Optimal cross-validation in density estimation

Alain Celisse

Optimal cross-validation in density estimation

Alain Celisse

Laboratoire de Mathématiques Painlevé,
UMR 8524 CNRS-Université Lille 1, Modal *team-project INRIA*
F-59 655, Villeneuve d'Ascq Cedex
e-mail: celisse@math.univ-lille1.fr

Abstract: The performance of cross-validation (CV) is analyzed in two contexts: (*i*) *risk estimation* and (*ii*) *model selection* in the density estimation framework. The main focus is given to one CV algorithm called leave-p-out (Lpo), where p denotes the cardinality of the test set. Closed-form expressions are settled for the Lpo estimator of the risk of projection estimators, which makes V-fold cross-validation completely useless.

From a theoretical point of view, these closed-form expressions enable to study the Lpo performances in terms of risk estimation. For instance, the optimality of leave-one-out (Loo), that is Lpo with $p = 1$, is proved among CV procedures. Two model selection frameworks are also considered: *estimation*, as opposed to *identification*.

Unlike risk estimation, Loo is proved to be suboptimal as a model selection procedure. In the estimation framework with finite sample size n, optimality is achieved for p large enough (with $p/n = o(1)$) to balance overfitting. A link is also identified between the optimal p and the structure of the model collection. These theoretical results are strongly supported by simulation experiments. When performing identification, model consistency is also proved for Lpo with $p/n \to 1$ as $n \to +\infty$.

AMS 2000 subject classifications: Primary 62G09; secondary 62G07, 62E17.
Keywords and phrases: Cross-validation, leave-p-out, resampling, risk estimation, model selection, density estimation, oracle inequality, projection estimators, concentration inequalities.

1. Introduction

1.1. Model selection

For estimating a target quantity denoted by s, let $\{S_m\}_{m\in\mathcal{M}}$ denote a collection of sets of candidate parameters and $\mathcal{M}$ denote a set of index. From each S_m called a *model*, an estimator $\widehat{s}_m$ of s is computed. The goal of model selection is to design a criterion crit : $\mathcal{M} \to \mathbb{R}^+$ such that minimizing crit$(\cdot)$ over $\mathcal{M}$ provides a final estimator $\widehat{s}_{\widehat{m}}$ that is "optimal". Among various strategies of model selection, *model selection via penalization* has been introduced in the seminal papers by Mallows (1973); Akaike (1973); Schwarz (1978) on respectively AIC, C_p, and BIC criteria. However since AIC and BIC are derived from asymptotic arguments, their performances crucially depend on model collection and sample size (see Baraud et al., 2009).

More recently Birgé and Massart (1997, 2001, 2006) have developed a non-asymptotic approach inspired from the pioneering work of Barron and Cover (1991). It relies on concentration inequalities (Talagrand, 1996; Ledoux, 2001) and aims at deriving *oracle inequalities* such as

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq C \inf_{m\in\mathcal{M}} \left\{\ell\left(s, \widehat{s}_m\right)\right\} + r_n$$

with probability larger than $1 - c/n^2$, where $c > 0$ is a constant, $\ell(s,t)$ is a measure of the gap between parameters s and t, r_n is a remainder term with respect to $\inf_m \ell\left(s, \widehat{s}_m\right)$, and $C \geq 1$ denotes a constant independent of s. The closer C to 1 and the smaller r_n, the better the model selection procedure. If $C = C_n \to 1$ and $r_n \to 0$ as $n \to +\infty$, the model selection procedure is said *asymptotically optimal* (or efficient) (see Arlot and Celisse, 2010, for instance). Note that other asymptotic optimality properties

have been studied in the literature. For instance, a model selection procedure satisfying

$$\mathbb{P}\left[\,\widehat{m} = m_0\,\right] \xrightarrow[n\to+\infty]{} 1 \ ,$$

where m_0 denotes a fixed given model is said *model consistent* (see Shao, 1997, for a study of various model selection procedures in terms of model consistency).

In the density estimation framework, model selection with deterministic penalties has been developed: (i) for Kullback-Leibler divergence and histograms by Barron et al. (1999); Castellan (1999, 2003) and further studied in Birgé and Rozenholc (2006), and (ii) for quadratic risk and projection estimators by Birgé and Massart (1997) and Barron et al. (1999).

1.2. Cross-validation

The aforementioned approaches rely on some deterministic penalties such as AIC or BIC. These penalties are derived in some specific settings (for instance Birgé and Massart, 2006, assume a Gaussian noise), which makes their performances setting dependent.

Conversely, cross-validation (CV) is a *resampling* procedure based on a *universal heuristics* which makes it applicable in a wide range of settings. CV algorithms have been first studied in a regression context by Stone (1974, 1977) for the leave-one-out (Loo) and Geisser (1974, 1975) for the V-fold cross-validation (VFCV), and in the density estimation framework by Rudemo (1982); Stone (1984). Since these algorithms can be computationally demanding or even intractable, Rudemo (1982); Bowman (1984) derived closed-form formulas for the Loo estimator of the risk of histograms or kernel estimators. These results have been recently extended to the leave-p-out cross-validation (Lpo) by Celisse and Robin (2008).

Although CV algorithms are extensively used in practice, only few theoretical results exist on their performances, most of them being of asymptotic nature. For instance in the regression framework, Burman (1989, 1990) proves Loo is asymptotically the best CV algorithm in terms of risk estimation. Several papers are dedicated to show the equivalence between some CV algorithms and penalized criteria in terms of asymptotic optimality properties: (*i*) *efficiency* in Li (1987); Zhang (1993), and (*ii*) *model consistency* in Shao (1993); Yang (2007). We refer interested readers to Shao (1997) for an extensive review about asymptotic optimality properties in terms of efficiency and model consistency of some penalized criteria as well as CV algorithms.

As for non-asymptotic results in the density framework, Birgé and Massart (1997) have settled an oracle inequality that relies on a conjecture and may be applied to Loo. However to the best of our knowledge, no result of this type has already been proved for Lpo in the density estimation framework. Recently in the regression setting, Arlot (2007) established oracle inequalities for V-fold penalties, while Arlot and Celisse (2011) have carried out an extensive simulation study in the change-point detection problem with heteroscedastic observations.

1.3. Main contributions

In the present paper, we derive closed-form expressions for the Lpo risk estimator of the broad class of projection estimators (Section 2). Such closed-form expressions make V-FCV completely useless since it is more variable and computationally demanding than Lpo (Section 2.3). They also enable to study the theoretical performance of CV in two respects: (*i*) for risk estimation (Section 2.4), and (ii) for model selection (Section 3). For instance, it is proved that Loo is the best CV algorithm for risk estimation (Theorem 2.1), while it is suboptimal for model selection (Corollary 3.1 and Theorem 3.3).

Moreover, two aspects of model selection *via* CV have been explored. In Section 3.1, the *estimation* point of view is described where it is shown that Lpo is optimal as long as $p/n = o(1)$ and p is large enough to balance the influence of the model collection structure. All these new theoretical results are supported

by simulation experiments detailed in Section 3.1.4. Conversely, *identification* is studied in Section 3.2, where the optimal performance is obtained for $p/n \xrightarrow[n\to+\infty]{} 1$, which is consistent with previous results settled in the regression framework for instance by Shao (1993). However, our result is more precise since we were able to localize the optimal rate of convergence of $1 - p/n$ toward 0 between $1/n$ and $1/\sqrt{n}$ as n tends to $+\infty$. Finally, proofs and technical lemmas have been collected in Appendix A.

2. Cross-validation and risk estimation

2.1. Statistical framework

2.1.1. Notation

Throughout the paper, $X_1, \ldots, X_n \in [0,1]$ are independent and identically distributed (*i.i.d.*) random variables drawn from a probability distribution P of density $s \in L^2([0,1])$ with respect to Lebesgue's measure on $[0,1]$, and $X_{1,n} = (X_1, \ldots, X_n)$.

Let $\mathcal{S}^*$ denote the set of mesurable functions on $[0,1]$. The distance between s and any $u \in \mathcal{S}^*$ is measured thanks to the quadratic *loss* denoted by

$$\ell : \ (s,u) \mapsto \ell(s,u) := \|s-u\|^2 = \int_{[0,1]} [\, s(t) - u(t) \,]^2 \, dt \ .$$

It is related to the *contrast* function

$$\gamma : \ (u,x) \mapsto \gamma(u;x) := \|u\|^2 - 2u(x) \ , \quad \text{with} \quad \ell(s,u) = P\gamma(u) - P\gamma(s) \tag{1}$$

where $P\gamma(u) = P(\gamma(u;\cdot))$ and $Pf := \mathbb{E}[\, f(X_1) \,]$ for every $f \in \mathcal{S}^*$. The performance of every estimator $\widehat{s} = \widehat{s}(X_1, \ldots, X_n)$ of s is assessed thanks to the *quadratic risk*

$$R_n(\widehat{s}) := \mathbb{E}[\, \ell(s, \widehat{s}\,) \,] = \mathbb{E}\left[\, \|s - \widehat{s}\,\|^2 \right].$$

Estimating $P\gamma(u)$ is made through the *empirical contrast* defined by

$$P_n\gamma(u) := \frac{1}{n} \sum_{i=1}^{n} \gamma(u; X_i) \, , \quad \text{where} \quad P_n = 1/n \sum_{i=1}^{n} \delta_{X_i}$$

denotes the empirical measure and $P_n f := 1/n \sum_{i=1}^n f(X_i)$ for every $f \in \mathcal{S}^*$.

Let us further introduce $\mathcal{M}_n$ a countable set of indices and for every $m \in \mathcal{M}_n$, S_m denote a set of functions, called *model*, used to estimate s. To each S_m, an estimator $\widehat{s}_m$ corresponds that is defined as the *empirical contrast minimizer*

$$\widehat{s}_m := \mathrm{Argmin}_{u \in S_m} P_n\gamma(u) \ . \tag{2}$$

It results a collection $\{\widehat{s}_m\}_{m \in \mathcal{M}_n}$ of estimators of s depending on the choice of models S_ms. Instances of such models and estimators are described in Section 2.1.2.

2.1.2. Projection estimators

Let Λ_n be a set of countable indices and $\{\varphi_\lambda\}_{\lambda \in \Lambda_n}$ a family of vectors in $L^2([0,1])$ such that for every $m \in \mathcal{M}_n$, $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$ denotes an orthonormal family of $L^2([0,1])$ with $\Lambda(m) \subset \Lambda_n$. For every $m \in \mathcal{M}_n$,

S_m denotes the linear space spanned by $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$, $D_m = \dim(S_m)$, and s_m is the *orthogonal projection* of s onto S_m

$$s_m := \mathrm{Argmin}_{u\in S_m} P\gamma(u) = \sum_{\lambda\in\Lambda(m)} P\varphi_\lambda \ \varphi_\lambda, \qquad \text{with} \qquad P\varphi_\lambda = \mathbb{E}\left[\,\varphi_\lambda(X)\,\right].$$

Definition 2.1. *An estimator $\widehat{s} \in L^2([0,1])$ is a projection estimator if there exists a family $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ of orthonormal vectors of $L^2([0,1])$ such that*

$$\widehat{s} = \sum_{\lambda\in\Lambda} \alpha_\lambda \ \varphi_\lambda, \qquad \text{with} \qquad \alpha_\lambda = \frac{1}{n}\sum_{i=1}^{n} H_\lambda(X_i),$$

where $\{H_\lambda(\cdot)\}_{\lambda\in\Lambda}$ depends on the family $\{\varphi_\lambda\}_{\lambda\in\Lambda}$.

As a consequence, it is straightforward to check that the empirical contrast minimizer defined by Eq. (2) over $S_m = \mathrm{Span}\,(\varphi_\lambda,\ \lambda\in\Lambda(m))$ is a projection estimator since

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda \ \varphi_\lambda \ , \qquad \text{with} \qquad P_n\varphi_\lambda = \frac{1}{n}\sum_{i=1}^{n} \varphi_\lambda(X_i) \ . \tag{3}$$

Here are a few examples of projection estimators (see DeVore and Lorentz, 1993):

- *Histograms*: For every $m \in \mathcal{M}_n$, let $\{I_\lambda\}_{\lambda\in\Lambda(m)}$ be a partition of $[0,1]$ in $D_m = \mathrm{Card}(\Lambda(m))$ intervals. Set $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ for every $\lambda \in \Lambda(m)$, with $|I_\lambda|$ the Lebesgue measure of I_λ, and $\mathbb{1}_{I_\lambda}(x) = 1$ if $x \in I_\lambda$ and 0 otherwise. Then,

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\mathbb{1}_{I_\lambda} \frac{\mathbb{1}_{I_\lambda}}{|I_\lambda|} \ . \tag{4}$$

- *Trigonometric polynomials*: For every $\lambda \in \mathcal{Z}$, let $\varphi_\lambda : \ t \mapsto \varphi_\lambda(t) = e^{2\pi i\lambda t}$. Then for any finite $\Lambda(m) \subset \mathbb{Z}$,

$$\widehat{s}_m(t) = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda \ e^{2\pi i\lambda t}, \qquad \forall t \in [0,1] \tag{5}$$

  is a trigonometric polynomial.
- *Wavelet basis*: Let $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ be an orthonormal basis of $L^2([0,1])$ made of compact supported wavelets, where $\Lambda_n = \left\{(j,k) \mid j \in \mathbb{N}^* \text{ and } 1 \le k \le 2^j\right\}$. Then for every subset $\Lambda(m)$ of Λ_n,

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda \ \varphi_\lambda \ . \tag{6}$$

Some of these estimators can take negative values for finite sample size. The same phenomenon arises with kernel estimators (Tsybakov, 2003). A possible solution to avoid negative values is truncating and normalizing the preliminary projection estimator

$$\widetilde{s}_m = \widehat{s}_m \mathbb{1}_{\widehat{s}_m \ge 0} \left(\int_{[0,1]} \mathbb{1}_{\widehat{s}_m \ge 0}(t)\widehat{s}_m(t)\,dt \right)^{-1} .$$

Note that if $s(x_0) > 0$ at a given $x_0 \in [0,1]$ and $\widehat{s}_m(x) \xrightarrow[n\to+\infty]{P} s(x)$ for every $x \in [0,1]$, then $\widehat{s}_m(x_0) \ge 0$ for large enough values of n.

2.2. Leave-p-out cross-validation

In the literature, several cross-validation (CV) algorithms have been successively introduced to overcome the defects of already existing ones. The purpose of the present section is to briefly describe the main CV algorithms that will be used throughout the paper with some emphasis to computational aspects.

2.2.1. Cross-validation

For every $1 \leq p \leq n-1$, let us define $\mathcal{E}_p = \{e \subset \{1,\dots,n\},\ \mathrm{Card}(e) = p\}$ and for any such $e \in \mathcal{E}_p$, set $X^e = \{X_i,\ i \in e\}$ (test set) and $X^{(e)} = \{X_i,\ i \in \{1,\dots,n\} \setminus e\}$ (training set). Let also $P_n^e := 1/p \sum_{i\in e} \delta_{X_i}$ and $P_n^{(e)} := 1/(n-p) \sum_{i\in(e)} \delta_{X_i}$ denote the empirical measures defined respectively from the test set X^e and the training set $X^{(e)}$.

Hold-out *Simple validation* also called *Hold-out* has been introduced at the early 30s (Larson, 1931). For every $1 \leq p \leq n-1$, it consists in randomly splitting observations into a training set $X^{(e)}$ of cardinality $n-p$ and a test set X^e of cardinality p. Random data splitting is only made once and introduces additional variability. For every $e \in \mathcal{E}_p$ (randomly chosen), the hold-out estimator of $R_n(\widehat{s})$ is

$$\widehat{R}_{\mathrm{Ho},p}(\widehat{s}) := P_n^e \gamma\left(\widehat{s}(X^{(e)})\right) = \frac{1}{p}\sum_{i\in e} \gamma\left(\widehat{s}(X^{(e)}); X_i\right) \quad . \tag{7}$$

Hold-out has been studied for instance by Bartlett et al. (2002); Blanchard and Massart (2006) in classification and by Lugosi and Nobel (1999); Wegkamp (2003) in regression.

Leave-p-out Unlike Eq. (7) where a single split e of the data is randomly chosen, which introduces additional unwanted variability, *leave-p-out* (Lpo) considers all the $\binom{n}{p} = \mathrm{Card}\left(\mathcal{E}_p\right)$ splits. The Lpo estimator of $R_n(\widehat{s})$ is defined by

$$\widehat{R}_p(\widehat{s}) = \binom{n}{p}^{-1} \sum_{e\in\mathcal{E}_p} P_n^e \gamma\left(\widehat{s}(X^{(e)})\right) \quad . \tag{8}$$

For instance, it has been studied by Shao (1993), Zhang (1993), and Arlot and Celisse (2011) in the regression framework. With $p = 1$, Lpo reduces to the celebrated *leave-one-out* (Loo) cross-validation introduced by Mosteller and Tukey (1968) and further studied by Stone (1974). Note that computing the Lpo estimator requires a computational complexity of order $\binom{n}{p}$ times that of computing $\widehat{s}$, which quickly becomes intractable as n grows.

