Submitted to the Annals of Statistics

OPTIMAL CROSS-VALIDATION IN DENSITY ESTIMATION WITH THE L^2-LOSS

By Alain Celisse

Laboratoire de Mathématiques Painlevé, UMR 8524 CNRS–Université Lille 1
MODAL *INRIA project-team*

We analyze the performance of cross-validation (CV) in the density estimation framework with two purposes: (*i*) *risk estimation* and (*ii*) *model selection*. The main focus is given to the so-called leave-p-out CV procedure (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. These expressions provide a great improvement upon V-fold cross-validation in terms of variability and computational complexity.

From a theoretical point of view, closed-form expressions also enable to study the Lpo performance in terms of risk estimation. The optimality of leave-one-out (Loo), that is Lpo with $p = 1$, is proved among CV procedures used for risk estimation. Two model selection frameworks are also considered: *estimation*, as opposed to *identification*. For estimation with finite sample size n, optimality is achieved for p large enough (with $p/n = o(1)$) to balance the overfitting resulting from the structure of the model collection. For identification, model selection consistency is settled for Lpo as long as p/n is conveniently related to the rate of convergence of the best estimator in the collection: (*i*) $p/n \to 1$ as $n \to +\infty$ with a parametric rate, and (*ii*) $p/n = o(1)$ with some nonparametric estimators. These theoretical results are validated by simulation experiments.

1. Introduction. For estimating a target quantity denoted by s, let $\{S_m\}_{m\in\mathcal{M}}$ denote a collection of sets of candidate parameters indexed by $\mathcal{M}$. 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 Akaike (1973); Mallows (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 (Ledoux, 2001; Talagrand, 1996) and aims at deriving *oracle inequalities* such as

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq C \inf_{m\in\mathcal{M}} \left\{\ell\left(s, \widehat{s}_m\right)\right\} + r_n \tag{1}$$

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

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 $\inf_m \ell(s, \widehat{s}_m)$, 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 $C = C_n \to 1$ 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 selection consistent* (see Shao, 1997, for a study of various model selection procedures in terms of model selection consistency).

In the density estimation framework, model selection with deterministic penalties has been addressed: (i) for Kullback-Leibler divergence by Barron et al. (1999); Castellan (1999, 2003); Yang and Barron (1998) 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).

The aforementioned approaches rely on some deterministic penalties such as AIC or BIC. These penalties are derived in some specific settings (for instance a Gaussian noise is assumed by Birgé and Massart, 2006) and remain unjustified and even sometimes misleading in more general settings.

Conversely, cross-validation (CV) is a *resampling* procedure based on a *universal heuristics* which makes it applicable in a wide range of settings. CV procedures 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 procedures can be computationally demanding or even intractable, Bowman (1984); Rudemo (1982) 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 procedures 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 procedure in terms of risk estimation. Several papers are dedicated to show the equivalence between some CV procedures and penalized criteria in terms of asymptotic optimality properties: (*i*) *efficiency* in Li (1987); Zhang (1993), and (*ii*) *model selection consistency* in Shao (1997); Yang (2006, 2007). Let us notice that in the parametric setting, Yang (2007) proved that efficiency and model selection consistency are contradictory objectives that cannot be achieved simultaneously. We refer interested readers to Shao (1997) for an extensive review about asymptotic optimality properties in terms of efficiency and model selection consistency of some penalized criteria as well as CV procedures.

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 such result 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.

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 considerably improve upon V-FCV in terms of (i) variability (Celisse and Robin, 2008), and (ii) computational complexity (Section 2.3). A second improvement allowed by these formulas is the deep new understanding of the theoretical performance of CV in two respects: first for risk estimation (Section 2.4), and second for model selection (Section 3). For instance, it is proved that Loo is the best CV procedure for risk estimation (Theorem 2.1), while the story can be different for model selection (Corollary 3.1 and Theorems 3.3 and 3.4).

In Section 3, two aspects of model selection *via* CV have been explored. The *estimation* point of view is described in Section 3.1. 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. This phenomenon is supported by simulation experiments detailed in Section 3.1.4. Finally Section 3.2 deals with the *identification* point of view. CV is proved to be model selection consistent in various settings where the choice of p is related to the convergence rate (parametric and nonparametric) of the best estimator one tries to recover. Simulation results illustrate these different behaviors in Section 3.2.2. Finally, a discussion is provided in Section 4 to give some guidelines towards a better understanding of CV procedures. The main proofs have been postponed to the appendix. For reasons owing to space constraints, more technical ones are provided in the supplementary material (from Section S.1 and following ones).

