 \sqrt{v_m^2 D_m x} + e_m (40.3x)^2}{n-1} \right) \leq 3.8\mathrm{e}^{-x}. \tag{35}$$

Proof. The exchangeability property ensures that $\mathbb{E}^W(\bar{W}_n^2) = n^{-1} \sum_{i=1}^n \mathbb{E}(W_i \bar{W}_n) = \mathbb{E}(W_1 \bar{W}_n)$. We deduce that

$$\mathbb{E}^W\big((P_n^W - \bar{W}_n P_n)(\bar{W}_n \hat{s}_m) \big) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}^W\big((W_i - \bar{W}_n) \bar{W}_n \big) \hat{s}_m(X_i) = 0.$$

Let us recall that $\hat{s}_m^W = \arg\min_{t \in S_m} \{ \|t\|^2 - 2 P_n^W t \} = \sum_{\lambda \in m} (P_n^W \psi_\lambda) \psi_\lambda$. Hence

$$\mathbb{E}^W\big(\nu_n^W(\hat{s}_m^W) \big) = \mathbb{E}^W\big(\nu_n^W(\hat{s}_m^W - \bar{W}_n \hat{s}_m) \big) = \sum_{\lambda \in m} \mathbb{E}^W\big((\nu_n^W \psi_\lambda)^2 \big).$$

We apply Lemma A.11 with $L = id$ and $\Lambda = m$. By definition of $p(m)$ and D_m^W,

$$p(m) - \frac{D_m^W}{n} = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^n \sum_{\lambda \in m} \big(\psi_\lambda(X_i) - P\psi_\lambda \big) \big(\psi_\lambda(X_j) - P\psi_\lambda \big).$$

Thus, from Lemma A.6, for all $x > 0$,

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31 D_m^{3/4} (e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3 v_m^2 x + e_m (19.1x)^2}{n-1} \right) \leq 2\mathrm{e}^{-x},$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) > \frac{9 D_m^{3/4} (e_m x^2)^{1/4} + 7.61 \sqrt{v_m^2 D_m x} + e_m (40.3x)^2}{n-1} \right) \leq 3.8\mathrm{e}^{-x}.$$

This proves (34) and (35).

In order to obtain (32) and (33), we introduce, for all m in $\mathcal{M}_n$, the function $T_m = \sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2$ and the random variable

$$U_m = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in m} \big(\psi_\lambda(X_i) - P\psi_\lambda\big)\big(\psi_\lambda(X_j) - P\psi_\lambda\big).$$

We apply Lemma A.11 with $L = id$, we obtain

$$D_m^W - D_m = \nu_n(T_m) - U_m.$$

From Bernstein's inequality (see Proposition A.3), for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi\,\nu_n(T_m) > \sqrt{\frac{2\,\mathrm{Var}(T_m(X))x}{n}} + \frac{\|T_m\|_\infty x}{3n}\right) \leq \mathrm{e}^{-x}.$$

From Cauchy–Schwarz inequality, $T_m = \sup_{t\in B_m}(t - Pt)^2$, thus $\|T_m\|_\infty/n = 4e_m$ and $\mathrm{Var}(T_m(X))/n \leq \|T_m\|_\infty PT_m/n = 4e_m D_m$, therefore, for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi\,\nu_n(T_m) > \sqrt{8e_m D_m x} + \frac{4e_m x}{3}\right) \leq \mathrm{e}^{-x}.$$

Moreover, from Lemma A.6, for all $x > 0$,

$$\mathbb{P}\left(U_m > \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1}\right) \leq 2\mathrm{e}^{-x},$$

$$\mathbb{P}\left(U_m < -\frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1}\right) \leq 3.8\mathrm{e}^{-x}.$$

We deduce that, for all $x > 0$, with probability larger than $1 - 4.8\mathrm{e}^{-x}$,

$$D_m^W - D_m \leq \sqrt{8e_m D_m x} + e_m\left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1}\right) + \frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x}}{n-1}.$$

Moreover, for all $x > 0$, on an event of probability larger than $1 - 3\mathrm{e}^{-x}$,

$$D_m^W - D_m \geq -\sqrt{8e_m D_m x} - e_m\left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1}\right) - \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x}{n-1}.$$

□

Lemma A.13. *Let*

$$\Omega_{\mathrm{u}} = \bigcap_{m\in\mathcal{M}_n} \left\{\frac{D_m^W}{n} - p(m) \leq 10\varepsilon_n \frac{R_m}{n}\right\},$$

$$\Omega_{\mathrm{l}} = \bigcap_{m\in\mathcal{M}_n} \left\{\frac{D_m^W}{n} - p(m) \geq -12\varepsilon_n \frac{R_m}{n}\right\}$$

and $\tilde{\Omega}_p = \Omega_{\mathrm{u}} \cap \Omega_{\mathrm{l}}$. *There exists a constant* $C > 0$ *such that* $\mathbb{P}(\tilde{\Omega}_p^c) \leq C\mathrm{e}^{-(1/2)(\ln n)^\gamma}$.