V-fold cross-validation To overcome the high computational burden of Lpo (Eq. (8)), Geisser (1974, 1975) introduced the *V-fold cross-validation* (V-FCV). Instead of considering all the $\binom{n}{p}$ possible splits, one (randomly or not) chooses a partition of $X_1,\dots,X_n$ into V subsets $X^{e_1},\dots,X^{e_V}$ of approximately equal size $p = n/V = \mathrm{Card}(e_i)$, $i = 1,\dots,V$. Every X^{e_i}, $i = 1,\dots,V$ is successively used as a test set leading to the V-fold risk estimator of $R_n(\widehat{s})$

$$\widehat{R}_{\mathrm{V-FCV}}(\widehat{s}) = \frac{1}{V}\sum_{v=1}^{V} P_n^{e_v} \gamma\left(\widehat{s}(X^{(e_v)})\right) \quad . \tag{9}$$

V-FCV has been studied in the regression framework by Burman (1989, 1990) who suggests a correction to remove its bias.

2.2.2. Lpo versus V-FCV

As explained in Section 2.2.1, the Lpo computational complexity is roughly $\binom{n}{p}$ times that of computing $\widehat{s}$, which can be highly time-consuming. Unlike Lpo (and even Loo when $p = 1$), V-FCV involves only V such computations, which is less demanding as long as $V \ll n$. Note that usual values for V are 3, 5, and 10 (except $V = n$ where V-FCV and Loo coincide).

However, V-FCV relies on a preliminary (possibly random) partitioning of $X_1, \ldots, X_n$ into V subsets. Unlike Lpo where an exhaustive splitting is performed, this preliminary partitioning induces some additional variability, which could be misleading. For instance, Celisse and Robin (2008) have theoretically quantified the amount of additional variability induced by V-FCV with respect to Lpo.

On the one hand, Lpo can be seen as a "gold standard" among CV algorithms since it relies on exhaustive splitting and does not introduce any additional variability. On the other hand, V-FCV appears as an approximation to the "ideal Lpo" that cannot be achieved due to a prohibitive computational cost. Note that other approximations to Lpo have been proposed such as the *repeated learning-testing cross-validation* (Breiman et al., 1984; Burman, 1989; Zhang, 1993).

2.3. Closed-form expressions for the Lpo risk estimator

In Section 2.2.2 it is claimed that as long as Lpo cannot be computed V-FCV is preferable. Closed-form formulas for the Lpo estimator are proved in the present section, which makes Lpo fully effective in practice and always better than V-FCV. Besides, closed-form formulas also enable a more accurate theoretical analysis of CV algorithms both in terms of risk estimation (Section 2.4) and model selection (Section 3).

With the notation introduced at the beginning of Section 2.2.1, let us consider projection estimators $\widehat{s}_m$ defined by Eq. (3). Closed-form formulas for the Lpo risk estimator are derived exploiting the "linearity" of projection estimators. Sums over $\mathcal{E}_p$ (which cannot be computed in general) then reduce to binomial coefficients. Recalling the expression of the contrast $\gamma(\cdot\,;\cdot)$ (Eq. (1)), one has to compute both quadratic and linear terms.

Lemma 2.1. *For every $m \in \mathcal{M}_n$, let $\widehat{s}_m = \widehat{s}_m(X_{1,n})$ denote a projection estimator defined by Eq. (3) and set $X^e = \{X_i,\ i \in e\}$ for every $e \in \mathcal{E}_p$. Then for every $p \in \{1, \ldots, n-1\}$,*

$$\sum_{e \in \mathcal{E}_p} \left\| \widehat{s}_m(X^{(e)}) \right\|^2 = \frac{1}{(n-p)^2} \sum_{\lambda \in \Lambda(m)} \left[\binom{n-1}{p} \sum_{k=1}^{n} \varphi_\lambda^2(X_k) + \binom{n-2}{p} \sum_{k \neq \ell} \varphi_\lambda(X_k)\varphi_\lambda(X_\ell) \right] ,$$

$$\sum_{e \in \mathcal{E}_p} \sum_{i \in e} \widehat{s}\,(X^{(e)})(X_i) = \frac{1}{n-p} \sum_{\lambda \in \Lambda(m)} \binom{n-2}{p-1} \sum_{i \neq j} \varphi_\lambda(X_i)\varphi_\lambda(X_j) \ .$$

Proof of Lemma 2.1. For every $e \in \mathcal{E}_p$, and $t \in [0,1]$,

$$\widehat{s}_m(X^{(e)})(t) = \sum_\lambda (P_n^{(e)}\varphi_\lambda)\varphi_\lambda(t) = \frac{1}{n-p} \sum_{j=1}^{n} \sum_\lambda \varphi_\lambda(X_j)\varphi_\lambda(t) \mathbb{1}_{(j \in (e))} \ ,$$

which implies

$$\sum_{i \in e} \widehat{s}_m(X^{(e)})(X_i) = \frac{1}{n-p} \sum_{i \neq j} \sum_\lambda \varphi_\lambda(X_j)\varphi_\lambda(X_i) \mathbb{1}_{(j \in (e))} \mathbb{1}_{(i \in e)} \ .$$

It remains to sum over $e \in \mathcal{E}_p$, which is made thanks to Lemma A.1.

□

Lemma 2.1 enables to derive closed-form formulas for the Lpo risk estimator, which makes Lpo algorithm fully efficient in practice.

Proposition 2.1. *For every $m \in \mathcal{M}_n$, let $\widehat{s}_m = \widehat{s}_m(X_{1,n})$ denote a projection estimator defined by Eq. (3). Then for every $p \in \{1, \dots, n-1\}$,*

$$\widehat{R}_p(m) = \widehat{R}_p(\widehat{s}_m) = \frac{1}{n(n-p)} \sum_{\lambda \in \Lambda(m)} \left[\sum_{j=1}^{n} \varphi_\lambda^2(X_j) - \frac{n-p+1}{n-1} \sum_{j \neq k} \varphi_\lambda(X_j)\varphi_\lambda(X_k) \right] . \tag{10}$$

Proposition 2.1 enjoys a great interest. First it applies to the broad family of projection estimators. Second, it allows to reduce the computation time from an exponential to a linear complexity since computing (10) is of order $\mathcal{O}(n)$. Note that in the more specific setting of histograms and kernel estimators, such closed-form formulas have been derived by Celisse and Robin (2008).

Proof of Proposition 2.1. From definitions of the contrast (Eq. (1)) and the Lpo estimator Eq. (8), it comes

$$\widehat{R}_p(m) = \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_p} \left\| \widehat{s}_m(X_{1,n}^{(e)}) \right\|^2 - \frac{2}{p} \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_p} \sum_{i \in e} \widehat{s}_m(X_{1,n}^{(e)})(X_i) \ .$$

Then, Lemma 2.1 provides the expected conclusion. □

Let us now specify the Lpo estimator expressions for the three examples of projection estimators in Section 2.1.2.

1.

Corollary 2.1 (Histograms)**.** *For $\widehat{s}_m$ given by Eq. (4) and for $p \in \{1, \dots, n-1\}$,*

$$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{\lambda=1}^{D_m} \frac{1}{|I_\lambda|} \left[(2n-p) \frac{n_\lambda}{n} - n(n-p+1) \left(\frac{n_\lambda}{n} \right)^2 \right] ,$$

where $n_\lambda = \mathrm{Card}\left(\{ i \,|\, X_i \in I_\lambda \} \right)$.

2.

Corollary 2.2 (Trigonometric polynomials)**.** *For every $k \in \mathbb{N}$, let φ_λ denote either $t \mapsto \cos(2\pi k t)$, if $\lambda = 2k$ or $t \mapsto \sin(2\pi k t)$, if $\lambda = 2k+1$. Let us further assume $\Lambda(m) = \{0, \dots, 2K\}$ for $K \in \mathbb{N}^*$. Then for every $p \in \{1, \dots, n-1\}$,*

$$\widehat{R}_p(m) = \alpha(n,p) - \beta(n,p) \sum_{k=0}^{K} \left[\left\{ \sum_{j=1}^{n} \cos(2\pi k X_j) \right\}^2 + \left\{ \sum_{j=1}^{n} \sin(2\pi k X_j) \right\}^2 \right] ,$$

where $\alpha(n,p) = (p-2)(K+1)\left[(n-1)(n-p) \right]^{-1}$ and $\beta(n,p) = (n-p+1)\left[n(n-1)(n-p) \right]^{-1}$.

3.

Corollary 2.3 (Haar basis)**.** *Let us define* $\varphi : t \mapsto \mathbb{1}_{[0,1]}$ *and* $\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j \cdot -k)$*, where* $j \in \mathbb{N}$ *and* $0 \leq k \leq 2^j - 1$*, and assume* $\Lambda(m) \subset \left\{(j,k) \mid j \in \mathbb{N},\ 0 \leq k \leq 2^j - 1\right\}$ *for every* $m \in \mathcal{M}_n$*. Then,*

$$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{(j,k)\in\Lambda(m)} 2^j \left[(2n-p)\frac{n_{j,k}}{n} - n(n-p+1)\left(\frac{n_{j,k}}{n}\right)^2 \right],$$

where $n_{j,k} = \mathrm{Card}\left(\left\{i \mid X_i \in [k/2^j, (k+1)/2^j]\right\}\right)$.

2.4. Risk estimation: Leave-one-out optimality

From the general closed-form formula given by Eq. (10), one derives closed-form expressions for the expectation and variance of the Lpo risk as well. These expressions will be useful to analyze the theoretical behavior of CV in terms of risk estimation and model selection (see Section 3). In the present section for instance, they are used to prove the optimality of Loo for estimating the risk of any projection estimator (Theorem 2.1).

Proposition 2.2. *For every* $m \in \mathcal{M}_n$*, let* $\widehat{s}_m = \widehat{s}_m(X_{1,n})$ *denote a projection estimator defined by Eq.* (3)*. Then for every* $1 \leq p \leq n-1$*,*

$$\mathbb{E}\left[\widehat{R}_p(m)\right] = \frac{1}{n-p} \sum_{\lambda\in\Lambda(m)} \left[\mathbb{E}\varphi_\lambda^2(X) - \left(\mathbb{E}\varphi_\lambda(X)\right)^2\right] - \sum_{\lambda\in\Lambda(m)} \left(\mathbb{E}\varphi_\lambda(X)\right)^2 \ ,$$

and

$$\mathrm{Var}\left[\widehat{R}_p(m)\right] = \frac{1}{(n-1)^2}\left[a_n + \frac{b_n}{(n-p)} + \frac{c_n}{(n-p)^2}\right] \ , \tag{11}$$

where $a_n = \mathrm{Var}\left[\sum_{\lambda\in\Lambda(m)} \left(n(P_n\varphi_\lambda)^2 - P_n\varphi_\lambda^2\right)\right]$ *,* $c_n = \mathrm{Var}\left[n\sum_{\lambda\in\Lambda(m)} \left(P_n\varphi_\lambda^2 - (P_n\varphi_\lambda)^2\right)\right]$ *, and* $b_n = -2\ \mathrm{Cov}\left[\sum_{\lambda\in\Lambda(m)} \left(n(P_n\varphi_\lambda)^2 - P_n\varphi_\lambda^2\right), \sum_{\lambda\in\Lambda(m)} n\left(P_n\varphi_\lambda^2 - (P_n\varphi_\lambda)^2\right)\right]$.

The proof is a straightforward application of Proposition 2.1 and has been omitted. Note that the above quantities do exist as long as $P\left|\varphi_\lambda\right|^3 < +\infty$ for any $\lambda \in \Lambda(m)$, which holds true if s is bounded for instance and $\int \left|\varphi_\lambda\right|^3 < +\infty$ (φ_λ continuous and compact supported for instance). In the variance expression, a_n, b_n, and c_n do not depend on p. Then knowing the behavior of the variance with respect to p only depends on the magnitude of a_n, b_n, and c_n, which is clarified by Corollary 2.5.

Let us first focus on the bias $\mathbb{B}\left[\widehat{R}_p(m)\right] := \mathbb{E}\widehat{R}_p(m) - \mathbb{E}\left[\left\|\widehat{s}_m\right\|^2 - 2\int_{[0,1]} s\,\widehat{s}_m\right]$ of the Lpo estimator.

Corollary 2.4 (Bias)**.** *For every* $m \in \mathcal{M}_n$*, let* $\widehat{s}_m = \widehat{s}_m(X_{1,n})$ *denote a projection estimator defined by Eq.* (3)*. Then for every* $m \in \mathcal{M}_n$ *and* $1 \leq p \leq n-1$*,*

$$\mathbb{B}\left[\widehat{R}_p(m)\right] = \frac{p}{n(n-p)} \sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left[\,\varphi_\lambda(X_1)\,\right] \geq 0 \ .$$

The bias is nonnegative and increases with p, which means Loo ($p = 1$) has the smallest bias among CV algorithms. Besides if $p = p_n$ satisfies $p_n/n \xrightarrow[n\to+\infty]{} q \in [0,1)$, then $\mathbb{B}\left[\widehat{R}_p(m)\right] \xrightarrow[n\to+\infty]{} 0$. Thus, Loo is asymptotically unbiased.

Let us now describe the behavior of the variance with respect to p.

Corollary 2.5 (Variance)**.** *With the same notation as Proposition 2.2, for every $m \in \mathcal{M}_n$ and $1 \leq p \leq n-1$,*

$$\mathrm{Var}\left[\widehat{R}_p(m)\right] = \frac{n}{(n-1)^2}\left[A + \frac{B}{n-p} + \frac{C}{(n-p)^2} + O\left(\frac{1}{n}\right)\right] \ ,$$

where

$$A = 4\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \geq 0 \ ,$$

$$B = 8\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] - 4\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \ ,$$

$$C = 4\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] - 4\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]$$
$$+ \mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right] \geq 0 \ .$$

In the more specific case of histogram and kernel density estimators, Celisse and Robin (2008) derived a similar (non asymptotic) result for the variance. Note that the monotonicity of the variance with respect to p depends on the sign of B since $x \mapsto f(x) = Ax^2 + Bx + C$ has for derivative $x \mapsto f'(x) = 2Ax + B$ and $A \geq 0$. However in full generality, the sign of B is unknown.

Proof of Corollary 2.5. Combining Proposition 2.2, Lemmas A.2 and A.3, and Proposition A.1, it comes

$$a_n = 4n\beta + O(1) \ ,$$
$$b_n = 8n\beta - 4n\gamma + O(1) \ ,$$
$$c_n = 4n\beta - 4n\gamma + n\delta + O(1) \ ,$$

where $\beta = \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]$, $\gamma = \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]$, and $\delta = \mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right]$. This provides the expected conclusion with $A = 4\beta$, $B = 8\beta - 4\gamma$, and $C = 4\beta - 4\gamma + \delta$. □

The purpose of the following proposition is to describe the monotonicity of the variance depending on the sign of B

Proposition 2.3. *Let us define $p_{0,n} = \mathrm{Argmin}_{1\leq p\leq n-1}\mathrm{Var}\left[\widehat{R}_p(m)\right]$ in Eq. (11). Then,*

$$p_{0,n} = n + \left(1 - \frac{\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]}{2\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]}\right)(1 + o(1)) \ .$$

Furthermore,

1. *if*

$$2\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \geq \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \ , \quad (12)$$

$p \in \{1, \ldots, n-1\} \mapsto \mathrm{Var}\left[\widehat{R}_p(m)\right]$ is increasing.

2. Otherwise, $p \mapsto \mathrm{Var}\left[\widehat{R}_p(m)\right]$ *is decreasing on* $[1, p_{0,n}]$ *and increasing on* $[p_{0,n}, n-1]$.

Eq. (12) is related to the sign of B in Corollary 2.5 and to the minimum location value $p_{0,n}$. In particular if it holds true, then $p_{0,n} \notin \{2, \ldots, n-1\}$, which means Loo has the smallest variance among CV algorithms.

Theorem 2.1. *For every* $m \in \mathcal{M}_n$, *let us define the mean-square error* (MSE) *of* $\widehat{s}_m$ *by* $\mathrm{MSE}(m;p) = \left(\mathbb{B}\left[\widehat{R}_p(m)\right]\right)^2 + \mathrm{Var}\left[\widehat{R}_p(m)\right]$, *for every* $p \in \{1, \ldots, n-1\}$.

1. *If* (12) *holds true, then for every* $m \in \mathcal{M}_n$, $p \mapsto \mathrm{MSE}(m;p)$ *is minimum for* $p = 1$.
2. *Otherwise, for every* $p = p_n \in \{1, \ldots, n-1\}$ *such that* $p_n/n \xrightarrow[n\to+\infty]{} q \in [0,1)$, *then*

$$\mathrm{MSE}(m;p) = \frac{A}{n} + O\left(\frac{1}{n^2}\right), \qquad \text{as } n \to +\infty \ .$$

If (12) holds true, Loo is the best CV algorithm in terms of MSE when estimating the risk of an estimator. Otherwise as long as $p_n/n \not\to 1$ as $n \to +\infty$, choosing a value of $p \neq 1$ is useless since any value in $\{1, \ldots, n-1\}$ asymptotically leads to the same performance in terms of MSE. But since Loo has a minimum bias (Corollary 2.4), one concludes *Loo is optimal among CV algorithms for estimating the risk of an estimator*. This result confirms what has been previously stated by Burman (1989) in the regression framework.

3. Optimal cross-validation for model selection

In Section 2.4, the optimality of Loo among CV algorithms has been proved in the context of risk estimation. However, the best possible algorithm for risk estimation is not necessarily the best one for *model selection*. For instance, empirical contrast minimization (2) is used to design an estimator $\widehat{s}_m \in S_m$. But using empirical contrast minimization to choose one $\widehat{m} \in \mathcal{M}_n$ (without penalizing) would systematically lead to overfitting. The purpose of the present section is to study the performance of CV for model selection with respect to the cardinality p of the test set.

In model selection, two (contradictory) purposes can be pursued: *Estimation* and *Identification* (see Shao, 1997; Yang, 2005, for an extensive presentation). With the Estimation point of view, one focuses on minimizing the risk over a collection of models without assuming the targeted s belongs to one of them. Conversely in Identification, one assumes s belongs to at least one model of the collection and the goal is to recover the smallest model containing s.