2. Cross-validation and risk estimation.

2.1. *Statistical framework.*

2.1.1. *Notation.* Throughout the paper, $X_1, \dots, X_n \in [0,1]$ 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 $X_{1,n} = (X_1, \dots, X_n)$.

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

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

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

$$P_n\gamma(u) := \frac{1}{n}\sum_{i=1}^{n}\gamma\left(u; X_i\right), \quad \text{where} \quad P_n = 1/n\sum_{i=1}^{n}\delta_{X_i} \tag{3}$$

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

For a collection of models $\{S_m\}_{m\in\mathcal{M}_n}$ indexed by a countable set $\mathcal{M}_n$, the *empirical contrast minimizer* is defined by

$$\widehat{s}_m := \text{Argmin}_{u\in S_m} P_n\gamma(u). \tag{4}$$

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

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

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*

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

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

It is straightforward to check that the empirical contrast minimizer over $S_m = \text{Span}\left(\varphi_\lambda,\ \lambda \in \Lambda(m)\right)$, defined by Eq. (4), is a projection estimator since

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\ \varphi_\lambda, \qquad \text{with} \qquad P_n\varphi_\lambda = \frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i). \tag{5}$$

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 $D_m = \text{Card}(\Lambda(m))$ 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,

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\mathbb{1}_{I_\lambda}\,\frac{\mathbb{1}_{I_\lambda}}{|I_\lambda|}. \tag{6}$$

- *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 $\Lambda(m) \subset \mathbb{Z}$,

$$\widehat{s}_m(t) = \sum_{\lambda \in \Lambda(m)} P_n \varphi_\lambda \, e^{2\pi i \lambda t}, \qquad \forall t \in [0,1] \tag{7}$$

  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 $\Lambda_n = \left\{(j,k) \mid j \in \mathbb{N}^* \text{ and } 1 \leq k \leq 2^j\right\}$. Then for every subset $\Lambda(m)$ of Λ_n,

$$\widehat{s}_m = \sum_{\lambda \in \Lambda(m)} P_n \varphi_\lambda \, \varphi_\lambda. \tag{8}$$

Some of these estimators can take negative values. A possible solution is truncating and normalizing the preliminary projection estimator

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

However the closed-form expressions provided in Section 2.3 are not available for these truncated and normalized estimators.

2.2. *Leave-p-out cross-validation.* In the literature, several cross-validation (CV) procedures have been successively introduced to overcome the defects of already existing ones. Let us describe the main CV procedures with some emphasis to computational aspects.

2.2.1. *Cross-validation.* For $1 \leq p \leq n-1$, let us define $\mathcal{E}_p = \{e \subset \{1, \ldots, n\}, \ \mathrm{Card}(e) = p\}$ and for $e \in \mathcal{E}_p$, set $X^e = \{X_i, \ i \in e\}$ (test set) and $X^{(e)} = \{X_i, \ i \in \{1, \ldots, n\} \setminus e\}$ (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 respectively defined 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

$$\widehat{R}_{\mathrm{Ho},p}(\widehat{s}) := P_n^e \gamma\left(\widehat{s}(X^{(e)})\right) = \frac{1}{p} \sum_{i \in e} \gamma\left(\widehat{s}(X^{(e)}); X_i\right). \tag{9}$$

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

$$\widehat{R}_p(\widehat{s}) = \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_p} P_n^e \gamma\left(\widehat{s}(X^{(e)})\right). \tag{10}$$

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 becomes intractable as n grows.

V-fold cross-validation. To overcome the high computational burden of Lpo (Eq. (10)), 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 $X_1, \ldots, X_n$ into V subsets $X^{e_1}, \ldots, X^{e_V}$ of approximately equal size $p = n/V = \mathrm{Card}(e_i)$, $i = 1, \ldots, V$. Every X^{e_i}, $i = 1, \ldots, V$ is successively used as a test set leading to the V-fold risk estimator of $R_n(\widehat{s})$

$$\widehat{R}_{\mathrm{V-FCV}}(\widehat{s}) = \frac{1}{V} \sum_{v=1}^{V} P_n^{e_v} \gamma\left(\widehat{s}(X^{(e_v)})\right). \tag{11}$$

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. Several surrogates of Lpo have been proposed such as V-FCV and the *repeated learning-testing cross-validation* (Breiman et al., 1984; Burman, 1989; Zhang, 1993). 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. 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.