Proof. From Assumption [V] applied with $m = m'$ (see (14)), if $l_m = l_{n,\gamma}(R_m, R_m)$, for all $K > 0$,

$$D_m^{3/4}\big(e_m\big(K^2 l_m\big)^2\big)^{1/4} \leq K\varepsilon_n R_m, \qquad \sqrt{v_m^2 D_m\big(K^2 l_m\big)} \leq K\varepsilon_n R_m,$$

$$v_m^2\big(K^2 l_m\big) \leq (K\varepsilon_n)^2 R_m, \qquad e_m(Kl_m)^2 \leq (K\varepsilon_n)^4 R_m.$$

We apply Proposition A.12 with $x = K^2 l_m$ and we obtain

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) > \left(8.31K\varepsilon_n + 3(K\varepsilon_n)^2 + (19.1)^2(K\varepsilon_n)^4\right)\frac{R_m}{n-1}\right) \leq 2\mathrm{e}^{-K^2 l_m}.$$

Thus, for all $K > 1/(\sqrt{2})$, if $e_n(K) = n(8.31K\varepsilon_n + 3(K\varepsilon_n)^2 + (19.1)^2(K\varepsilon_n)^4)/(n-1)$, from (28)

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n, \frac{D_m^W}{n} - p(m) > e_n(K)\frac{R_m}{n}\right) \leq 2\Sigma\left(2K^2\right)\mathrm{e}^{-K^2(\ln n)^\gamma}.$$

Take $K = 8/8.31$ and $n \geq 10$ sufficiently large to ensure that $3K^2\varepsilon_n + (19.1)^2K^4\varepsilon_n^3 \leq 1$, then

$$e_n(K) \leq \frac{10}{9}(8\varepsilon_n + \varepsilon_n) \leq 10\varepsilon_n.$$

We deduce that, for sufficiently large n,

$$\mathbb{P}\left(\Omega_{\mathrm{u}}^c\right) \leq 2\Sigma\left(2K^2\right)\mathrm{e}^{-K^2(\ln n)^\gamma}.$$

We also apply Proposition A.12 with $x = K^2 l_m$, and we use the same arguments to prove that, for $K = 16/16.61$, for all $n \geq 10$ sufficiently large to ensure that $(40.3)^2K^4\varepsilon_n^3 \leq 2$

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n, \frac{D_m^W}{n} - p(m) < -20\varepsilon_n\frac{R_m}{n}\right) \leq 3.8\Sigma\left(2K^2\right)\mathrm{e}^{-K^2(\ln n)^\gamma}.$$

Hence, the conclusion of Lemma A.13 holds for sufficiently large n. It holds in general, provided that we increase the constant C if necessary. □

References

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** (1970) 203–217. MR0286233

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. Akadémiai Kiadó, Budapest, 1973. MR0483125

[3] S. Arlot. Resampling and model selection. Ph.D. thesis, Université Paris-Sud 11, 2007.

[4] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.* **3** (2009) 557–624. MR2519533

[5] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** (2009) 245–279.

[6] A. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**(3) (1999) 301–413. MR1679028

[7] J.-P. Baudry, K. Maugis and B. Michel. Slope heuristics: Overview and implementation. Report *INRIA*, 2010. Available at http://hal.archives-ouvertes.fr/hal-00461639/fr/.

[8] L. Birgé. Model selection for density estimation with l^2-loss. Preprint, 2008.

[9] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* 55–87. Springer, New York, 1997. MR1462939

[10] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**(1–2) (2007) 33–73. MR2288064

[11] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334**(6) (2002) 495–500. MR1890640

[12] F. Bunea, A. B. Tsybakov and M. H. Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning Theory* 530–543. *Lecture Notes in Comput. Sci.* **4539**. Springer, Berlin, 2007. MR2397610

[13] A. Célisse. Density estimation via cross validation: Model selection point of view. Preprint, 2008. Available at arXiv.org:08110802.

[14] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.* **24**(2) (1996) 508–539. MR1394974

[15] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**(1) (1979) 1–26. MR0515681

[16] I. Gannaz and O. Wintenberger. Adaptive density estimation under dependence. *ESAIM Probab. Stat.* **14** (2010) 151–172. MR2654551

[17] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications* 55–69. *Progr. Probab.* **56**. Birkhäuser, Basel, 2003. MR2073426

[18] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**(3) (2005) 1060–1077. MR2135312

[19] C. L. Mallows. Some comments on c_p. *Technometrics* **15** (1973) 661–675.

[20] P. Massart. *Concentration Inequalities and Model Selection*. *Lecture Notes in Mathematics* **1896**. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879
[21] P. Rigollet. Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12**(2) (2006) 351–370. MR2218559
[22] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16**(3) (2007) 260–280. MR2356821
[23] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Stat.* **9**(2) (1982) 65–78. MR0668683
[24] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.* **6** (1978) 461–464. MR0468014
[25] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** (1974) 111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking and A. S. Young and with a reply by the authors. MR0356377
[26] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.* **126**(3) (1996) 505–563. MR1419006