3.1. Optimal cross-validation for Estimation

Model selection by CV pursuing Estimation is our main concern here. First, the performance of CV with respect to p is characterized through a *sharp oracle inequality* (Theorem 3.1) where constants are related to the difficulty of the estimation problem. In particular, a leading constant converging to 1 as $n \to +\infty$ is achieved for given values of p. Second, Loo is theoretically shown to be suboptimal for model selection (Corollary 3.1), which is also empirically supported by simulation experiments (Section 3.1.4).

3.1.1. Estimation point of view

With the notation of Section 2.1, let us consider a family of projection estimators $\{\widehat{s}_m\}_{m\in\mathcal{M}_n}$, where $\mathcal{M}_n$ denotes an (at most countable) index set allowed to depend on n. The best possible model, called *the*

oracle model, is denoted by S_{m^*}, where

$$m^* := \mathrm{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m) - P\gamma(s) = \mathrm{Argmin}_{m\in\mathcal{M}_n} \left\|s-\widehat{s}_m\right\|^2$$
$$= \mathrm{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m) \ .$$

Since $P\gamma(\widehat{s}_m)$ has to be estimated, one uses CV (Lpo) to choose a candidate model. So for every $1 \leq p \leq n-1$,

$$\widehat{m}(p) := \mathrm{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(m) \ , \tag{13}$$

and the final candidate model is denoted by $S_{\widehat{m}(p)}$. The purpose is now to study the properties of $\widehat{s}_{\widehat{m}(p)}$ with respect to $p \in \{1,\dots,n-1\}$ in terms of an oracle inequality, that is an inequality such that an event of large probability exists on which

$$\left\|s-\widehat{s}_{\widehat{m}}\right\|^2 \leq C_n \inf_{m\in\mathcal{M}_n} \left\{\left\|s-\widehat{s}_m\right\|^2\right\} + r_n \ , \tag{14}$$

where $\widehat{s}_{\widehat{m}}$ is the final estimator provided by the considered model selection procedure, the constant $C_n \geq 1$ does not depend on s, and r_n is a remainder term. When $C_n \xrightarrow[n\to+\infty]{} 1$ on an event of probability larger than $1-K/n^2$ (for some $K>0$), the model selection procedure is said *efficient* (Arlot and Celisse, 2010).

3.1.2. Main oracle inequality

Let us first introduce some notation and detail the main assumptions used along the following sections.

Square-integrable density:

$$s \in L^2([0,1]) \ . \tag{**SqI**}$$

Unlike Castellan (2003) for instance, it is not assumed that $s \geq \rho$ for a constant $\rho > 0$.

Polynomial collection: There exists $a_{\mathcal{M}} \geq 0$ such that

$$\mathrm{Card}(\mathcal{M}_n) \leq n^{a_{\mathcal{M}}} \ . \tag{**Pol**}$$

In particular, this holds true if there exists $\alpha \geq 0$ such that $\mathrm{Card}\left(\{m\in\mathcal{M}_n,\ D_m = D\}\right) \leq D^\alpha$, for every $1 \leq D \leq n$.

Model regularity:

$$\exists \Phi > 0, \quad \sup_{m\in\mathcal{M}_n} \frac{\left\|\phi_m\right\|_\infty}{D_m} \leq \Phi \ , \quad \text{with} \quad \phi_m = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 \ . \tag{**RegD**}$$

It relates the regularity of the orthonormal basis (measured in terms of sup-norm) to the dimension of the model. For instance using (4), (**RegD**) requires $|I_\lambda| \geq (\Phi D_m)^{-1}$ for every $\lambda \in \Lambda(m)$. Thus, the length of intervals I_λ cannot be too different from one another.

Maximal dimension:

$$\exists \Gamma > 0, \quad \sup_{m\in\mathcal{M}_n} D_m \leq \Gamma \frac{n}{(\log n)^2} \ . \tag{**Dmax**}$$

In the sequel, we always use $\Gamma = 1$ to simplify the notation. Note that proofs and conclusions are not changed by this particular choice.

Estimation error and dimension:

$$\exists \xi > 0, \quad \inf_{m \in \mathcal{M}_n} \frac{\sqrt{n}\mathbb{E}\left(\|s_m - \widehat{s}_m\|\right)}{\sqrt{D_m}} \geq \sqrt{\xi} \ . \tag{\textbf{LoEx}}$$

This assumption makes the estimation error $\mathbb{E}\left(\|s_m - \widehat{s}_m\|^2\right)$ and D_m comparable. For instance, (**LoEx**) is fulfilled if $s \geq \rho > 0$.

Richness of the collection: There exist $m_0 \in \mathcal{M}_n$ and $c_{rich} \geq 1$ such that,

$$\sqrt{n} \leq D_{m_0} \leq c_{rich}\sqrt{n} \ . \tag{\textbf{Rich}}$$

Such an assumption only depends on our choice of model collection and can always be fulfilled.

Approximation property: There exist $c_\ell, c_u > 0$ and $\ell > u > 0$ such that, for every $m \in \mathcal{M}_n$,

$$c_\ell D_m^{-\ell} \leq \|s - s_m\|^2 \leq c_u D_m^{-u} \ . \tag{\textbf{Bias}}$$

This assumption quantifies the bias (approximation error) incurred by model S_m in estimating s. It therefore relies on a smoothness assumption on s. Such an upper bound is classical for α-Hölderian functions with $\alpha \in (0,1]$ and regular histograms (4) for instance. Note that Stone (1985) uses the same assumption (lower bound), which is the *finite sample counterpart* of the classical assumption $\|s - s_m\| > 0$ for every $m \in \mathcal{M}_n$ usually made to prove *asymptotic optimality* for a model selection procedure (see Birgé and Massart, 2006).

Rate of convergence for the oracle model:

$$nR_n^*(\log n)^{-2} \xrightarrow[n\to\infty]{} +\infty, \quad \text{with} \quad R_n^* = \inf_{m \in \mathcal{M}_n} R_n(\widehat{s}_m) \ , \tag{\textbf{OrSp}}$$

This assumption implies the risk of the oracle model R_n^* does not decrease to 0 faster than $(\log n)^2/n$. In particular, this holds true for densities in $\mathcal{H}(L,\alpha)$ with $L > 0$ and $\alpha \in (0,1]$ for instance (see Section A.6).

The performance of the Lpo estimator with respect to p is described by the following oracle inequality where the leading constant $C_n(p)$ relates the complexity of the collection of models $\{S_m\}_{m\in\mathcal{M}_n}$ to p.

Theorem 3.1 (Optimal CV)**.** *Let s denote a density on $[0,1]$ such that* (**SqI**) *holds true, set $\{S_m\}_{m\in\mathcal{M}_n}$ a collection of models defined in Section 2.1.2, and assume* (**Pol**), (**RegD**), (**Dmax**), (**Rich**), (**LoEx**), (**Bias**)*, and* (**OrSp**)*. Let $\widehat{m} = \widehat{m}(p)$ denote the model minimizing $\widehat{R}_p(m)$ over $\mathcal{M}_n$ for every $p \in \{1,\dots,n-1\}$. Then, there exist a sequence $(\delta_n)_\mathbb{N}$ such that $\delta_n \to +\infty$, and $n\delta_n \to +\infty$ as $n \to +\infty$, and an event $\widetilde{\Omega}$ with $\mathbb{P}(\widetilde{\Omega}) \geq 1 - 6/n^2$ on which, for large enough values of n,*

$$\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \leq C_n(p) \inf_{m\in\mathcal{M}_n} \left\{\|s - \widehat{s}_m\|^2\right\} \qquad \text{with} \qquad C_n(p) = \frac{T_B^+ \vee T_V^+}{T_B^- \wedge T_V^-} \geq 1 \ ,$$

where

$$T_B^- = 1 - \delta_n K(n,p) \ , \qquad T_V^- = \frac{1}{1-p/n}(1-\delta_n)\left[1 - 4\delta_n\right] - 2\delta_n K(n,p)\left[3 - 4\delta_n\right] \ ,$$

$$T_B^+ = 1 + \delta_n K(n,p) \ , \qquad T_V^+ = \frac{1}{1-p/n}(1+\delta_n)\left[1 + 4\delta_n\right] + 2\delta_n K(n,p)\left[3 + 4\delta_n\right] \ ,$$

and $K(n,p) = 1 + \frac{2}{n-1} + \frac{p}{n-p}\frac{1}{n-1}$.

First if $p = p_n = o(n)$, then $p/n \to 0$ and $C_n(p) \to 1$ as $n \to +\infty$. Then, such values of p lead to *efficient* (asymptotically optimal) model selection procedures. In particular, this holds true for $p = 1$, that is, *Loo is asymptotically optimal* since

$$\frac{\left\| s - \widehat{s}_{\widehat{m}(1)} \right\|^2}{\inf_{m \in \mathcal{M}_n} \left\{ \left\| s - \widehat{s}_m \right\|^2 \right\}} \xrightarrow[n \to +\infty]{a.s.} 1 \ .$$

Second, $C_n(p)$ can be optimized as a function of p at each finite sample size n. Since $C_n(p)$ also depends on δ_n, which is related to the structure of $\{S_m\}_{m \in \mathcal{M}_n}$ and the probability of the event $\widetilde{\Omega}$, minimizing $C_n(p)$ with respect to p enables to take into account the difficulty of the estimation problem at hand.

Proof of Theorem 3.1.
First let us use Proposition A.2 applied with $m, m' \in \mathcal{M}_n$ such that $\widehat{R}_p(m') \leq \widehat{R}_p(m)$. Then, it comes

$$\begin{aligned}
&\frac{n}{n-p}\mathbb{E}\left[Z_{m'}^2 \right] + \left\| s - s_{m'} \right\|^2 - K(n,p) \left[Z_{m'}^2 - \mathbb{E}\left[Z_{m'}^2 \right] \right] \\
&\leq \frac{n}{n-p}\mathbb{E}\left[Z_m^2 \right] + \left\| s - s_m \right\|^2 - K(n,p) \left[Z_m^2 - \mathbb{E}\left[Z_m^2 \right] \right] \\
&- 2K(n,p)\, \nu_n \left(s_{m'} - s_m \right) + \frac{1}{n} \left(K(n,p) + \frac{n}{n-p} \right) \nu_n \left(\phi_{m'} - \phi_m \right) \ ,
\end{aligned}$$

where $K(n,p) = 1 + \frac{2}{n-1} + \frac{p}{n-p}\frac{1}{n-1}$.

Then, combining Propositions A.3 and A.4 to control the remainder terms, there exist a sequence $(\delta_n)_{\mathbb{N}}$ with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as $n \to +\infty$ and an event $\Omega = \Omega_{\text{rem},1} \cap \Omega_{\text{rem},2}$ of probability $1 - 4/n^2$ on which

$$\begin{aligned}
&\frac{n}{n-p}\mathbb{E}\left[Z_{m'}^2 \right] + \left\| s - s_{m'} \right\|^2 - K(n,p) \left[Z_{m'}^2 - \mathbb{E}\left[Z_{m'}^2 \right] \right] \\
&\leq \frac{n}{n-p}\mathbb{E}\left[Z_m^2 \right] + \left\| s - s_m \right\|^2 - K(n,p) \left[Z_m^2 - \mathbb{E}\left[Z_m^2 \right] \right] \\
&+ \delta_n K(n,p) \left(\left\| s - s_{m'} \right\|^2 + \mathbb{E}\left[Z_{m'}^2 \right] + \left\| s - s_m \right\|^2 + \mathbb{E}\left[Z_m^2 \right] \right) \\
&+ \delta_n \left(K(n,p) + \frac{n}{n-p} \right) \left[\mathbb{E}\left[Z_{m'}^2 \right] + \mathbb{E}\left[Z_m^2 \right] \right] \ .
\end{aligned}$$

In the following, δ_n always denotes such a sequence even if the precise expression of δ_n can differ from line to line.

Let us now use concentration results stated in Corollaries A.1 and A.2 on the events Ω_{left} and Ω_{right}. The important point in this proof is given by Lemmas A.4 and A.5, where it is proved that on the event $\Omega = \Omega_{\text{left}} \cap \Omega_{\text{right}} \cap \Omega_{\text{rem},1} \cap \Omega_{\text{rem},1}$, $\min\left\{ D_{m^*}, D_{\widehat{m}(p)} \right\} \geq (\log n)^4$ for large enough values of n. Therefore, one can apply Lemma A.8 and Corollaries A.1 and A.2 with $L_m = 0 = r_n(m)$ to get

$$\begin{aligned}
&Z_{m'}^2 \left[\left(\frac{n}{n-p}(1-\delta_n) - 2\delta_n K(n,p) \right)(1 - 4\delta_n) - 4K(n,p)\delta_n \right] + \left[1 - \delta_n K(n,p) \right] \left\| s - s_{m'} \right\|^2 \\
&\leq Z_m^2 \left[\left(\frac{n}{n-p}(1+\delta_n) + 2\delta_n K(n,p) \right)(1 + 4\delta_n) + 4K(n,p)\delta_n \right] + \left[1 + \delta_n K(n,p) \right] \left\| s - s_m \right\|^2 \ .
\end{aligned}$$

Choosing $m' = \widehat{m}$, it comes

$$T_V^- Z_{\widehat{m}}^2 + T_B^- \left\| s - s_{\widehat{m}} \right\|^2 \leq T_V^+ Z_m^2 + T_B^+ \left\| s - s_m \right\|^2 \ ,$$

where

$$T_B^- = 1 - \delta_n K(n,p) \ , \qquad T_V^- = \frac{n}{n-p}(1-\delta_n)\left[1-4\delta_n\right] - 2K(n,p)\left[3\delta_n - 4\delta_n^2\right] \ ,$$
$$T_B^+ = 1 + \delta_n K(n,p) \ , \qquad T_V^+ = \frac{n}{n-p}(1+\delta_n)\left[1+4\delta_n\right] + 2K(n,p)\left[3\delta_n + 4\delta_n^2\right] \ .$$

Finally on the event Ω, the following oracle inequality holds true for every $p \in \{1, n-1\}$

$$\left\| s - \widehat{s}_{\widehat{m}(p)} \right\|^2 \leq C_n(p) \inf_{m \in \mathcal{M}_n} \left\{ \left\| s - \widehat{s}_m \right\|^2 \right\} \ , \qquad \text{with} \qquad C_n(p) = \frac{T_B^+ \vee T_V^+}{T_B^- \wedge T_V^-} \ .$$

Moreover, on the event Ω, Lemmas A.4 and A.5 show $\min\left\{D_{m^*}, D_{\widehat{m}(p)}\right\} \geq (\log n)^4$. Then, it is enough to apply Propositions A.6 and A.5 to models satisfying this constraint, which leads to the new event $\widetilde{\Omega}$ (where models with dimension smaller than $(\log n)^4$ have been omitted) of probability at least $1 - 6/n^2$. □

While *asymptotic optimality* is proved in Theorem 3.1 for any CV procedure as long as $p = o(n)$, it is also desirable to analyze the performance of CV for finite samples. Minimizing $C_n(p)$ as a function of p for each n provides the value $p^* = p_n^*$ for which $\widehat{m}(p^*)$ reaches the best performance among $\{\widehat{m}(p),\ 1 \leq p \leq n-1\}$. The following Corollary 3.1 proves Loo is suboptimal in terms of rate of convergence, which can lead to overfitted models.

Corollary 3.1 (Suboptimality of Leave-one-out)**.** *With the notation and assumptions of Theorem 3.1, the constant $C_n(p)$ is minimized over $p \in \{1, \ldots, n-1\}$ for*

$$0 < q_n^* := \frac{p_n^*}{n} = 1 - \frac{1 - 5\delta_n + 4\delta_n^2 - \frac{2}{n-1}(3\delta_n - 4\delta_n^2) + \frac{\delta_n}{n-1}}{1 + 2(1+\frac{1}{n-1})(3\delta_n - 4\delta_n^2) - \delta_n(1+\frac{1}{n-1})} < 1 \ \cdot$$

Furthermore, the optimal ratio $q_n^* = p^*/n$ is slowly decreasing to 0 as n tends to $+\infty$

$$q_n^* \sim_{+\infty} 10\delta_n \ , \qquad \text{and} \qquad p_n^* \sim_{+\infty} 10n\delta_n \xrightarrow[n\to+\infty]{} +\infty \ .$$

In particular, Loo ($p = 1$) is suboptimal in terms of rate of convergence with respect to n.

Whereas Theorem 3.1 settles Loo (and any CV algorithm with $p = o(n)$) is asymptotically optimal, Corollary 3.1 proves it is nevertheless suboptimal in terms of rate of convergence. Indeed, the optimal rate is achieved when p_n/n is slowly decreasing to 0 like δ_n as n grows. Let us also recall that δ_n is strongly related to the structure of the model collection, so that the more complex the collection, the slower δ_n, and the larger p_n should be to balance overfitting arising with too large models. As a consequence, Loo ($p = 1$) does not adapt to the model collection $\{S_m\}_{m \in \mathcal{M}_n}$, which results in overfitting, that is, choosing too large models (see simulation experiments in Section 3.1.4).