2.3. *Closed-form expressions for the Lpo risk estimator.* Closed-form formulas for the Lpo estimator are proved in the present section, which makes Lpo fully effective in practice and better than V-FCV. Such formulas also enable a more accurate theoretical analysis of CV procedures 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. (5). 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. In the present section, proofs have been deferred to

Section S.1 (Supplementary material). Recalling the expression of the contrast $\gamma(\cdot\,;\cdot)$ (Eq. (2)), 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. (5) and set $X^e = \{X_i,\ i \in e\}$ for every $e \in \mathcal{E}_p$. Then for every $p \in \{1, \dots, n-1\}$,*

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

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

Lemma 2.1 enables to derive closed-form formulas for the Lpo risk estimator, which makes Lpo procedure 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. (5). Then for every $p \in \{1, \dots, n-1\}$,*

$$\widehat{R}_p(m) = \widehat{R}_p(\widehat{s}_m) = \frac{1}{n(n-p)} \sum_{\lambda \in \Lambda(m)} \left[\sum_{j=1}^{n} \varphi_\lambda^2(X_j) - \frac{n-p+1}{n-1} \sum_{j \neq k} \varphi_\lambda(X_j)\varphi_\lambda(X_k) \right]. \tag{12}$$

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 (12) 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).

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

COROLLARY 2.1 (Histograms). *For $\widehat{s}_m$ given by Eq. (6) 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 $n_\lambda = \mathrm{Card}\left(\{i \,|\, X_i \in I_\lambda\}\right)$.

COROLLARY 2.2 (Trigonometric polynomials). *For every $k \in \mathbb{N}$, let φ_λ denote either $t \mapsto \cos(2\pi kt)$, if $\lambda = 2k$ or $t \mapsto \sin(2\pi kt)$, 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 $\beta(n,p) = (n-p+1)\left[\,n(n-1)(n-p)\,\right]^{-1}$.

COROLLARY 2.3 (Haar basis). *Let us define* $\varphi(\cdot) = \mathbb{1}_{[0,1]}(\cdot)$ *and* $\varphi_{j,k}(\cdot) = 2^{j/2}\varphi(2^j \cdot -k)$, *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 formula (12), one derives closed-form expressions for the expectation and variance of the Lpo risk. These expressions allow to analyze the theoretical behavior of CV in terms of risk estimation and model selection (see Section 3). In the present section we 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.* (5). *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

$$\mathrm{Var}\left[\widehat{R}_p(m)\right] = \frac{1}{(n-1)^2}\left[a_n + \frac{b_n}{(n-p)} + \frac{c_n}{(n-p)^2}\right], \tag{13}$$

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 (13), 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 $\mathbb{B}\left[\widehat{R}_p(m)\right] := \mathbb{E}\widehat{R}_p(m) - \mathbb{E}\left[\|\widehat{s}_m\|^2 - 2\int_{[0,1]} s\,\widehat{s}_m\right]$ 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.* (5). *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 procedures. If $p = p_n$ satisfies $p_n/n \xrightarrow[n\to+\infty]{} q \in [0,1)$, then $\mathbb{B}\left[\widehat{R}_p(m)\right] \xrightarrow[n\to+\infty]{} 0$, and 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 $1 \le p \le n-1$,*

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

where the big $O(\cdot)$ does not depend on p, but depends on S_m and P, and

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

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

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

In the more specific case of histogram and kernel density estimators, Celisse and Robin (2008) derived a similar (non asymptotic) result for the variance.

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 $A \ge 0$. However in full generality, the sign of B is unknown. The following proposition relates the monotonicity of $p \mapsto \mathrm{Var}\left[\widehat{R}_p(m)\right]$ to this sign.

PROPOSITION 2.3. *Let us define $p_{0,n} = \mathrm{Argmin}_{1\le p\le n-1}\mathrm{Var}\left[\widehat{R}_p(m)\right]$ in Eq. (13). Then,*

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

where the little $o(\cdot)$ only depends on S_m and P. Furthermore, if

$$2\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right] \ge \mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_3)\right], \tag{14}$$

$p \in \{1,\dots,n-1\} \mapsto \mathrm{Var}\left[\widehat{R}_p(m)\right]$ is increasing. 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. (14) is related to the sign of B (Corollary 2.5) and to the minimum location value $p_{0,n}$. If it holds true, then $p_{0,n} \notin \{2, \dots, n-1\}$, which means Loo has the smallest variance among CV procedures. In particular let us notice (14) holds true with any density estimated by regular histograms since $\sum_\lambda \varphi_\lambda^2(X_1)$ is then a constant and the covariance in the left-hand side is a variance by independence of X_1, X_2, and X_3. On the contrary, (14) is not fulfilled when using histograms based on a partition such that $\mathbb{P}[X \in I_\lambda] = C < 1/2$ for every λ, where C denotes a constant.

We are now in position to provide the main result of this section, which describes the behavior of $\widehat{R}_p$ as a risk estimator in terms of mean-square error (MSE).

THEOREM 2.1. *For every $m \in \mathcal{M}_n$, let us define the* 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 $p \in \{1, \dots, n-1\}$.*

1. *If (14) 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 $p = p_n \in \{1, \dots, n-1\}$ such that $\limsup_{n\to+\infty} p_n/n < 1$, then*

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

where A is given in Corollary 2.5 and the big $O(\cdot)$ depends on S_m and P.

If (14) holds true, Loo is the best CV procedure in terms of MSE when estimating the risk of an estimator. Otherwise as long as $\limsup_{n\to+\infty} p_n/n < 1$, choosing a value of $p \neq 1$ is useless since any value in $\{1, \dots, n-1\}$ asymptotically leads to the same performance in terms of MSE. Therefore since Loo has the smallest bias (Corollary 2.4), *Loo is optimal among CV procedures 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. From Section 2.4, Loo is proved to be the best CV procedures in the context of risk estimation. However the best procedure for risk estimation is not necessarily the best one for *model selection*. Although the empirical risk $P_n\gamma(\widehat{s}_m)$ (4) is a reliable estimator of $\mathbb{E}[P_n\gamma(\widehat{s}_m)]$, using empirical risk 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.

3.1. *Optimal cross-validation for Estimation.* The performance of CV with respect to p is first characterized through a *sharp oracle inequality* (Theorem 3.1). A leading constant converging to 1 as $n \to +\infty$ is achieved for some values of p, highlighting the asymptotic optimality of corresponding CV procedures. From a theoretical point of view, Corollary 3.1 explores the link between (a proxy to) the optimal p and influential quantities related to the difficulty of the estimation problem for finite sample size. These results are further validated 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

$$\begin{aligned} m^* &:= \operatorname{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m) - P\gamma(s) = \operatorname{Argmin}_{m\in\mathcal{M}_n} \|s-\widehat{s}_m\|^2 \\ &= \operatorname{Argmin}_{m\in\mathcal{M}_n} P\gamma(\widehat{s}_m). \end{aligned}$$

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

$$\widehat{m}(p) := \operatorname{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(m), \tag{15}$$

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

3.1.2. *Main oracle inequality.* Let us introduce some notation and detail the main assumptions used along the following sections.

Square-integrable density:.

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

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

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

(Pol) $$\operatorname{Card}(\mathcal{M}_n)\le n^{a_{\mathcal{M}}}.$$

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

Model regularity:.

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

It relates the regularity of the orthonormal basis (measured in terms of sup-norm) to the dimension of the model. For instance using (6), (**RegD**) requires $|I_\lambda|\ge(\Phi D_m)^{-1}$ for every $\lambda\in\Lambda(m)$. The length of intervals I_λ cannot be too different from one another to some extent.

Maximal dimension:.

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

In the sequel, $\Gamma=1$ is always considered to simplify expressions. Note that proofs and conclusions remain unchanged with this particular choice.

Estimation error and dimension:.

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

This assumption makes the estimation error $\mathbb{E}\left(\|s_m - \widehat{s}_m\|^2\right)$ and D_m comparable. For instance, Lemma S.6 (supplementary material) proves (**LoEx**) is fulfilled with Hölder densities estimated by regular histograms defined by (6) such that $|\{x \in [0,1] \mid s(x) \geq \eta\}| \geq \ell$ for some $\eta \in (0,1)$, where $0 < \ell < 1$ satisfies

$$\ell > \left(\inf_{m \in \mathcal{M}_n} D_m\right)^{-1}.$$

Note the latter inequality amounts to exclude too small models for which the support of s is included in one single interval I_λ.

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

$$\textbf{(Rich)} \qquad \sqrt{n} \leq D_{m_0} \leq c_{rich}\sqrt{n}.$$

This requirement is rather mild since one can add such a model in our collection.

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

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

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 (6) 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:.

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

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 $\alpha \in (0,1]$ estimated by regular histograms (see Section S.5).

The performance of the Lpo estimator with respect to p is described by the following oracle inequality from which the CV optimality is deduced for some values of p. The proof is given in Appendix A.1.

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* $\delta_n \to 0$, *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 $K(n,p) = 1 + \frac{2}{n-1} + \frac{p}{n-p}\frac{1}{n-1}$.