Proof of Corollary 3.1. Let us recall the expression of the leading constant

$$C_n(p) = \frac{T_B^+ \vee T_V^+}{T_B^- \wedge T_V^-} \ ,$$

with

$$T_B^- = 1 - \delta_n K(n,p) \ , \qquad T_V^- = \frac{1}{1-p/n}(1-\delta_n)\left[1-4\delta_n\right] - 2\delta_n K(n,p)\left[3 - 4\delta_n\right] \ ,$$
$$T_B^+ = 1 + \delta_n K(n,p) \ , \qquad T_V^+ = \frac{1}{1-p/n}(1+\delta_n)\left[1+4\delta_n\right] + 2\delta_n K(n,p)\left[3 + 4\delta_n\right] \ ,$$

and $K(n,p) = 1 + \frac{2}{n-1} + \frac{p}{n-p}\frac{1}{n-1}$.

First as long as n is large enough, simple calculations when $p = 1$ show $T_V^-(1) \leq T_B^-(1)$. Noticing moreover that $T_V^+(p) \geq T_B^+(p)$ for every p, it comes for p close to 1

$$C_n(p) = \frac{T_V^+}{T_V^-} = \frac{(1+\delta_n)\left[1+4\delta_n\right] + 2(1-p/n)\delta_n K(n,p)\left[3+4\delta_n\right]}{(1-\delta_n)\left[1-4\delta_n\right] - 2(1-p/n)\delta_n K(n,p)\left[3-4\delta_n\right]} \ .$$

It is then easy to show that $p \mapsto C_n(p)$ is increasing on $\{1,\ldots,p^*\}$, where p^* denotes the value of p such that $T_V^-(p) = T_B^-(p)$. Hence,

$$\frac{p_n^*}{n} = 1 - \frac{1 - 5\delta_n + 4\delta_n^2 - \frac{2}{n-1}(3\delta_n - 4\delta_n^2) + \frac{\delta_n}{n-1}}{1 + 2(1+\frac{1}{n-1})(3\delta_n - 4\delta_n^2) - \delta_n(1+\frac{1}{n-1})} \ .$$

It results that for every $p \geq p^*$

$$C_n(p) = \frac{T_V^+}{T_B^-} \ ,$$

which is increasing with respect to p.

In the same way, it is easy to check that $p_n^*/(10n\delta_n) \xrightarrow[n\to+\infty]{} 1$, which enables to conclude.

□

3.1.3. Adaptivity in the minimax sense

Adaptivity in the minimax sense is a desirable property for model selection procedures. It means the considered procedure automatically adapts to the unknown smoothness of the target function s to estimate (see Barron et al., 1999, for an extensive presentation).

Several adaptivity in the minimax sense results are provided in the present section. Deriving such results from oracle inequalities such as (14) is somewhat classical. However, the novelty is first that CV as model selection procedure enjoys such a desirable property, second that the leading constant $C_n(p)$ in Theorem 3.1 when converging to 1 as n tends to $+\infty$ provides accurate results.

Let us first provide a general result from which all adaptivity results will be immediate corollaries.

Theorem 3.2. *Let s denote a density on $[0,1]$ such that* (**SqI**) *holds true, set $\{S_m\}_{m\in\mathcal{M}_n}$ a collection of models defined in Section 2.1.2, and assume* (**Pol**), (**RegD**), (**Dmax**), (**Rich**), (**LoEx**), (**Bias**), *and* (**OrSp**). *Let $\widehat{m} = \widehat{m}(p)$ denote the model minimizing $\widehat{R}_p(m)$ over $\mathcal{M}_n$ for every $p \in \{1,\ldots,n-1\}$. Then for every $1 \leq p \leq n-1$,*

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right] \leq C_n(p)\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\|s - \widehat{s}_m\right\|^2\right] + (\Phi + \|s\|^2)\frac{12}{n(\log n)^2} + \frac{6c_u}{n^2} \ , \tag{15}$$

where $C_n(p) = \frac{T_B^+ \vee T_V^+}{T_B^- \wedge T_V^-}$, with

$$T_B^- = 1 - \delta_n K(n,p) \ , \qquad T_V^- = \frac{1}{1-p/n}(1-\delta_n)\left[1-4\delta_n\right] - 2\delta_n K(n,p)\left[3-4\delta_n\right] \ ,$$

$$T_B^+ = 1 + \delta_n K(n,p) \ , \qquad T_V^+ = \frac{1}{1-p/n}(1+\delta_n)\left[1+4\delta_n\right] + 2\delta_n K(n,p)\left[3+4\delta_n\right] \ ,$$

and $K(n,p) = 1 + \frac{2}{n-1} + \frac{p}{n-p}\frac{1}{n-1}$.

The last two terms in the right-hand side of (15) are remainder terms. They results from Assumptions (**RegD**), (**Dmax**), and (**Bias**). From remarks following Theorem 3.1, one deduces $p = p_n = o(n)$ implies $C_n(p) \to 1$ as $n \to +\infty$ and

$$\frac{\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right]}{\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\|s - \widehat{s}_m\right\|^2\right]} \xrightarrow[n\to+\infty]{} 1 \ .$$

Proof of Theorem 3.2. Introducing the event $\widetilde{\Omega}$ of Theorem 3.1, it comes

$$\begin{aligned}\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right] &= \mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}}\right] + \mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] \\ &\leq \mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\|s - \widehat{s}_m\right\|^2\right] + \mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] \ .\end{aligned}$$

Applying (**Bias**), one gets

$$\mathbb{E}\left[\left\|s - s_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] \leq \mathbb{E}\left[\frac{c_u}{D^u_{\widehat{m}(p)}}\mathbb{1}_{\widetilde{\Omega}^c}\right] \leq c_u\mathbb{P}\left(\widetilde{\Omega}\right) \leq \frac{6c_u}{n^2} \ ,$$

and (**RegD**) and (**Dmax**) provide

$$\begin{aligned}\mathbb{E}\left[\left\|s_{\widehat{m}(p)} - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] &= \mathbb{E}\left[\sum_{\lambda\in\Lambda(\widehat{m}(p))}\left(P_n\varphi_\lambda - P\varphi_\lambda\right)^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] \\ &\leq 2\mathbb{E}\left[\sum_{\lambda\in\Lambda(\widehat{m}(p))}\left(P_n\varphi_\lambda\right)^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] + 2\mathbb{E}\left[\sum_{\lambda\in\Lambda(\widehat{m}(p))}\left(P\varphi_\lambda\right)^2 \mathbb{1}_{\widetilde{\Omega}^c}\right] \\ &\leq 2\mathbb{E}\left[\sum_{\lambda\in\Lambda(\widehat{m}(p))}\frac{1}{n^2}\sum_{i,j=1}^{n}\varphi_\lambda(X_i)\varphi_\lambda(X_j)\mathbb{1}_{\widetilde{\Omega}^c}\right] + 2\left\|s\right\|^2\mathbb{E}\left[D_{\widehat{m}(p)}\mathbb{1}_{\widetilde{\Omega}^c}\right] \\ &\leq 2(\Phi + \left\|s\right\|^2)\frac{n}{(\log n)^2}\mathbb{P}\left(\widetilde{\Omega}^c\right) \leq (\Phi + \left\|s\right\|^2)\frac{12}{n(\log n)^2} \ .\end{aligned}$$

□

Applying Theorem 3.2 to the collection of regular histograms defined by (4), the following corollary settles an adaptivity property with respect to Hölder balls (see DeVore and Lorentz, 1993).

Corollary 3.2. *Let us consider the model collection of Section 2.1.2 made of piecewise constant functions and the associated histograms defined by* (4) *such that, for every* $m \in \mathcal{M}_n$ *and* $\lambda \in \Lambda(m)$, $|I_\lambda| = D_m^{-1}$ *(regular histograms). Let us also assume* (**Dmax**) *and* (**LoEx**) *hold true.*
If the target density s *belongs to the Hölder ball* $\mathcal{H}(L,\alpha)$ *for some* $L > 0$ *and* $\alpha \in (0,1]$*, then for every* $1 \leq p \leq n-1$ *there exist constants* $0 < K_\alpha^- \leq K_\alpha^+$ *such that*

$$K_\alpha^- L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \leq \sup_{s\in\mathcal{H}(L,\alpha)}\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right] \leq C_n(p)K_\alpha^+ L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} + O\left(\frac{1}{n(\log n)^2}\right) \ ,$$

K_α^- *and* K_α^+ *only depend on* α *(not on* n *or* s*).*
Furthermore since this property holds for every $L > 0$ *and* $\alpha \in (0,1]$*, then* $\left\{\widehat{s}_{\widehat{m}(p)}\right\}_{n\in\mathbb{N}^*}$ *is adaptive in the minimax sense with respect to* $\{\mathcal{H}(L,\alpha)\}_{L>0,\alpha\in(0,1]}$ *for every* $1 \leq p \leq n-1$.

The proof has been deferred to Section A.6. The lower bound is not new and has been proved earlier by Ibragimov and Khas'minskij (1981). Besides, the upper bound is tight since the rate $n^{-\frac{2\alpha}{2\alpha+1}}$ and the dependence on the radius $L^{\frac{2}{2\alpha+1}}$ are the same as in the lower bound. Note that similar results can be easily proved for instance for Besov balls $\mathcal{B}^{\alpha}_{\infty,2}(L)$, with $\alpha, L > 0$ (see DeVore and Lorentz, 1993) by using an appropriate collection of models such as trigonometric polynomials defined by (5).

3.1.4. Simulation experiments

Results of simulation experiments are provided to support the theoretical analysis developed in Section 3.1.2. Samples of size $n = 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 10\,000$ have been generated from a mixture of Beta distributions

$$\forall x \in [0,1], \qquad s(x) = \frac{\beta(3,7;x) + \beta(10,5;x)}{2} \ , \tag{16}$$

which is a Hölderian density on $[0,1]$. For each n, every $p \in \{1, \dots, n-1\}$ have been considered. Note that in these experiments, (**Dmax**) is fulfilled with $\Gamma = 1$ (Figure 1) and $\Gamma = 2$ (Figure 2).

The model collection we used is made of piecewise constant functions described in Section 2.1.2 leading to regular histogram estimators defined by (4). For every $1 \leq p \leq n-1$, $\widehat{m}(p)$ is defined by (13).

Let us also introduce

$$C_{or,n}(p) := \mathbb{E}\left[\frac{\left\| s - \widehat{s}_{\widehat{m}(p)} \right\|^2}{\inf_{m \in \mathcal{M}_n} \left\{ \left\| s - \widehat{s}_m \right\|^2 \right\}} \right] \quad \text{and} \quad p_0 := \text{Argmin}_{1 \leq p \leq n-1} C_{or,n}(p) \ , \tag{17}$$

which measures the average performance of $\widehat{s}_{\widehat{m}(p)}$ with respect to that of the oracle estimator $\widehat{s}_{m^*}$. Thus the closer $C_{or,n}(p)$ to 1, the better $\widehat{s}_{\widehat{m}(p)}$. Then, minimizing $C_{or,n}(p)$ as a function of p for various values of n allows us to check whether the conclusions drawn from minimizing $C_n(p)$ with respect to p (Theorem 3.1 and Corollary 3.1) hold true or not, that is whether $C_n(p)$ is an accurate approximation of $C_{or,n}(p)$. For each curve $p \mapsto C_{or,n}(p)$, a confidence band has been also displayed. It is delimited by $p \mapsto C^-_{or,n}(p)$ and $p \mapsto C^+_{or,n}(p)$ respectively defined by

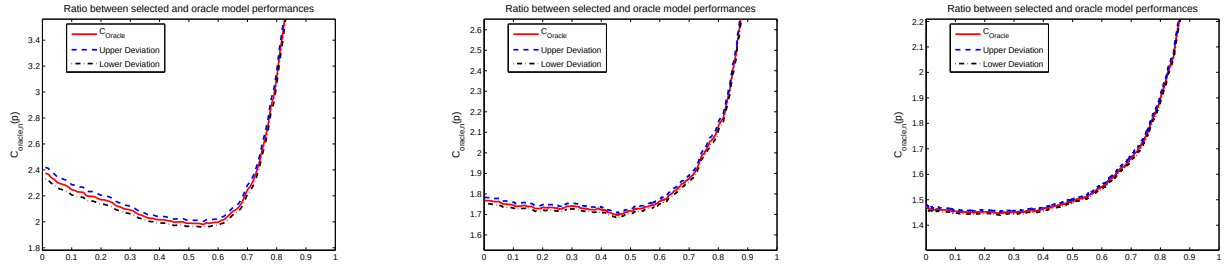
$$C^-_{or,n}(p) = C_{or,n}(p) - \frac{\widehat{\sigma}}{\sqrt{N}}, \quad \text{and} \quad C^+_{or,n}(p) = C_{or,n}(p) + \frac{\widehat{\sigma}}{\sqrt{N}} \ , \tag{18}$$

where $\widehat{\sigma}$ denotes the empirical standard deviation.

First from Figure 1, curves $p/n \mapsto C_{or,n}(p)$ (plain red lines) decrease to 1 uniformly with p as n grows. This confirms Theorem 3.1 where $C_n(p) \to 1$ as $n \to +\infty$ when p is kept fixed. Furthermore, $p \mapsto C_n(p)$ and $p \mapsto C_{or,n}(p)$ have a similar behavior since, as suggested by Corollary 3.1 when n is fixed, $C_{or,n}(p)$ is minimized for $p > 1$ but increases as p/n gets closer to 1. Recalling $C_{or,n}(p)$ measures the accuracy of $\widehat{s}_{\widehat{m}(p)}$, previous remarks show Theorem 3.1 is accurate enough to make $C_n(p)$ a reliable measure of the performance of $\widehat{s}_{\widehat{m}(p)}$ with respect to p. In particular, optimizing $C_n(p)$ as a function of p actually amounts to finding the best estimator among $\left\{ \widehat{s}_{\widehat{m}(p)} \right\}_{1 \leq p \leq n-1}$.

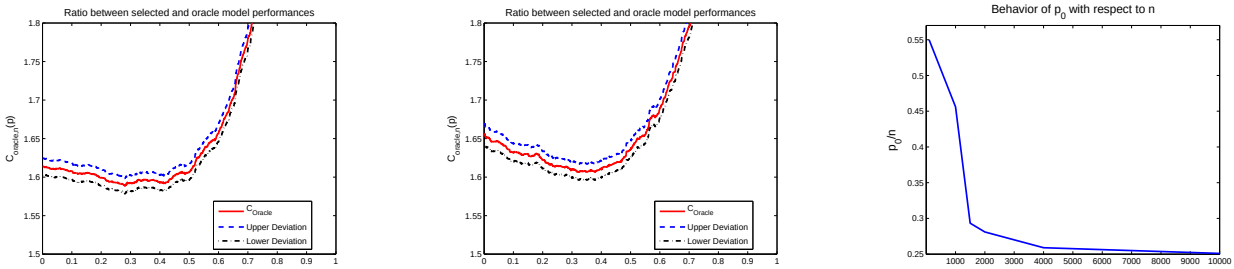
Second from (a) to (c) (Figure 1), the shape of $p/n \mapsto C_{or,n}(p)$ changes, its minimum location becoming less clear as n grows from $n = 100$, to $n = 10\,000$. According to Corollary 3.1, p is chosen large enough to balance the deviations due to δ_n (model collection complexity). Since $\delta_n \to 0$ as $n \to +\infty$, this requirement on p vanishes as n grows. This explanation is also supported by Figure 2 (c) where p_0/n (see Eq. (17)) has been displayed for different values of n. It shows p_0/n slowly decreases as n grows, which has been proved in Corollary 3.1 ($p^*/n \sim 10\delta_n$ with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as n tends to $+\infty$).

Finally, (a) and (b) in Figure 2 display $p/n \mapsto C_{or,n}(p)$ for $n = 2\,000$ and (a) $\Gamma = 1$, (b) $\Gamma = 2$. As indicated by (**Dmax**), an increase of Γ results in a more complex collection of models, inducing larger



(a) (b) (c)

FIG 1. *From (a) to (c),* $p/n \mapsto C_{or,n}(p)$ *(plain red line) is plotted for* $\Gamma = 1$ *(see* (**Dmax**)*) and different values of* n*: (a)* $n = 100$*, (b)* $n = 1000$*, (c) and* $n = 10\,000$*.* $p/n \mapsto C^{+}_{oracle,n}(p)$ *(blue dashed line) and* $p/n \mapsto C^{-}_{oracle,n}(p)$ *(black dot-dashed line) have been plotted on the same graph as well (see* (18)*). In each setting,* $N = 1000$ *samples have been drawn from the mixture of Beta distributions defined by* (16)*.*



(a) (b) (c)

FIG 2. *For (a) and (b),* $p/n \mapsto C_{or,n}(p)$ *(plain red line) is plotted for* $n = 2000$ *and different values of* Γ*(see* (**Dmax**)*): (a)* $\Gamma = 1$ *, (b)* $\Gamma = 2$*.* $p/n \mapsto C^{+}_{oracle,n}(p)$ *(blue dashed line) and* $p/n \mapsto C^{-}_{oracle,n}(p)$ *(black dot-dashed line) have been plotted on the same graph as well (see* (18)*). In each setting,* $N = 1000$ *samples have been drawn from the mixture of Beta distributions defined by* (16)*. For (c),* $n \mapsto p_0/n$ *is displayed, where* p_0 *denotes the minimizer of* $C_{or,n}(p)$ *as a function of* p*.*

deviations (δ_n slower). On the one hand, the curve in (b) ($\Gamma = 2$) is above that in (a) ($\Gamma = 1$). The performance of $\widehat{s}_{\widehat{m}(p)}$ worsens as the collection of modes becomes more complex. On the other hand, the minimum location is also larger for $\Gamma = 2$ than for $\Gamma = 1$. Since Γ is larger, so is δ_n. Then, p has to be chosen larger to balance the effect of δ_n. Since the same phenomena have been observed in other simulation experiments (not reported here), one concludes the optimal p is strongly linked with the model collection structure: the more complex the model collection, the larger the optimal p.