If $p/n \to 0$ then $C_n(p) \to 1$ as $n \to +\infty$, which leads to *efficient* (asymptotically optimal) model selection procedures. In particular this holds true for $p = 1$ that is, *Loo is asymptotically optimal* since

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

From the proof, it also arises the slowly decreasing sequence δ_n is related to the model collection structure. An increase of $\mathcal{M}_n$ makes the model selection problem more difficult and δ_n larger.

While *asymptotic optimality* is deduced from Theorem 3.1 for any CV procedure as long as $p = o(n)$, it is also desirable to analyze the performance of CV as p depends on the finite sample size. From a theoretical point of view, this will provide the rate at which p/n has to decrease to 0 to reach efficiency. Based on Figure 1 (Panel (c)) where $C_n(p)$ appears as a reliable proxy to the optimal $C_{or,n}(p)$ (given by Eq. (19)), we suggest to optimize $C_n(p)$ with respect to p to get a surrogate optimal p depending on influential parameters such as n and δ_n. This strategy has been validated by simulation experiments of Section 3.1.4. The following Corollary 3.1 proves the best (surrogate) p/n slowly decreases to 0.

Corollary 3.1 (Optimizing upper bound). *With the notation and assumptions of Theorem 3.1, the constant* $C_n(p)$ *is minimized over* $p \in \{1, \dots, n-1\}$ *for*

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

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

$$p_n^* \sim_{+\infty} 10 n \delta_n \xrightarrow[n \to +\infty]{} +\infty. \tag{16}$$

The proof has ben deferred to Appendix A.1. Corollary 3.1 describes the rate (up to constant) at which $p = p_n^*$ has to grow with n to achieve finite-sample optimality. In particular p_n^*/n in (16) is related to δ_n that is strongly connected to the structure of the model collection as explained in the following of Theorem 3.1. A more complex collection leads to a larger δ_n and then to a larger optimal p_n^*. In other words, p must be chosen large enough to balance the overfitting induced by the structure of the model collection. This phenomenon is observed in practice in the simulation experiments of Section 3.1.4 (Figure 2).

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 (1) is somewhat classical. Here the novelty is first that CV enjoys such a desirable property as a model selection procedure, 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 start providing a general theorem from which any adaptivity result will be immediate corollary. The proof is given in Appendix A.1.

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 $p \in \{1,\dots,n-1\}$. Then for every $1 \le p \le n-1$,*

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right] \le C_n(p)\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\|s - \widehat{s}_m\|^2\right] + (\Phi + \|s\|^2)\frac{12}{n(\log n)^2} + \frac{6c_u}{n^2}, \tag{17}$$

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 (17) are remainder terms by Assumptions (**RegD**), (**Dmax**), and (**Bias**).

Applying Theorem 3.2 to the collection of regular histograms defined by (6), 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 (6) 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 there exist constants $0 < K_\alpha^- \leq K_\alpha^+$ such that for every $p = o(n)$,

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

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 $p = o(n)$.

The proof has been deferred to Section S.5. 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, which has been stated by Ibragimov and Khas'minskij (1981). The main contribution of this result is to prove $p = o(n)$ leads to adaptivity. Note that similar results can also be proved for Besov balls $\mathcal{B}^\alpha_{\infty,2}(L)$, with $\alpha, L > 0$ for instance (see DeVore and Lorentz, 1993), by using an appropriate collection of models such as trigonometric polynomials defined by (7).

3.1.4. *Simulation experiments.* Results of simulation experiments are provided to check the conclusions drawn (from theory) in Section 3.1.2. A mixture of Beta distributions

$$\forall x \in [0,1], \qquad s(x) = \frac{\beta(3,7;x) + \beta(10,5;x)}{2} \tag{18}$$

has been used to generate samples of size $n = 100, 500, 1000, 2000, 3000, 4000, 5000, 10\,000, 20\,000$. Note that (18) defines a Hölder density on $[0,1]$. For each n, every $p \in \{1,\dots,n-1\}$ have been considered and (**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 (6). Only regular histograms with dimension $D_m \geq 2$ are used so that (**LoEx**) holds true (Lemma S.6). For every $1 \leq p \leq n-1$, $\widehat{m}(p)$ is defined by (15).

Let us also introduce

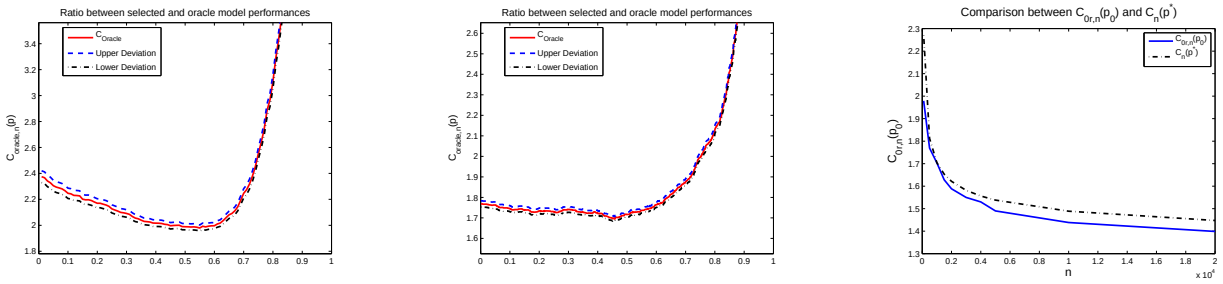
$$C_{or,n}(p) := \mathbb{E}\left[\frac{\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2}{\inf_{m\in\mathcal{M}_n}\left\{\left\|s - \widehat{s}_m\right\|^2\right\}}\right] \quad \text{and} \quad p_0 := \operatorname{Argmin}_{1\leq p\leq n-1} C_{or,n}(p), \tag{19}$$

which measures the average performance of $\widehat{s}_{\widehat{m}(p)}$ with respect to that of $\widehat{s}_{m^*}$ (oracle estimator). The closer $C_{or,n}(p)$ to 1, the better $\widehat{s}_{\widehat{m}(p)}$. Minimizing $C_{or,n}(p)$ as a function of p for various values of n enables 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 to $C_{or,n}(p)$. For each curve $p \mapsto C_{or,n}(p)$, a confidence band has been displayed. It is delimited by $p \mapsto C^-_{or,n}(p)$ and $p \mapsto C^+_{or,n}(p)$ respectively defined by

$$C^-_{or,n}(p) = C_{or,n}(p) - \frac{\widehat{\sigma}}{\sqrt{N}}, \quad \text{and} \quad C^+_{or,n}(p) = C_{or,n}(p) + \frac{\widehat{\sigma}}{\sqrt{N}}, \tag{20}$$

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



(a) (b) (c)

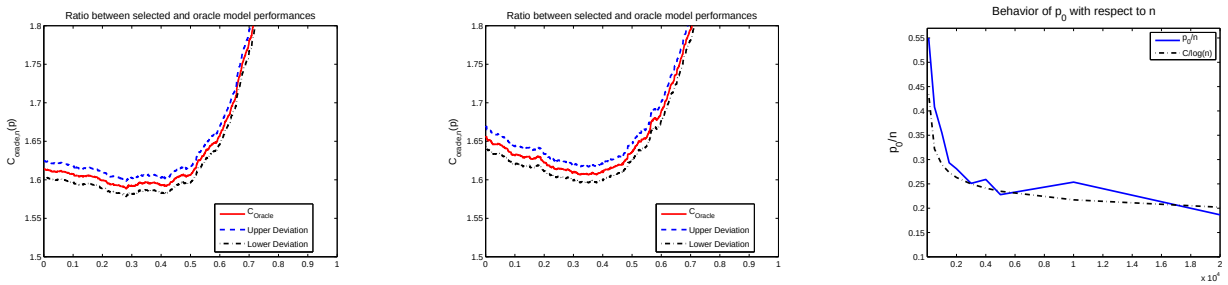
FIG 1. *Panels (a) and (b): $p/n \mapsto C_{or,n}(p)$ (plain red line) is plotted for $\Gamma = 1$ (see* (**Dmax**)*) and different values of n: (a) $n = 100$, and (b) $n = 1000$. $p/n \mapsto C^+_{oracle,n}(p)$ (blue dashed line) and $p/n \mapsto C^-_{oracle,n}(p)$ (black dot-dashed line) have been plotted on the same graph as well (see* (20)*). Panel (c): $n \mapsto C_{or,n}(p_0)$ (plain blue line) and $n \mapsto C_n(p^*)$ (black dot-dashed line) are displayed. $N = 1000$ samples have been drawn from the mixture of Beta distributions* (18).

First from Panels (a) and (b) of Figure 1, $C_{or,n}(p)$ (plain red lines) decreases pointwise as n grows. This is confirmed by Panel (c) of Figure 1 at the particular value $p = p_0$ as n grows. This is in accordance with Theorem 3.1 and $C_n(p) \to 0$ as n increases. Second the optimization strategy at the basis of Corollary 3.1 is empirically validated by Panel (c) of Figure 1 where $C_{or,n}(p_0)$ and its proxy $C_n(p_0)$ remain very close to each other. Furthermore the optimal rate derived in Corollary 3.1 is supported up to constant by simulation results displayed in Panel (c) of Figure 2 where p_0/n is almost equal to the predicted $\delta_n \approx C/(\log n)$ ($C > 0$) from the proof of Theorem 3.1.