3.2. Optimal cross-validation for Identification

3.2.1. Identification point of view

With the notation of Section 2.1, $\{\widehat{s}_m\}_{m\in\mathcal{M}_n}$ denotes a collection of projection estimators (Section 2.1.2) which is allowed to depend on n. From the Identification point of view, one assumes

$$\{m \in \mathcal{M}_n,\ s \in S_m\} \neq \emptyset \ .$$

The purpose is to find the smallest model containing s, denoted by $S_{\bar{m}}$ and defined by

$$\bar{m} := \mathrm{Argmin}_{m \in \mathcal{M}_n} \mathbb{E}\left[\left\| s_m - \widehat{s}_m \right\|^2 \right] \ , \tag{19}$$

where s_m denotes the orthogonal projection of s onto S_m. Note that Assumptions (**LoEx**) and (**RegD**) imply $\xi D_m \leq \mathbb{E}\left[\left\| s_m - \widehat{s}_m \right\|^2 \right] \leq \left\| s \right\|^2 \Phi D_m$. For every $m \in \mathcal{M}_n$, $\mathbb{E}\left[\left\| s_m - \widehat{s}_m \right\|^2 \right]$ is related to D_m as a measure of the size of S_m. However unlike the dimension D_m, $\mathbb{E}\left[\left\| s_m - \widehat{s}_m \right\|^2 \right]$ measures the size of S_m through s. Thus, a model S_m is not simply "too large" because it depends on more parameters, but rather because the estimation error $\mathbb{E}\left[\left\| s_m - \widehat{s}_m \right\|^2 \right]$ incurred by S_m is too large.

3.2.2. Main model consistency result

In the following analysis, one further assumes $\bar{m}$ does not depend on n for large enough values of n. First, it entails $\bar{m} \in \mathcal{M}_n$ for large enough values of n. Second, letting $\mathcal{M}_n$ grow with n amounts for instance to include too large models in $\{S_m\}_{m \in \mathcal{M}_n}$ without modifying $\bar{m}$. In particular, it is not required that $\{S_m\}_{m \in \mathcal{M}_n}$ is nested.

Let us first describe the asymptotic behavior of $\widehat{R}_p$ as a function of $1 \leq p \leq n-1$.

Theorem 3.3 (Asymptotic behavior of $\widehat{R}_p$)**.** *Let $\cup_{m \in \mathcal{M}_n} S_m$ be a collection of models satisfying* (**Pol**)*, $\bar{m} \in \mathcal{M}$ be defined by* (19) *such that $\bar{m}$ does not depend on n, and assume* (**SqI**)*,* (**RegD**)*,* (**Dmax**)*,* (**LoEx**) *hold true. Then, an event Ω_n exists with $\mathbb{P}\left[\Omega_n \right] \geq 1 - 8/n^2$ on which for every $p = p_n$ such that*

$$n\left(1 - \frac{p}{n}\right) \xrightarrow[n \to +\infty]{} +\infty \ , \tag{20}$$

1. *if $s \notin S_m$,*

$$\widehat{R}_p(m) - \widehat{R}_p(\bar{m}) = \left\| s - s_m \right\|^2 + o_{\mathbb{P}}(1) > 0 \ ,$$

2. *if $s \in S_m$,*

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq \frac{n}{n-p}\left[(1+\delta_n)\mathbb{E}\left[Z_{\bar{m}}^2 \right] - (1-\delta_n)\mathbb{E}\left[Z_m^2 \right] \right] + 4L_n\left(\mathbb{E}\left[Z_{\bar{m}}^2 \right] + \mathbb{E}\left[Z_m^2 \right] \right) \tag{21}$$

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \geq \frac{n}{n-p}\left[(1-\delta_n)\mathbb{E}\left[Z_{\bar{m}}^2 \right] - (1+\delta_n)\mathbb{E}\left[Z_m^2 \right] \right] - 4L_n\left(\mathbb{E}\left[Z_{\bar{m}}^2 \right] + \mathbb{E}\left[Z_m^2 \right] \right) \ , \tag{22}$$

where $\delta_n \to 0$ with $n\delta_n \to +\infty$ as $n \to +\infty$, and $L_n = \sqrt{4\sqrt{2\Phi}} n^{-1/2}$.

On the one hand, the constraint (20) ensures $\widehat{R}_p(m)$ is a consistent estimator of $R_n(\widehat{s}_m)$. It only requires $1 - p/n$ converges to 0 slower than $1/n$. Any $p = p_n$ satisfying (20) enables to discard too small models S_m such that $s \notin S_m$ since $\widehat{R}_p(m) > \widehat{R}_p(\bar{m})$. On the other hand when $s \in S_m$, upper and lower bounds (21) and (22) give possible deviations of $\widehat{R}_p(m) - \widehat{R}_p(\bar{m})$ with high probability for large models. These bounds relate p to δ_n and L_n which are determined by the structure of the model collection $\{S_m\}_{m \in \mathcal{M}_n}$ at hand and the probability of the event Ω_n.

Proof of Theorem 3.3. From Proposition A.2, for every $m, m' \in \mathcal{M}$,

$$\begin{aligned}\widehat{R}_p(m') - \widehat{R}_p(m) = \mathbb{E}\left[\, \|s - \widehat{s}_{\,m'}\|^2 \,\right] - \mathbb{E}\left[\, \|s - \widehat{s}_{\,m}\|^2 \,\right] \\ + \frac{p}{n-p}\left(\mathbb{E}\left[\, Z_{m'}^2 \,\right] - \mathbb{E}\left[\, Z_m^2 \,\right]\right) \\ + K(n,p)\left[\, Z_m^2 - \mathbb{E}\left[\, Z_m^2 \,\right]\,\right] - K(n,p)\left[\, Z_{m'}^2 - \mathbb{E}\left[\, Z_{m'}^2 \,\right]\,\right] \\ - 2K(n,p)\,\nu_n\left(s_{m'} - s_m\right) + \frac{1}{n}\left(K(n,p) + \frac{n}{n-p}\right)\nu_n\left(\phi_{m'} - \phi_m\right) \quad ,\end{aligned}$$

where $K(n,p) = 1 + 2/(n-1) + p/\left[(n-1)(n-p)\right]$.

Setting $\Delta(m) = \widehat{R}_p(m) - \mathbb{E}\left[\, \|s - \widehat{s}_{\,m}\|^2 \,\right] - \frac{p}{n-p}\mathbb{E}\left[\, Z_m^2 \,\right]$, it results

$$\begin{aligned}|\Delta(m') - \Delta(m)| \leq K(n,p)\left|\left[\, Z_m^2 - \mathbb{E}\left[\, Z_m^2 \,\right]\,\right] - \left[\, Z_{m'}^2 - \mathbb{E}\left[\, Z_{m'}^2 \,\right]\,\right]\right| \\ + 2K(n,p)\,\left|\nu_n\left(s_{m'} - s_m\right)\right| + \frac{1}{n}\left(K(n,p) + \frac{n}{n-p}\right)\left|\nu_n\left(\phi_{m'} - \phi_m\right)\right| \quad .\end{aligned}$$

First, Propositions A.5 and A.6 imply there exist an event $\Omega_{\text{right}} \cap \Omega_{\text{left}}$ of probability at least $1 - 2/n^2 - \beta_1 - \beta_2$ on which

$$K(n,p)\left|\left[\, Z_m^2 - \mathbb{E}\left[\, Z_m^2 \,\right]\,\right] - \left[\, Z_{m'}^2 - \mathbb{E}\left[\, Z_{m'}^2 \,\right]\,\right]\right| \leq K(n,p)\left[\,(\delta_n + L_{m'})\mathbb{E}\left[\, Z_m^2 \,\right] + (\delta_n + L_m)\mathbb{E}\left[\, Z_m^2 \,\right]\,\right] \quad . \tag{23}$$

Let us notice that since D_m does not depend from n, $L_m > 0$ for large enough values of n. Furthermore choosing $\beta_1 = \beta_2 = \beta_n$ with $\beta_n \to 0$ as $n \to +\infty$, it comes $L_m = L_{m'} = L_n = \sqrt{4\sqrt{2\Phi}}\beta_n^{-1/4} \to +\infty$ as n tends to $+\infty$. Subsequently, choosing β_n such that $n^4\beta_n \to +\infty$ as $n \to +\infty$ results in $L_n/n \to 0$ as $n \to +\infty$.

Second, Proposition A.3 entails there exists an event $\Omega_{\text{rem},1}$ with probability at least $1 - 2/n^2$ on which

$$\begin{aligned}\frac{1}{n}\left(K(n,p) + \frac{n}{n-p}\right)\left|\nu_n\left(\phi_{m'} - \phi_m\right)\right| &\leq \delta_n\left(K(n,p) + \frac{n}{n-p}\right)\left(\mathbb{E}(Z_m^2) + \mathbb{E}(Z_{m'}^2)\right) \\ &\leq \delta_n\left(3 + \frac{n}{n-p}\right)\left(\mathbb{E}(Z_m^2) + \mathbb{E}(Z_{m'}^2)\right) \quad ,\end{aligned} \tag{24}$$

since $K(n,p) \leq 3$ for $n \geq 4$.

Third for $m' = \bar{m}$, Proposition A.8 entails there exists an event $\Omega_{\text{rem},3}$ with probability at least $1 - 2/n^2$ on which

$$2K(n,p)\,\left|\nu_n\left(s_{\bar{m}} - s_m\right)\right| \leq 3\delta_n\left\|s_m - s_{\bar{m}}\right\|^2 + 3\delta_n\|s\|\sqrt{\Phi}\sqrt{\frac{D_m + D_{\bar{m}}}{n}} + 3\delta_n\Phi\frac{D_m + D_{\bar{m}}}{n} \quad . \tag{25}$$

Combining (23), (24), and (25), there exist an event $\Omega_n := \Omega_{\text{right}} \cap \Omega_{\text{left}} \cap \Omega_{\text{rem},1} \cap \Omega_{\text{rem},3}$ with probability at least $1 - 6/n^2 - 2\beta_n$ on which two settings occur:

1. If $s \notin S_m$,

$$\begin{aligned}|\Delta(\bar{m}) - \Delta(m)| \leq \left(3L_n + 6\delta_n + \frac{n}{n-p}\delta_n\right)\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] + \mathbb{E}\left[\, Z_m^2 \,\right]\right) \\ + 3\delta_n\left\|s_m - s_{\bar{m}}\right\|^2 + 3\delta_n\|s\|\sqrt{\Phi}\sqrt{\frac{D_m + D_{\bar{m}}}{n}} + 3\delta_n\Phi\frac{D_m + D_{\bar{m}}}{n} \quad .\end{aligned} \tag{26}$$

2. If $s \in S_m$,

$$|\Delta(\bar{m}) - \Delta(m)| \leq \left(3L_n + 6\delta_n + \frac{n}{n-p}\delta_n\right)\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] + \mathbb{E}\left[\, Z_m^2 \,\right]\right) \ . \tag{27}$$

In these two settings for every $m \in \mathcal{M}$, (**RegD**) and $\delta_n \xrightarrow[n\to+\infty]{} 0$ imply

$$|\Delta(\bar{m}) - \Delta(m)| = o_{\mathbb{P}}(1) \ ,$$

hence,

$$\widehat{R}_p(m) - \widehat{R}_p(\bar{m}) = \mathbb{E}\left[\, \|s - \widehat{s}_m\|^2 \,\right] - \mathbb{E}\left[\, \|s - \widehat{s}_{\bar{m}}\|^2 \,\right] + \frac{p}{n-p}\left(\mathbb{E}\left[\, Z_m^2 \,\right] - \mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right]\right) + o_{\mathbb{P}}(1) \ .$$

Hence, requiring $\widehat{R}_p(m) - \widehat{R}_p(\bar{m}) - \left(\mathbb{E}\left[\, \|s - \widehat{s}_m\|^2 \,\right] - \mathbb{E}\left[\, \|s - \widehat{s}_{\bar{m}}\|^2 \,\right]\right) \to 0$ as $n \to +\infty$ implies the necessary constraint $p/\left[\,(n-p)n\,\right] \to 0$, which amounts to

$$n\left(1 - \frac{p}{n}\right) \xrightarrow[n\to+\infty]{} +\infty \ . \tag{28}$$

On the one hand, it is then straightforward to check that Eq. (26) leads, for every $m \in \mathcal{M}$ such that $s \notin S_m$, to

$$\widehat{R}_p(m) - \widehat{R}_p(\bar{m}) = \|s - s_m\|^2 + o_{\mathbb{P}}(1) \ .$$

On the other hand, for every $m \in \mathcal{M}$ such that $s \in S_m$, Eq. (27) provides

$$\left|\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) - \frac{n}{n-p}\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] - \mathbb{E}\left[\, Z_m^2 \,\right]\right)\right| \leq \left(4L_n + \frac{n}{n-p}\delta_n\right)\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] + \mathbb{E}\left[\, Z_m^2 \,\right]\right) \ ,$$

hence

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq \frac{n}{n-p}\left[\,(1+\delta_n)\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] - (1-\delta_n)\mathbb{E}\left[\, Z_m^2 \,\right]\right] + 4L_n\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] + \mathbb{E}\left[\, Z_m^2 \,\right]\right) \ , \tag{29}$$

and

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \geq \frac{n}{n-p}\left[\,(1-\delta_n)\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] - (1+\delta_n)\mathbb{E}\left[\, Z_m^2 \,\right]\right] - 4L_n\left(\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] + \mathbb{E}\left[\, Z_m^2 \,\right]\right) \ .$$

Using $\beta_n = 1/n^2$ enables to conclude.

□

From the upper bound (21), one derives a sufficient condition on p to discard too large models. This condition enables to determine the minimal rate at which p/n has to decrease to 0 as n tends to $+\infty$, which ensures *model consistency* for $\widehat{m}$.

Corollary 3.3 (Model consistency)**.** *With the same notation and assumptions as Theorem 3.3, let us define* $\widehat{m} = \widehat{m}(p) = \mathrm{Argmin}_{m\in\mathcal{M}}\widehat{R}_p(m)$ *for every* $1 \leq p \leq n-1$*. Then any* $p = p_n$ *such that*

$$n\left(1 - \frac{p}{n}\right) \xrightarrow[n\to+\infty]{} +\infty, \quad \text{and} \quad 0 < 1 - \frac{p}{n} < \frac{K}{\sqrt{n}} \quad \text{with} \quad K = \left(8\sqrt{\sqrt{2\Phi}}\right)^{-1} \ ,$$

leads to

$$\mathbb{P}\left[\,\widehat{m} = m^*\,\right] \xrightarrow[n\to+\infty]{} 1 \ .$$

First the main conclusion is that model consistency results from requiring $p_n/n \to 1$ as $n \to +\infty$. One therefore recovers previous results by Shao (1993) and Yang (2007) established in the regression framework. However, our Corollary 3.3 is more precise than Shao (1993) since it localizes the optimal convergence rate of $1 - p/n$ between $1/\sqrt{n}$ and $1/n$. In particular, Loo (and Lpo with any $p = o(n)$) is completely misleading for identifying the model $S_{\bar{m}}$. Second, p has to be chosen large enough to balance the deviations (L_n) in (30). Indeed, the rate $1/\sqrt{n}$ is determined by the structure of the model collection $\{S_m\}_{m\in\mathcal{M}_n}$ and the probability of the event Ω_n. Another collection of models could have produced another minimal rate.

Proof of Corollary 3.3. Applying Eq. (29) for every $m \neq \bar{m} \in \mathcal{M}$ such that $s \in S_m$, it results a sufficient condition on p such that $\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) < 0$, that is

$$0 < \left(\frac{n}{n-p}(1+\delta_n) + 4L_n\right)\mathbb{E}\left[\, Z_{\bar{m}}^2 \,\right] < \left(\frac{n}{n-p}(1-\delta_n) - 4L_n\right)\mathbb{E}\left[\, Z_m^2 \,\right] \quad . \tag{30}$$

This leads to require $\frac{n}{n-p} > 4L_n(1-\delta_n)^{-1} > 4L_n$, which can be reformulated as

$$\frac{1}{1-\frac{p}{n}} > 4L_n = 8\sqrt{\sqrt{2\Phi}}\beta_n^{-1/4} \quad \Leftrightarrow \quad 0 < 1 - \frac{p}{n} < \frac{\beta_n^{1/4}}{8\sqrt{\sqrt{2\Phi}}} \cdot$$

The conclusion results from choosing $\beta_n = 1/n^2$ and $\mathbb{P}\left[\,\Omega_n\,\right] \xrightarrow[n\to+\infty]{} 1$.

□

4. Discussion

From the present analysis of CV algorithms in te density estimation framework, we were able to prove the optimality of leave-one-out cross-validation for risk estimation. Besides when CV is used as model selection procedure, the optimal p strongly depends on the structure of the model collection and on our goal (estimation or identification). However this characterization of the behavior of the optimal p provides some guidelines, but does not result in a data-driven choice of p.