The conclusion of Corollary 3.1 about the dependence of the optimal p_0 on the complexity of the model collection (through δ_n) is also illustrated by Panels (a) and (b) in Figure 2 where Γ (**Dmax**) respectively equals 1 and 2. As Γ grows the model collection becomes more complex, leading to a worse performance and a larger p_0 in Panel (b). The need for a larger p_0 is all the more strong as the curve in Panel (b) is less flat than in Panel (a), indicating the problem becomes more difficult as Γ increases and any misspecification of p_0 leads to a stronger loss in accuracy. One concludes the more complex the model collection, the larger the optimal p.

Note that this conclusion does not apply to Loo ($p = 1$) (see (16)), suggesting Loo may be suboptimal for finite sample size. This is supported by Figure 1 (panels (a) and (b)) and Figure 2 (panels (a) and (b)) where the minimum of each curve is not reached at $p = 1$.

3.2. *Optimal cross-validation for Identification.* 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.



(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* (20)*). $N = 1000$ samples have been drawn from the mixture of Beta distributions* (18)*. For (c), $n \mapsto p_0/n$ (blue plain line) and $n \mapsto C/(\log n)$ (black dot-dashed line) are displayed, where p_0 denotes the minimizer of $C_{or,n}(p)$ as a function of p and C is a constant.*

The purpose is now to recover the best model denoted by $S_{\bar{m}}$ and defined by

$$\bar{m} := \mathrm{Argmin}_{m \in \mathcal{M}_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2 \right], \tag{21}$$

where $\bar{m}$ is a deterministic quantity unlike m^* from Section 3.1. Since this goal cannot be reached if other models can perform as well as $S_{\bar{m}}$ (even asymptotically), one also requires there exist $\mu > 0$ and $n_0 \in \mathbb{N}^*$ such that for every integer $n > n_0$,

$$(1+\mu)\mathbb{E}\left[\|s - \widehat{s}_{\bar{m}}\|^2 \right] \leq \inf_{m \in \mathcal{M}_n \setminus \{\bar{m}\}} \mathbb{E}\left[\|s - \widehat{s}_m\|^2 \right]. \tag{**BeMo**}$$

A similar assumption (in probability rather than in expectation) has been made by Yang (2007). Let us further assume the collection $\{S_m\}_{m \in \mathcal{M}_n}$ can be split into

- *parametric models* indexed by $\mathcal{M}_{n,P}$ for which there exist constants $\pi, \tau > 0$ (independent of n) such that

$$\sup_{m \in \mathcal{M}_{n,P}} \left\{ n\mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right] \right\} \leq \pi \ , \quad \text{and} \quad \inf_{m \in \mathcal{M}_{n,P}, s \notin S_m} \left\{ \|s - s_m\|^2 \right\} \geq \tau. \tag{22}$$

- *nonparametric models* indexed by $\mathcal{M}_{n,NP}$ such that

$$n\,(\log n)^{-2} \inf_{m \in \mathcal{M}_{n,NP}} \mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right] \xrightarrow[n \to +\infty]{} +\infty. \tag{23}$$

Then,

$$\{S_m\}_{m \in \mathcal{M}_n} = \{S_m\}_{m \in \mathcal{M}_{n,P}} \cup \{S_m\}_{m \in \mathcal{M}_{n,NP}}. \tag{**P-NP**}$$

Parametric models are models with convergence rate of order $1/n$. Since $\mathbb{E}\left[\|s - \widehat{s}_m\|^2 \right] \approx \|s - s_m\|^2 + C \cdot D_m/n$, allowing D_m to depend on n makes the rate of the corresponding model

slower than $1/n$ (nonparametric model). Consistently with this remark, (22) requires the largest dimension over parametric models is bounded by a constant independent of n, and that the bias of parametric models such that $s \notin S_m$ cannot decrease with n toward 0. Otherwise, such a model would be nonparametric. Conversely (23) only requires the dimension of nonparametric models must be larger than $(\log n)^2$. In particular, this does not prevent nonparametric models from containing s or having their bias decreasing to 0 as n grows.