A possible way to design such a data-driven choice is to follow the same idea as Shao (1997) exploiting the deep connection between CV and penalized criteria. Let us describe the heuristic argument leading to this choice. Arlot (2008) introduced the ideal penalty defined for every $m \in \mathcal{M}_n$ by

$$\mathrm{pen}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m) \quad ,$$

with the notation of Section 2.1. It enables to rephrase $\ell(s, \widehat{s}_m)$ in terms of a penalized criterion

$$\ell(s, \widehat{s}_m) = P\gamma(\widehat{s}_m) - P\gamma(s) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{id}}(m) - P\gamma(s) \quad .$$

Similarly for the Lpo risk estimator,

$$\widehat{R}_p(\widehat{s}_m) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{Lpo}}(m) \quad ,$$

where $\mathrm{pen}_{\mathrm{Lpo}}(m) = \widehat{R}_p(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)$ is called the Lpo-penalty (see Celisse, 2008). Then in our setting, some simple algebra provides

$$\mathbb{E}\left[\,\mathrm{pen}_{\mathrm{Lpo}}(m)\,\right] = \frac{2n-p}{2(n-p)}\mathbb{E}\left[\,\mathrm{pen}_{\mathrm{id}}(m)\,\right] \quad ,$$

showing that on average $\mathrm{pen}_{\mathrm{Lpo}}$ is equal to $\mathrm{pen}_{\mathrm{id}}$ up to a multiplicative constant. Thus, using the so-called slope heuristics (Arlot and Massart, 2009, in the regression framework) could serve to calibrate the optimal p of CV algorithms.

Appendix A: Proofs of Sections 2 and 3

A.1. Closed-form expressions

Lemma A.1. *With the notation of Section 2.2.1, for any $i \neq j \neq k \in \{1, \ldots, n\}$,*

$$\begin{aligned}
\sum_{e \in \mathcal{E}_p} \mathbb{1}_{(j \in (e))} &= \binom{n-1}{p} \qquad \text{and} \qquad \sum_{e \in \mathcal{E}_p} \mathbb{1}_{(j \in (e))} \mathbb{1}_{(k \in (e))} = \binom{n-2}{p-1}, \\
\sum_{e \in \mathcal{E}_p} \mathbb{1}_{(i \in e)} \mathbb{1}_{(j \in (e))} \mathbb{1}_{(k \in (e))} &= \binom{n-3}{p-1} \qquad \text{and} \qquad \sum_{e \in \mathcal{E}_p} \mathbb{1}_{(i \in e)} \mathbb{1}_{(j \in (e))} = \binom{n-2}{p-1} \ .
\end{aligned}$$

Lemma A.2. *With the same notation as Proposition 2.2, it comes*

$$\begin{aligned}
a_n &= n^2 \mathrm{Var}\left[\sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] + \mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2 \right] - 2n \mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] \ , \\
b_n &= 2n^2 \left(\mathrm{Var}\left[\sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] - \mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] \left(1 + \frac{1}{n} \right) + \frac{1}{n} \mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2 \right] \right) \ , \\
c_n &= n^2 \mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2 \right] + n^2 \mathrm{Var}\left[\sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] - 2n^2 \mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2 \right] \ .
\end{aligned}$$

Lemma A.3. *With the notation of Section 2.3, simple algebra leads to*

$$\begin{aligned}
\mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2 \right] &= \frac{1}{n} \mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1) \right] \ , \\
\mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_\lambda (P_n \varphi_\lambda)^2 \right] &= \frac{1}{n^2} \mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1) \right] + 2 \frac{n-1}{n^2} \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1) \varphi_\lambda(X_2) \right] \\
\mathrm{Var}\left[\sum_\lambda (P_n \varphi_\lambda)^2 \right] &= \frac{\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1) \right]}{n^3} + 4 \frac{n-1}{n^3} \mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1) \varphi_\lambda(X_2) \right] \\
&\quad + 4 \frac{(n-1)(n-2)}{n^3} \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1) \varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1) \varphi_\lambda(X_3) \right] \\
&\quad + 4 \frac{n-1}{n^3} \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1) \varphi_\lambda(X_3) \right] \ .
\end{aligned}$$

Proposition A.1. *With the notation of Lemma A.2,*

$$\begin{aligned}
a_n &= 4 \frac{n-1}{n} \alpha + 4 \frac{(n-1)(n-2)}{n} \beta \\
b_n &= 8 \frac{n-1}{n} \alpha + 8 \frac{(n-1)(n-2)}{n} \beta - 4(n-1) \left(1 - \frac{1}{n} \right) \gamma \\
c_n &= 4 \frac{n-1}{n} \alpha + 4 \frac{(n-1)(n-2)}{n} \beta - 4 (n-1) \left(1 - \frac{1}{n} \right) \gamma + \left(n - 2 + \frac{1}{n} \right) \delta \ .
\end{aligned}$$

where $\alpha = \mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right]$, $\beta = \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]$, $\gamma = \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]$, *and* $\delta = \mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right]$.

Proof of Proposition A.1. Using Lemmas A.2 and A.3, it comes

$$\begin{aligned}
a_n &= n^2 \left[\frac{\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right]}{n^3} + 4\frac{n-1}{n^3}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right] \right. \\
&\quad + 4\frac{(n-1)(n-2)}{n^3}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \\
&\quad \left. + 4\frac{n-1}{n^3}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right]\right] + \frac{1}{n}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right] \\
&\quad - 2n\left[\frac{1}{n^2}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right] + 2\frac{n-1}{n^2}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right]\right] \\
&= 4\frac{n-1}{n}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right] \\
&\quad + 4\frac{(n-1)(n-2)}{n}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \quad .
\end{aligned}$$

In the same way,

$$\begin{aligned}
b_n &= 2n^2\left(\mathrm{Var}\left[\sum_{\lambda\in\Lambda(m)} (P_n\varphi_\lambda)^2\right] - \mathrm{Cov}\left[\sum_\lambda P_n\varphi_\lambda^2, \sum_{\lambda\in\Lambda(m)} (P_n\varphi_\lambda)^2\right]\left(1+\frac{1}{n}\right) + \frac{1}{n}\mathrm{Var}\left[\sum_\lambda P_n\varphi_\lambda^2\right]\right) \\
&= 8\frac{n-1}{n}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right] \\
&\quad + 8\frac{(n-1)(n-2)}{n}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \\
&\quad - 4(n-1)\left(1-\frac{1}{n}\right)\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \quad .
\end{aligned}$$

Finally,

$$\begin{aligned}
c_n &= n^2\mathrm{Var}\left[\sum_\lambda P_n\varphi_\lambda^2\right] + n^2\mathrm{Var}\left[\sum_{\lambda\in\Lambda(m)} (P_n\varphi_\lambda)^2\right] - 2n^2\mathrm{Cov}\left[\sum_\lambda P_n\varphi_\lambda^2, \sum_{\lambda\in\Lambda(m)} (P_n\varphi_\lambda)^2\right] \\
&= \left(n-2+\frac{1}{n}\right)\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right] + 4\frac{n-1}{n}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right] \\
&\quad + 4\frac{(n-1)(n-2)}{n}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \\
&\quad - 4\,(n-1)\left(1-\frac{1}{n}\right)\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \quad .
\end{aligned}$$

□

Proposition A.2. *For every $m, m' \in \mathcal{M}$ and $p \in \{1, \dots, n-1\}$, it comes*

$$\begin{aligned}
&\widehat{R}_p(m') - \widehat{R}_p(m) \\
&= \left(\frac{n}{n-p}\right)\left(\mathbb{E}\left[\left\|s_{m'} - \widehat{s}_{m'}\right\|^2\right] - \mathbb{E}\left[\left\|s_m - \widehat{s}_m\right\|^2\right]\right) + \left[\left\|s - s_{m'}\right\|^2 - \left\|s - s_m\right\|^2\right] \\
&- K(n,p)\left[\left\|s_{m'} - \widehat{s}_{m'}\right\|^2 - \mathbb{E}\left[\left\|s_{m'} - \widehat{s}_{m'}\right\|^2\right]\right] + K(n,p)\left[\left\|s_m - \widehat{s}_m\right\|^2 - \mathbb{E}\left[\left\|s_m - \widehat{s}_m\right\|^2\right]\right] \\
&- 2K(n,p)\,\nu_n\left(s_{m'} - s_m\right) + \frac{1}{n}\left(K(n,p) + \frac{n}{n-p}\right)\nu_n\left(\phi_{m'} - \phi_m\right) \quad ,
\end{aligned}$$

where

$$K(n,p) = 1 + \frac{1}{n-1} + \frac{n}{n-p}\frac{1}{n-1} \quad .$$

A.2. Bounding remainder terms

Proposition A.3 (Bound on $\nu_n\left(\phi_m - \phi_{m'}\right)$)**.**
Let us assume (**RegD**) *and apply* (42) *with $t = \phi_m$ and $x = x_m = c_1 n \mathbb{E}(Z_m^2)$ ($c_1 > 0$). Then, an event $\Omega_{\mathrm{rem},1}$ exists with $\mathbb{P}\left[\Omega_{\mathrm{rem},1}\right] \geq 1 - 2\sum_{m \in \mathcal{M}} e^{-x_m}$, on which for every $m, m' \in \mathcal{M}_n$*

$$\left|\nu_n\left(\phi_m - \phi_{m'}\right)\right| \leq \frac{n\mathbb{E}(Z_m^2) + n\mathbb{E}(Z_{m'}^2)}{\log n} \quad ,$$

where $Z_m = \sup_{t \in S_m} \nu_n(t)$ for every m.

Proof of Proposition A.3. A straightforward use of (42) leads to the expected conclusion. □

Proposition A.4 (Bound on $\nu_n\left(s_m - s_{m'}\right)$)**.** *Let us assume* (**Pol**),(**SqI**), (**RegD**), (**LoEx**), *and* (**OrSp**) *hold true. Then, there exists a sequence $(\delta_n)_{\mathbb{N}}$ such that for every $m, m' \in \mathcal{M}$,*

$$\mathbb{P}\left[2\left|\nu_n(s_m - s_{m'})\right| > \delta_n\left(\mathbb{E}\left\|s - \widehat{s}_m\right\|^2 + \mathbb{E}\left\|s - \widehat{s}_{m'}\right\|^2\right)\right] \leq 2n^{-(2a_{\mathcal{M}}+2)} \quad ,$$

with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as $n \to +\infty$, and $0 \leq \delta_n \leq 1$ for n large enough.
Furthermore, an event $\Omega_{\mathrm{rem},2}$ exists with $\mathbb{P}\left[\Omega_{\mathrm{rem},2}\right] \geq 1 - 2/n^2$, on which for every $m, m' \in \mathcal{M}$

$$2\left|\nu_n(s_m - s_{m'})\right| \leq \delta_n\left(\mathbb{E}\left\|s - \widehat{s}_m\right\|^2 + \mathbb{E}\left\|s - \widehat{s}_{m'}\right\|^2\right) \quad .$$

Proof of Proposition A.4. For every $\eta > 0$,

$$\begin{aligned}
2\nu_n(s_m - s_{m'}) &= 2\left\|s_m - s_{m'}\right\| \nu_n(t_{m,m'}) \\
&\leq \eta\left\|s_m - s_{m'}\right\|^2 + \eta^{-1}\left[\nu_n(t_{m,m'})\right]^2 \quad ,
\end{aligned}$$

where $t_{m,m'} = \left(s_m - s_{m'}\right) / \left\|s_m - s_{m'}\right\|$.

Thanks to (42) where $t = t_{m,m'}$, it comes

$$|\nu_n(t_{m,m'})| > \sqrt{2\frac{\mathrm{Var}\left(t_{m,m'}(X_1)\right)}{n}x} + \frac{\|t_{m,m'}\|_\infty}{3n}x \ ,$$

with probability not larger than $2\exp(-x)$, $x > 0$. Hence with (**SqI**), one has

$$\begin{aligned} 2\nu_n(s_m - s_{m'}) &\leq \eta \|s_m - s_{m'}\|^2 + 4\eta^{-1}\frac{\mathrm{Var}\left(t_{m,m'}(X_1)\right)}{n}x + 2\eta^{-1}\left(\frac{\|t_{m,m'}\|_\infty}{3n}x\right)^2 \\ &\leq 2\eta\left(\|s - s_m\|^2 + \|s - s_{m'}\|^2\right) + 4\eta^{-1}\frac{\|s\|\,\|t_{m,m'}\|_\infty}{n}x + 2\eta^{-1}\left(\frac{\|t_{m,m'}\|_\infty}{3n}x\right)^2 \ . \end{aligned} \tag{31}$$

Moreover assuming (**RegD**), it comes

$$\|t_{m,m'}\|_\infty \leq \sqrt{\Phi\left(D_m + D_{m'}\right)} \ .$$

Then,

$$\begin{aligned} \frac{\mathrm{Var}\left(t_{m,m'}(X_1)\right)}{n}x &\leq \frac{\|s\|\sqrt{\Phi}\sqrt{(D_m + D_{m'})}}{n}x \ , \\ \left(\frac{\|t_{m,m'}\|_\infty}{3n}x\right)^2 &\leq \Phi\left(D_m + D_{m'}\right)\frac{x^2}{9n^2} \ . \end{aligned}$$

Let us take $x = (2a_{\mathcal{M}} + 2)\log n$. Then,

$$\begin{aligned} \frac{\mathrm{Var}\left(t_{m,m'}(X_1)\right)}{n}x &\leq \frac{\|s\|\sqrt{\Phi}\sqrt{(D_m + D_{m'})}}{n}(2a_{\mathcal{M}} + 2)\log n \ , \\ \left(\frac{\|t_{m,m'}\|_\infty}{3n}x\right)^2 &\leq \Phi\left(D_m + D_{m'}\right)\frac{\left((2a_{\mathcal{M}} + 2)\log n\right)^2}{9n^2} \ . \end{aligned}$$

Then,

$$\begin{aligned} &2\frac{\nu_n(s_m - s_{m'})}{\mathbb{E}\|s - \widehat{s}_m\|^2 + \mathbb{E}\|s - \widehat{s}_{m'}\|^2} \\ &\leq 2\eta + 4\eta^{-1}\frac{\|s\|\sqrt{\Phi}\sqrt{(D_m + D_{m'})}}{n\left(\mathbb{E}\|s - \widehat{s}_m\|^2 + \mathbb{E}\|s - \widehat{s}_{m'}\|^2\right)}(2a_{\mathcal{M}} + 2)\log n \\ &+ 2\eta^{-1}\Phi\frac{D_m + D_{m'}}{n\left(\mathbb{E}\|s - \widehat{s}_m\|^2 + \mathbb{E}\|s - \widehat{s}_{m'}\|^2\right)}\frac{\left((2a_{\mathcal{M}} + 2)\log n\right)^2}{9n} \\ &\leq 2\eta + 4\eta^{-1}\frac{\|s\|\sqrt{\frac{\Phi}{\xi}}}{\sqrt{n\left(\mathbb{E}\|s - \widehat{s}_m\|^2 + \mathbb{E}\|s - \widehat{s}_{m'}\|^2\right)}}(2a_{\mathcal{M}} + 2)\log n \\ &+ 2\eta^{-1}\frac{\Phi}{\xi}\frac{\left((2a_{\mathcal{M}} + 2)\log n\right)^2}{9n} \ , \end{aligned}$$

thanks to (**LoEx**). Moreover using that

$$n\left(\mathbb{E}\|s - \widehat{s}_m\|^2 + \mathbb{E}\|s - \widehat{s}_{m'}\|^2\right) \geq 2n\inf_m \mathbb{E}\|s - \widehat{s}_m\|^2 =: 2nR_n^*,$$

it comes

$$2\frac{\nu_n(s_m - s_{m'})}{\mathbb{E}\left\|s - \widehat{s}_m\right\|^2 + \mathbb{E}\left\|s - \widehat{s}_{m'}\right\|^2} \leq 2\eta + 4\eta^{-1}\left\|s\right\|\sqrt{\frac{\Phi}{2\xi}}(2a_{\mathcal{M}}+2)\frac{1}{\sqrt{nR_n^*(\log n)^{-2}}} + 2\eta^{-1}\frac{\Phi}{\xi}(2a_{\mathcal{M}}+2)^2\frac{(\log n)^2}{9n} \ .$$

Then, (**OrSp**) entails there exists a sequence $\delta_n \to 0$, $n\delta_n \to +\infty$ as $n \to +\infty$ ($0 < \delta_n < 1$ for n large enough) such that

$$2\nu_n(s_m - s_{m'}) \leq \delta_n\left(\mathbb{E}\left\|s - \widehat{s}_m\right\|^2 + \mathbb{E}\left\|s - \widehat{s}_{m'}\right\|^2\right) \ .$$

Finally, let us notice that $\sum_{m,m'\in\mathcal{M}} 2n^{-(2a_{\mathcal{M}}+2)} \leq 2n^{2a_{\mathcal{M}}}n^{-(2a_{\mathcal{M}}+2)} = 2/n^2$.

□

A.3. Deviations of $\sqrt{n}Z_m$

A.3.1. Right deviation

Proposition A.5 (Right deviation of $\sqrt{n}Z_m$)**.** *Let us assume* (**Pol**), (**SqI**), (**RegD**), *and* (**LoEx**) *hold true, and set* $Z_m = \sup_{t\in S_m}\nu_n(t)$, $\sigma_m^2 = \sup_{t\in S_m}\mathrm{Var}\left[t(X_1)\right]$ *and* $b_m = \sup_{t\in S_m}\left\|t\right\|_\infty$. *Then, there exists a sequence* $(\delta_n)_{n\geq 1}$ *with* $\delta_n \to 0$ *and* $n\delta_n \to +\infty$ *as* $n \to +\infty$ *such that for every* $m \in \mathcal{M}$,

$$\sqrt{n}Z_m \leq \sqrt{n}\mathbb{E}(Z_m)\left[1 + \delta_n + \sqrt{4\sqrt{\frac{\Phi}{\xi}}C\left\|s\right\|\mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}}\right]$$

on an event Ω_{right} *with* $\mathbb{P}\left[\Omega_{\mathrm{right}}\right] \geq 1 - 1/n^2 - \beta_1$, *for any* $\beta_1 \in (0,1)$ *and* $C \geq \sqrt{2\xi/\beta_1}$.

Proof of Proposition A.5.
Let us use Eq. (43) and upper bound the deviation terms. Assuming (**SqI**) and (**RegD**), Lemma A.6 leads to

$$\sigma_m^2 \leq \left\|s\right\|\sqrt{\Phi}\sqrt{D_m} \ , \qquad b_m \leq \sqrt{\Phi}\sqrt{D_m} \ .$$

Furthermore, (**LoEx**) entails

$$\sigma_m^2 \leq \left\|s\right\|\sqrt{\frac{\Phi}{\xi}}\sqrt{n}\mathbb{E}(Z_m) \ , \qquad b_m \leq \sqrt{\frac{\Phi}{\xi}}\sqrt{n}\mathbb{E}(Z_m) \ .$$

Let us first upperbound $\sqrt{2\left(\sigma_m^2 + 2b_m\mathbb{E}(Z_m)\right)x_m}$:

1. If $\sqrt{D_m} \geq (\log n)^2$:
   Then choosing $x_m = (a_{\mathcal{M}}+2)\log n$, there exists a sequence δ_n decreasing to 0, $n\delta_n \to +\infty$ as $n \to +\infty$ such that

$$\sqrt{2\left(\sigma_m^2 + 2b_m\mathbb{E}(Z_m)\right)x_m} \leq \sqrt{n}\mathbb{E}(Z_m)\delta_n \ .$$

2. Otherwise $\sqrt{D_m} < (\log n)^2$:
   Then, $\sqrt{2\left(\sigma_m^2 + 2b_m\mathbb{E}(Z_m)\right)x_m}$ is no longer negligible with respect to $\sqrt{n}\mathbb{E}(Z_m)$. So, choosing $x_m = C\sqrt{n}\mathbb{E}(Z_m)$ $(C > 0)$ leads to

$$\sqrt{2\left(\sigma_m^2 + 2b_m\mathbb{E}(Z_m)\right)x_m} \leq \sqrt{n}E(Z_m)\sqrt{2\sqrt{\frac{\Phi}{\xi}}C\left(\|s\| + 2E(Z_m)\right)} \leq \sqrt{n}E(Z_m)\sqrt{4\sqrt{\frac{\Phi}{\xi}}C\|s\|} \ ,$$

   as long as n is large enough.

Let us now upperbound $\frac{b_m x_m}{3\sqrt{n}}$:

$$\frac{b_m x_m}{3\sqrt{n}} \leq \sqrt{n}\mathbb{E}(Z_m)\sqrt{\frac{\Phi}{\xi}}\frac{(a_{\mathcal{M}+2})\log n \vee C(\log n)^2}{3\sqrt{n}} \ .$$

Finally, we can remark that

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \sum_{D_m \geq (\log n)^4} n^{-(a_{\mathcal{M}}+2)} + \sum_{D_m < (\log n)^4} e^{-C\sqrt{n}E(Z_m)} \leq \frac{1}{n^2} + 2\frac{\xi}{c^2} \ .$$

□

Corollary A.1. *For $Z_m = \sup_{t\in S_m}\nu_n(t)$, set $L_m = \sqrt{4\sqrt{\frac{\Phi}{\xi}}C\|s\|\mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}}$. Then on the event Ω_{right} defined in Proposition A.5,*

$$Z_m^2 \leq \mathbb{E}(Z_m^2)\left(1 + \delta_n + L_m\right)^2 \ .$$

A.3.2. Left deviation

Proposition A.6 (Left deviation of $\sqrt{n}Z_m$)**.** *Let us assume* (**Pol**), (**SqI**), (**RegD**), *and* (**LoEx**) *hold true, and set $Z_m = \sup_{t\in S_m}\nu_n(t)$, $\sigma_m^2 = \sup_{t\in S_m}\mathrm{Var}\left[t(X_1)\right]$ and $b_m = \sup_{t\in S_m}\|t\|_\infty$. Then, there exists a sequence $(\delta_n)_{n\geq 1}$ with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as $n \to +\infty$ such that for every $m \in \mathcal{M}$,*

$$\sqrt{n}Z_m \geq \sqrt{n}\mathbb{E}(Z_m)\left[1 - \delta_n - \sqrt{4\sqrt{\frac{\Phi}{\xi}}C\|s\|\mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}}\right] \ , \tag{32}$$

on an event Ω_{left} with $\mathbb{P}\left[\Omega_{\text{left}}\right] \geq 1 - 1/n^2 - \beta_2$ for any $\beta_2 \in (0,1)$ and $C \geq \sqrt{2\xi/\beta_2}$.

Proof of Proposition A.6. Similar to that of Proposition A.5 with the use of Eq. (44) and the additional Proposition A.7 which provides an upper bound of $\mathbb{E}(Z_m)^2$ depending on $\mathbb{E}(Z_m^2)$. □

Proposition A.7 (Upper bound on $\mathrm{Var}(Z)$)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables defined on a mesurable space $(\mathcal{X}, \mathcal{T})$. Let S denote a set of real valued functions such that $\sup_{t\in S} \|t\|_\infty \leq b$, $\sup_{t\in S} \mathrm{Var}\left(t(X_1)\right) = \sigma^2$, and set $Z = \sup_{t\in S} \nu_n(t)$. Then,*

$$\mathrm{Var}(Z) \leq \frac{2\sigma^2 + 32b\mathbb{E}(Z)}{n} \quad . \tag{33}$$

Let us assume (**SqI**)*,* (**RegD**)*, and* (**LoEx**)*. If S denotes a linear space of dimension D, then there exists a positive sequence $(\delta_n)_{n\geq 1}$ with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as $n \to +\infty$ ($0 < \delta_n < 1$ for n large enough), and every constant $\theta > 0$ such that*

$$\mathbb{E}(Z^2) \leq \left(\mathbb{E}(Z)\right)^2 \left(1 + \delta_n + \theta\sqrt{\frac{\Phi}{\xi}}\mathbb{1}_{\left(\sqrt{D_m} < (\log n)^2\right)}\right) + r_n \quad ,$$

where $r_n = \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{2\|s\|^2}{n}\mathbb{1}_{\left(\sqrt{D} < (\log n)^2\right)}$.

Proof of Proposition A.7. Assumptions (**SqI**), (**RegD**), and (**LoEx**) provide

$$\mathbb{E}(Z^2) - \left(\mathbb{E}(Z)\right)^2 \leq 2\sqrt{\frac{\Phi}{\xi}}\left(\mathbb{E}(Z)\right)^2 \left(\frac{\|s\|}{\sqrt{n}\mathbb{E}(Z)} + \frac{16}{\sqrt{n}}\right) \quad .$$

1. If $\sqrt{n}\mathbb{E}(Z) \geq \sqrt{\xi D} \geq \sqrt{\xi}(\log n)^2$:

$$\begin{aligned}\mathbb{E}(Z^2) - \left(\mathbb{E}(Z)\right)^2 &\leq 2\sqrt{\frac{\Phi}{\xi}}\left(\mathbb{E}(Z)\right)^2 \left(\frac{\|s\|}{\sqrt{\xi}(\log n)^2} + \frac{16}{\sqrt{n}}\right) \\ &\leq \delta_{1,n}\left(\mathbb{E}(Z)\right)^2 \quad ,\end{aligned}$$

   with $\delta_{1,n} = 2\sqrt{\frac{\Phi}{\xi}}\left(\frac{\|s\|}{\sqrt{\xi}(\log n)^2} + \frac{16}{\sqrt{n}}\right)$.
2. Otherwise $\sqrt{n}\mathbb{E}(Z) \leq \sqrt{\Phi D} < \sqrt{\Phi}(\log n)^2$:

$$\begin{aligned}\mathbb{E}(Z^2) - \left(\mathbb{E}(Z)\right)^2 &\leq 2\sqrt{\frac{\Phi}{\xi}}\mathbb{E}(Z)\left(\frac{\|s\|}{\sqrt{n}} + \frac{16\mathbb{E}(Z)}{\sqrt{n}}\right) \leq \theta\sqrt{\frac{\Phi}{\xi}}\left(\mathbb{E}(Z)\right)^2 + \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{1}{n}\left(\|s\| + 16\mathbb{E}(Z)\right)^2 \\ &\leq \theta\sqrt{\frac{\Phi}{\xi}}\left(\mathbb{E}(Z)\right)^2 + \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{1}{n}\left(2\|s\|^2 + 32\left(\mathbb{E}(Z)\right)^2\right) \\ &\leq \left(\delta_{2,n} + \theta\sqrt{\frac{\Phi}{\xi}}\right)\left(\mathbb{E}(Z)\right)^2 + r_n\end{aligned}$$

   for every $\theta > 0$, with $\delta_{2,n} = \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{32}{n}$ and $r_n = \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{2\|s\|^2}{n}$.

Then, there exists a positive sequence $(\delta_n)_{n\geq 1}$ with $\delta_n = \max\left\{\delta_{1,n}, \delta_{2,n}\right\}$ decreasing to 0 with $n\delta_n \to +\infty$ as $n \to +\infty$, such that

$$\frac{\mathbb{E}(Z^2) - r_n}{1 + \delta_n + \theta\sqrt{\frac{\Phi}{\xi}}\mathbb{1}_{\left(\sqrt{D} < (\log n)^2\right)}} \leq \left(\mathbb{E}(Z)\right)^2 \quad .$$

□

Corollary A.2. *For* $Z_m = \sup_{t\in S_m} \nu_n(t)$, *set* $L_m = \sqrt{4\sqrt{\frac{\Phi}{\xi}} C \|s\| \mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}}$ *and* $r_n(m) = \theta^{-1}\sqrt{\frac{\Phi}{\xi}}\frac{2\|s\|^2}{n}\mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}$. *Then on the event* Ω_{left} *defined in Proposition A.6,*

$$\mathbb{E}(Z_m^2) \leq Z_m^2 \left(1-\delta_n - L_m\right)^{-3} + r_n(m) \ .$$

Proof of Corollary A.2. From Propositions A.6 and A.7, it comes that

$$\mathbb{E}(Z_m^2) \leq Z_m^2 \frac{1+\delta_n + \theta\sqrt{\frac{\Phi}{\xi}}\mathbb{1}_{\left(\sqrt{D_m}<(\log n)^2\right)}}{\left(1-\delta_n - L_m\right)^2} + r_n(m) \ .$$

Then, Lemma A.9 enables to conclude.

□

A.4. Dimension behavior with respect to n

Lemma A.4 (Oracle dimension)**.** *Let us assume* (**Bias**), (**Rich**), *and* (**RegD**) *hold true. Then, on the event* $\Omega' = \Omega_{\text{left}} \cap \Omega_{\text{right}}$, *where* Ω_{left} *and* Ω_{right} *are respectively defined in Corollary A.1 and Corollary A.2, it comes*

$$D_{m^*} \geq (\log n)^4 \ , \tag{34}$$

for large enough values of n.

Proof of Lemma A.4. Since $m^* = \text{Argmin}_m \|s - \widehat{s}_m\|^2$, it comes

$$\|s - \widehat{s}_{m^*}\|^2 \leq \|s - s_{m_0}\|^2 + \|s_{m_0} - \widehat{s}_{m_0}\|^2 \ ,$$

with m_0 defined by (**Rich**).

First on the event Ω', using $\mathbb{E}\left(Z_{m_0}^2\right) \leq \Phi D_{m_0}/n$ by (**RegD**) and Corollaries A.1 and A.2, there exists δ_n such that

$$\left|Z_{m_0}^2 - \mathbb{E}\left(Z_{m_0}^2\right)\right| \leq \delta_n \mathbb{E}\left(Z_{m_0}^2\right) \leq \delta_n \Phi \frac{D_{m_0}}{n} \ .$$

Then by use of (**Bias**) and (**Rich**) on Ω',

$$c_\ell D_{m^*}^{-\ell} \leq \|s - s_{m^*}\|^2 \leq \|s - \widehat{s}_{m^*}\|^2 \leq c_u n^{-u/2} + c_{rich}(1+\delta_n)\Phi n^{-1/2} \ ,$$

which is contradictory with assuming $D_{m^*} < (\log n)^4$ as long as n is large enough.

□

Lemma A.5 (Chosen model dimension)**.** *Let us assume* (**Bias**), (**Rich**), (**LoEx**), *and* (**RegD**) *hold true. Then with the notation of Lemma A.4, on the event* $\Omega = \Omega' \cap (\Omega_{\text{rem},1} \cap \Omega_{\text{rem},2})$, *where* $\Omega_{\text{rem},1}$ *and* $\Omega_{\text{rem},2}$ *are respectively defined in Proposition A.3 and Proposition A.4, it comes*

$$D_{\widehat{m}} \geq (\log n)^4 \ , \tag{35}$$

for large enough values of n.

Proof of Lemma A.4. For any model m such that $\widehat{R}_p(m) \leq \widehat{R}_p(m_0)$, Proposition A.3, Proposition A.4, and Proposition A.2 lead to

$$\begin{aligned}
&[\,1 - K(n,p)\delta_n\,]\,\|s - s_m\|^2 + \left[\,\frac{n}{n-p} - K(n,p)\delta_n - \left(K(n,p) + \frac{n}{n-p}\right)\delta_n\right]\mathbb{E}\left(Z_m^2\right) \\
&- K(n,p)\left[\,Z_m^2 - \mathbb{E}\left(Z_m^2\right)\right] \\
&\leq [\,1 + K(n,p)\delta_n\,]\,\|s - s_{m_0}\|^2 + \left[\,\frac{n}{n-p} + K(n,p)\delta_n + \left(K(n,p) + \frac{n}{n-p}\right)\delta_n\right]\mathbb{E}\left(Z_{m_0}^2\right) \\
&- K(n,p)\left[\,Z_{m_0}^2 - \mathbb{E}\left(Z_{m_0}^2\right)\right] \quad .
\end{aligned}$$

First, assuming $D_{\widehat{m}} < (\log n)^4$ on Ω and combining (**LoEx**) and (**RegD**) entail for $m = \widehat{m}$ that there exists a constant $C > 0$ such that

$$\left|\left[\,\frac{n}{n-p} - K(n,p)\delta_n - \left(K(n,p) + \frac{n}{n-p}\right)\delta_n\right]\mathbb{E}\left(Z_m^2\right) - K(n,p)\left[\,Z_m^2 - \mathbb{E}\left(Z_m^2\right)\right]\right| \leq C\frac{(\log n)^4}{n} \quad .$$

Second, using (**Bias**) provides

$$[\,1 - K(n,p)\delta_n\,]\,\|s - s_m\|^2 \geq [\,1 - K(n,p)\delta_n\,]\,c_\ell(\log n)^{-4\ell} \quad ,$$

which is larger than $C\frac{(\log n)^4}{n}$ for large enough values of n.

Using the same arguments as in Lemma A.4 for upper bounding the terms depending on m_0, it results that $D_{\widehat{m}} \geq (\log n)^4$ on Ω.

□

A.5. Technical results

Lemma A.6. *Let $X_1, \ldots, X_n$ be i.i.d. random variables defined on a mesurable space $(\mathcal{X}, \mathcal{T})$. Let S denote a set of real valued functions such that $\sup_{t\in S}\|t\|_\infty \leq b$ and $\sup_{t\in S}\mathrm{Var}\,(t(X_1)) = \sigma^2$. Let us assume* (**SqI**)*. Then*

$$\sigma^2 \leq \|s\|\, b \quad . \tag{36}$$

Furthermore, (**RegD**) *leads to*

$$\sigma^2 \leq \|s\|\,\sqrt{\Phi D} \quad , \tag{37}$$

where D denotes the dimension of the vector space S.

Lemma A.7. (**Dmax**) *implies*

$$\begin{aligned}
\|\phi_m\|_\infty &\leq \sqrt{\Phi\frac{n}{(\log n)^2}} \quad , \\
\mathrm{Var}\,(\phi_m(X_1)) &\leq \left(n\mathbb{E}(Z^2) + \|s\|^2\right)\sqrt{\Phi\frac{n}{(\log n)^2}} \quad .
\end{aligned}$$

Proof.

$$\begin{aligned}\mathrm{Var}\left(\phi_m(X_1)\right) &\leq \mathbb{E}\left[\,\phi_m^2(X_1)\,\right] \leq \|\phi_m\|_\infty \mathbb{E}\left[\,\phi_m(X_1)\,\right] \\ &= \|\phi_m\|_\infty \left(n\mathbb{E}(Z^2) + \|s_m\|^2\right) \leq \|\phi_m\|_\infty \left(n\mathbb{E}(Z^2) + \|s\|^2\right) \\ &\leq \left(n\mathbb{E}(Z^2) + \|s\|^2\right) \sqrt{\Phi \frac{n}{(\log n)^2}} \ .\end{aligned}$$

□

Lemma A.8. *Let us assume that* $0 \leq \delta_n + L_m$ *for every* $m \in \mathcal{M}$. *Then on the event* $\Omega_{\mathrm{left}} \cap \Omega_{\mathrm{right}}$ *(with* Ω_{left} *and* Ω_{right} *defined in Proposition A.6 and Proposition A.5 respectively), for every* $m, m' \in \mathcal{M}$,

$$Z_{m'}^2 \left(1 - 4(\delta_n + L_{m'})\right) \leq \mathbb{E}(Z_{m'}^2), \qquad Z_{m'}^2 - \mathbb{E}(Z_{m'}^2) \leq 4Z_{m'}^2(\delta_n + L_{m'}) \ . \tag{38}$$

and

$$\mathbb{E}(Z_m^2) \leq Z_m^2 \left(1 + 4(\delta_n + L_m)\right) + r_n, \qquad \mathbb{E}(Z_m^2) - Z_m^2 \leq 4Z_m^2(\delta_n + L_m) + r_n \ . \tag{39}$$

Proof of Lemma A.8.
Proof of (38) From Corollary A.1, on the event Ω_{right}, it comes

$$Z_m^2 \leq \left(\mathbb{E}(Z_m)\right)^2 \left(1 + \delta_n + L_m\right)^2 \ .$$

Then assuming moreover $0 \leq \delta_n + L_m < 1$, Jensen's inequality and $(1-x)^{-2} < (1+x^2)$ for $x \in [0,1[$ lead to

$$Z_m^2 \leq \mathbb{E}(Z_m^2) \left(1 + \delta_n + L_m\right)^2 \leq \mathbb{E}(Z_m^2) \frac{1}{\left(1 - \delta_n - L_m\right)^2} \ .$$

Finally if $0 \leq \delta_n + L_m < 1/4$, then

$$\mathbb{E}(Z_m^2) \geq Z_m^2 \left(1 - \delta_n - L_m\right)^2 \geq Z_m^2 \left[\,1 - 2(\delta_n + L_m)\,\right] \geq Z_m^2 \left[\,1 - 4(\delta_n + L_m)\,\right] \ .$$

Proof of (39) Assuming $\delta_n + L_m < 1/4$, Corollary A.2 and Lemma A.10 provide

$$\mathbb{E}(Z_m^2) \leq Z_m^2 \left(1 + 4(\delta_n + L_m)\right) + r_n(m) \ .$$

□

Lemma A.9. *For every* $a, b \in (0,1)$ *such that* $a < b(1-b)^{-1}$,

$$\frac{1+a}{\left(1-b\right)^2} \leq \frac{1}{\left(1-b\right)^3} \ . \tag{40}$$

Moreover if $0 < a = b < 1$, *then* $a < a(1-a)^{-1}$ *and Eq.* (40) *holds true.*

Lemma A.10. *For every interval* $I \subset [0,1[$ *such that* $0 \in I$, *there exists a constant* $\Delta > 3$ *such that*

$$\forall x \in I, \qquad (1-x)^{-3} \leq 1 + \Delta x \ .$$

In particular for $I = [0, 1/4]$, *this property holds true with* $\Delta = 4$. *Furthermore for every* $x \in]1, +\infty[$,

$$(1-x)^{-3} \leq 1.$$

A.6. Adaptivity in the minimax sense

A.6.1. Proof of Corollary 3.2

The proof simply consists in combining Theorems 3.1 and 3.2 by checking their assumptions. First, $s \in \mathcal{H}(L,\alpha)$ implies (**SqI**). Combined with Lemma A.11, it shows (**Bias**) is fulfilled. Besides, (**OrSp**) holds true since

$$\inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\, \|s-\widehat{s}_m\|^2 \,\right] \approx n^{-\frac{2\alpha}{2\alpha+1}} \quad \Rightarrow \quad \frac{n}{(\log n)^2} \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\, \|s-\widehat{s}_m\|^2 \,\right] \approx n^{\frac{1}{2\alpha+1}} (\log n)^{-2} \ ,$$

where $a \approx b$ means there exist constants $0 < c_1 \leq c_2$ such that $c_1 b \leq a \leq c_2 b$.

Second, since the model collection is built from regular partitions of $[0,1]$, (**RegD**) is clearly satisfied, and (**Dmax**) entails (**Rich**) is fulfilled.

A.6.2. Technical Lemma

Lemma A.11. *Let s be a density such that $s \in \mathcal{H}(L,\alpha)$ for some $\alpha \in (0,1]$ and $L > 0$. For every $D \in \mathbb{N}^*$, let s_D denote the orthogonal projection of s defined in Section 2.1.2 onto piecewise constant functions built from a given regular partition of $[0,1]$ in D intervals. Then,*

$$\frac{c_\ell}{D^\ell} \leq \|s-s_D\|^2 \leq \frac{c_u}{D^u} \ , \tag{41}$$

where $u = 2\alpha$, $c_u = L^2$, $\ell = 1 + 1/\alpha$ and $c_\ell = \epsilon^{2+1/\alpha} 2^{-(5+2/\alpha)} L^{-1/\alpha}$, for some $\epsilon > 0$.

Proof of Lemma A.11. First, let us notice (41) excludes $s = \mathbb{1}_{[0,1]}$. Then, there exist $x < y \in [0,1]$ such that $|x-y| \leq \eta$ and $|s(x)-s(y)| \geq \epsilon$ for some $\eta, \epsilon > 0$. Besides for a regular partition of $[0,1]$ in intervals $I_1, \ldots, I_D$ of Lebesgue measure $|I_k| = 1/D$, it comes

$$\|s-s_D\|^2 = \sum_{k=1}^{D} \int_{I_k} \left[\, s(t) - s_D(t) \,\right]^2 dt = \sum_{k=1}^{D} \int_{I_k} \left[\, s(t) - s_{I_k} \,\right]^2 dt \ ,$$

where s_{I_k} denotes the mean of s on interval I_k.

Second, let $K(\eta) = \{1 \leq k \leq D,\ I_k \cap [x,y] \neq \emptyset\}$ and $N(\eta)$ denote the cardinality of $K(\eta)$. Then, $N(\eta) \leq 2 + \eta D$. Combined with Lemma A.12, it leads to

$$\|s-s_D\|^2 \geq \sum_{k\in K(\eta)} \int_{I_k} \left[\, s(t) - s_{I_k} \,\right]^2 dt \geq \frac{1}{2^{4+1/\alpha} L^{1/\alpha}} \sum_{k\in K(\eta)} \Delta_k^{2+1/\alpha} \ ,$$

where $\Delta_k := \sup_{I_k} s - \inf_{I_k} s$, for every $1 \leq k \leq D$. Applying Hölder's inequality, it comes

$$\sum_{k\in K(\eta)} \Delta_k^{2+1/\alpha} \geq N(\eta)^{-(1+1/\alpha)} \left(\sum_{k\in K(\eta)} \Delta_k \right)^{2+1/\alpha} \geq N(\eta)^{-(1+1/\alpha)} \epsilon^{2+1/\alpha} \ ,$$

since $\sum_{k\in K(\eta)} \Delta_k \geq \epsilon$. Hence,

$$\begin{aligned}\|s - s_D\|^2 \geq \sum_{k\in K(\eta)} \int_{I_k} [\, s(t) - s_{I_k}\,]^2 \, dt &\geq \frac{1}{2^{4+1/\alpha} L^{1/\alpha}} \sum_{k\in K(\eta)} \Delta_k^{2+1/\alpha} \\ &\geq \frac{1}{2^{4+1/\alpha} L^{1/\alpha}} N(\eta)^{-(1+1/\alpha)} \epsilon^{2+1/\alpha} \\ &\geq \frac{1}{2^{4+1/\alpha} L^{1/\alpha}} (1+\eta)^{-(1+1/\alpha)} D^{-(1+1/\alpha)} \epsilon^{2+1/\alpha} \\ &\geq \frac{\epsilon^{2+1/\alpha}}{2^{5+2/\alpha} L^{1/\alpha}} D^{-(1+1/\alpha)} \ .\end{aligned}$$

□

Lemma A.12. *Let s denote a density defined on $[0,1]$ such that $s \in \mathcal{H}(L,\alpha)$, for some $L > 0$ and $\alpha \in (0,1]$. Let us define an interval $I \subset [0,1]$ and $s_I = |I|^{-1} \int_I s(t)\, dt$ denotes the mean of s on I. Then,*

$$\int_I (s(t) - s_I)^2 \, dt \geq \frac{\Delta^{2+1/\alpha}}{2^{4+1/\alpha} L^{1/\alpha}} \ ,$$

where $\Delta = \sup_I s - \inf_I s$.

Proof of Lemma A.12. First, let us notice $s^- = \inf_I s \leq s_I \leq \sup_I s = s^+$, which implies

$$\max\left(s^+ - s_I, s_I - s^-\right) \geq \Delta/2 \ .$$

Without loss of generality, let us assume $\max\left(s^+ - s_I, s_I - s^-\right) = s^+ - s_I$. Then $s^+ - s_I \geq \Delta/2$.

Second, let us introduce $x^+ \in I$ such that $s^+ = s(x^+)$. By continuity of s, there exists an interval $J \subset I$ such that $x^+ \in J$ and

$$\forall x \in J, \quad 0 \leq s(x^+) - s(x) \leq \Delta/4 \ .$$

Then,

$$\forall x \in J, \quad s(x) - s_I \geq \Delta/2 - \Delta/4 = \Delta/4 \ .$$

Moreover,

$$|J|\,(\Delta/2)^2 \leq \int_J \left(s(x^+) - s_I\right)^2 \, dx \leq \int_J \left(s(x^+) - s(x)\right)^2 \, dx \leq \int_J L^2 \left|x^+ - x\right|^{2\alpha} \, dx \leq |J|^{2\alpha+1} \, L^2 \ ,$$

which implies

$$|J| \geq \left(\frac{\Delta}{2L}\right)^{1/\alpha} \ .$$

Finally,

$$\int_I (s(x) - s_I)^2 \, dx \geq \int_J (s(x) - s_I)^2 \, dx \geq (\Delta/4)^2 \, |J| \geq (\Delta/4)^2 \left(\frac{\Delta}{2L}\right)^{1/\alpha} \ .$$

□

A.7. Identification point of view

Proposition A.8 (Bound on $\nu_n(s_m - s_{\bar{m}})$)**.** *Let us assume* (**Pol**),(**SqI**), (**RegD**) *hold true. Then, there exists a sequence* $(\delta_n)_{\mathbb{N}}$ *and an event* $\Omega_{\mathrm{rem},3}$ *with* $\mathbb{P}[\Omega_{\mathrm{rem},3}] \geq 1 - 2/n^2$*, on which for every* $m \in \mathcal{M}$*,*

$$2\left|\nu_n(s_m - s_{\bar{m}})\right| \leq \delta_n \left\|s_m - s_{\bar{m}}\right\|^2 + \delta_n \left\|s\right\| \sqrt{\Phi}\sqrt{\frac{D_m + D_{\bar{m}}}{n}} + \delta_n \Phi \frac{D_m + D_{\bar{m}}}{n} \ ,$$

with $\delta_n \to 0$ *and* $n\delta_n \to +\infty$ *as* $n \to +\infty$*, and* $0 \leq \delta_n \leq 1$ *for* n *large enough.*

Proof of Proposition A.8. Combining Eq. (31), (**SqI**), and (**RegD**) it comes for every $\eta > 0$,

$$2\nu_n(s_m - s_{\bar{m}}) \leq \eta \left\|s_m - s_{\bar{m}}\right\|^2 4\eta^{-1} \frac{\left\|s\right\| \sqrt{\Phi}\sqrt{(D_m + D_{\bar{m}})}}{n} x + 2\eta^{-1}\Phi\left(D_m + D_{\bar{m}}\right) \frac{x^2}{9n^2} \ .$$

with probability not larger than $2\exp(-x)$, for any $x > 0$.

Let us further assume that (**Pol**) holds true. Then with $x = x_m = (a_{\mathcal{M}} + 2)\log n$, it comes

$$\begin{aligned} &2\nu_n(s_m - s_{\bar{m}}) \\ &\leq \eta \left\|s_m - s_{\bar{m}}\right\|^2 + 4\eta^{-1} \frac{\left\|s\right\| \sqrt{\Phi}\sqrt{(D_m + D_{\bar{m}})}}{\sqrt{n}} (a_{\mathcal{M}} + 2) \frac{\log n}{\sqrt{n}} + 2\eta^{-1}\Phi \frac{D_m + D_{\bar{m}}}{n} \frac{\left((a_{\mathcal{M}} + 2)\log n\right)^2}{9n} \ . \end{aligned}$$

Let us choose $\eta = 1/\log n$, then there exists a sequence $(\delta_n)_{\mathbb{N}}$ with $\delta_n \to 0$ and $n\delta_n \to +\infty$ as $n \to +\infty$ such that.

$$2\nu_n(s_m - s_{\bar{m}}) \leq \delta_n \left\|s_m - s_{\bar{m}}\right\|^2 + \delta_n \left\|s\right\| \sqrt{\Phi}\sqrt{\frac{D_m + D_{\bar{m}}}{n}} + \delta_n \Phi \frac{D_m + D_{\bar{m}}}{n} \ .$$

Finally, let us notice that $\sum_{m \in \mathcal{M}} 2e^{-x_m} = \sum_{m \in \mathcal{M}} 2n^{-(a_{\mathcal{M}}+2)} \leq 2n^{a_{\mathcal{M}}} n^{-(a_{\mathcal{M}}+2)} = 2/n^2$.

□

Appendix B: Key concentration inequalities

Theorem B.1 (Bernstein's inequality)**.** *Let* $X_1, \ldots, X_n$ *be i.i.d. random variables defined on a mesurable space* $(\mathcal{X}, \mathcal{T})$*, and let* t *denote a mesurable bounded real valued function. Then for every* $x > 0$*,*

$$\mathbb{P}\left[\nu_n(t) > \sqrt{\frac{2\mathrm{Var}\left(t(X_1)\right)x}{n}} + \frac{\left\|t\right\|_\infty x}{3n}\right] \leq e^{-x} \ . \tag{42}$$

Theorem B.2 (Bousquet's version of Talagrand's inequality (Bousquet, 2002))**.**
Let $X_1, \ldots, X_n$ *be i.i.d. random variables defined on a mesurable space* $(\mathcal{X}, \mathcal{T})$*. Let* S *denote a set of real valued functions such that* $\sup_{t \in S} \left\|t\right\|_\infty \leq b$ *and* $\sup_{t \in S} \mathrm{Var}\left(t(X_1)\right) = \sigma^2$*. Denoting* $Z = \sup_{t \in S} \nu_n(t)$*, then for every* $x > 0$

$$\mathbb{P}\left[\sqrt{n}Z \leq \sqrt{n}\mathbb{E}(Z) + \sqrt{2\left(\sigma^2 + 2b\mathbb{E}(Z)\right)x} + \frac{bx}{3\sqrt{n}}\right] \leq e^{-x} \ . \tag{43}$$

Theorem B.3 (Rio's version of Talagrand's inequality (Klein and Rio, 2005))**.**
Let $X_1, \ldots, X_n$ *be i.i.d. random variables defined on amesurable space* $(\mathcal{X}, \mathcal{T})$*. Let* S *denote a set of real valued functions such that* $\sup_{t \in S} \left\|t\right\|_\infty \leq b$ *and* $\sup_{t \in S} \mathrm{Var}\left(t(X_1)\right) = \sigma^2$*. Denoting* $Z = \sup_{t \in S} \nu_n(t)$*, then for every* $x > 0$

$$\mathbb{P}\left[\sqrt{n}Z \leq \sqrt{n}\mathbb{E}(Z) - \sqrt{2\left(\sigma^2 + 2b\mathbb{E}(Z)\right)x} - \frac{8bx}{3\sqrt{n}}\right] \leq e^{-x} \ . \tag{44}$$

References

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

Arlot, S. (2007). V-fold penalization: an alternative to v-fold cross-validation. In *Oberwolfach Reports*, volume 4 of *Mathematisches Forschungsinstitut*. European Mathematical Society (EMS), Zürich. Report No.50/2007. Workshop: Reassessing the Paradigms of Statistical Model Building.

Arlot, S. (2008). Model selection by resampling penalization. *Electronic journal of Statistics*, 00:00.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Arlot, S. and Celisse, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning*, 10:245–279.

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with unknown variance. *The Annals of Statistics*, 37(2):630–672.

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

Barron, A. and Cover, T. M. (1991). Minimum Complexity Density Estimation. *IEEE transactions on information theory*, 37(4):1034–1054.

Bartlett, P., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1–3):85–113.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In Pollard, D., Torgensen, E., and Yang, G., editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. European Math. Soc.*, 3(3):203–268.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*.

Birgé, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram? *ESAIM Probab. Statist.*, 10:24–45.

Blanchard, G. and Massart, P. (2006). Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2664–2671.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 1:495–500.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall.

Burman, P. (1989). Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and the repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514.

Burman, P. (1990). Estimation of optimal transformation using v-fold cross-validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–245.

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 99.61, Université de Paris-Sud.

Castellan, G. (2003). Density estimation via exponential model selection. *IEEE transactions on information theory*, 49(8):2052–2060.

Celisse, A. (2008). *Model selection via cross-validation in density estimation, regression and change-points detection. (In English)*. PhD thesis, University Paris-Sud 11. `http://tel.archives-ouvertes.fr/tel-00346320/en/`.

Celisse, A. and Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368.

DeVore, R. and Lorentz, G. (1993). *Constructive Approximation*. Springer.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1):101–107.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328.

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, Berlin.

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

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, 22:45–55.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Li, K.-C. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975.

Lugosi, G. and Nobel, A. (1999). Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics. In Lindzey, G. and Aronson, E., editors, *Handbook of Social Psychology, Vol. 2*. Addison-Wesley.

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, J. (1993). Model Selection by Cross-Validation. *Journal of the American Statis. Association*, 88(422):486–494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264. With comments and a rejoinder by the author.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297.

Stone, C. J. (1985). An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA. Wadsworth.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147. 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.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563.

Tsybakov, A. B. (2003). *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications. Springer-Verlag.

Wegkamp, M. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yang, Y. (2007). Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.

Zhang, P. (1993). Model selection via multifold cross-validation. *The Annals of Statistics*, 21(1):299–313.