3.2.1. *Main results.* Depending on whether s belongs or not to $\cup_{m\in\mathcal{M}_n} S_m$, the two following results prove model selection consistency for CV. Their main contribution is to relate the cardinality p of the test set to the rate of convergence of $\widehat{s}_{\bar{m}}$ and the model collection complexity. Note that in addition, the model consistency property is settled with a collection of models allowed to grow with n, which contrasts with earlier results (see for instance Yang, 2007).

Let us start with the setting where s belongs to $\cup_{m\in\mathcal{M}_n} S_m$, which implies the best estimator $\widehat{s}_{\bar{m}}$ achieves the parametric rate $1/n$.

THEOREM 3.3 (Model consistency with $s \in \cup_m S_m$). *Let $\cup_{m\in\mathcal{M}_n} S_m$ denote a collection of models satisfying* (**Pol**) *and* (**P-NP**), $\bar{m} \in \mathcal{M}_n$ *given by* (21) *be such that* (**BeMo**) *holds true, and assume* (**SqI**), (**RegD**), (**Dmax**), *and* (**LoEx**). *For every $1 \le p \le n-1$, let us also define $\widehat{m} = \widehat{m}(p) = \operatorname{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(m)$. If the target $s \in \cup_{m\in\mathcal{M}_n} S_m$, then every $1 \le p = p_n \le n-1$ such that*

$$\log(n)\left(1-\frac{p}{n}\right) \xrightarrow[n\to+\infty]{} 0, \qquad \text{and} \qquad n\left(1-\frac{p}{n}\right) \xrightarrow[n\to+\infty]{} +\infty, \tag{24}$$

leads to

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

The proof has been deferred to Appendix B.1. When s belongs to $\cup_{m\in\mathcal{M}_n} S_m$, the best estimator $\widehat{s}_{\bar{m}}$ in a polynomial collection can be recovered by CV provided p/n converges to 1 as n tends to $+\infty$. The proof establishes this rate (i) cannot exceed $1/n$ to allow distinguishing between parametric estimators (with convergence rate of order $1/n$), and (ii) has to be faster than $(\log n)^{-1}$ to allow dealing with the polynomial complexity of the model collection. For instance a finite collection would lead to replace the $(\log n)^{-1}$ rate by a slower one determined by the control level of $\mathbb{P}\left[\,\widehat{m} = \bar{m}\,\right]$. In the regression setting (Yang, 2007) already proved requiring $p/n \to 1$ enables to recover the best parametric estimator among parametric ones (see Corollary 1), while this requirement is no longer necessary when comparing parametric and nonparametric estimators. Our result is consistent with Yang's one, although our setting is somewhat different since we compare the best parametric estimator with both parametric and nonparametric ones in the same time.

Conversely when s does not belong to $\cup_m S_m$, every parametric model is biased according to (22) and $\widehat{s}_{\bar{m}}$ reaches a nonparametric rate, that is $nR_n(\bar{m}) \to +\infty$ as n tends to $+\infty$.

THEOREM 3.4 (Model consistency with $s \notin \cup_m S_m$). *Let $\cup_{m\in\mathcal{M}_n} S_m$ denote a collection of models satisfying* (**Pol**) *and* (**P-NP**), $\bar{m} \in \mathcal{M}_n$ *given by* (21) *be such that* (**BeMo**) *holds true,*

and assume (**SqI**), (**RegD**), (**Dmax**), *and* (**LoEx**). *For every* $1 \leq p \leq n-1$, *let us also define* $\widehat{m} = \widehat{m}(p) = \operatorname{Argmin}_{m \in \mathcal{M}_n} \widehat{R}_p(m)$. *Let us assume the target* $s \notin \cup_{m \in \mathcal{M}_n} S_m$ *and* $R_n(\bar{m}) \to 0$ *as* n *tends to* $+\infty$.

1. *If for large enough values of* n $D_{\bar{m}} \leq (\log n)^4$, *then every* $1 \leq p = p_n \leq n-1$ *such that*

$$\log(n)\left(1 - \frac{p}{n}\right) \xrightarrow[n \to +\infty]{} 0 \qquad \text{and} \qquad n\mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right] = o\,(n-p) \tag{25}$$

leads to

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

2. *If for large enough values of* n $D_{\bar{m}} > (\log n)^4$, *then every* $1 \leq p = p_n \leq n-1$ *such that*

$$\frac{(\log n)^5}{n\mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right]} = o\left(\frac{p/n}{1-p/n}\right) \qquad \text{and} \qquad \frac{p/n}{1-p/n} = o\left(1 \vee \frac{\| s - s_{\bar{m}} \|^2}{\mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right]}\right) \tag{26}$$

leads to

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

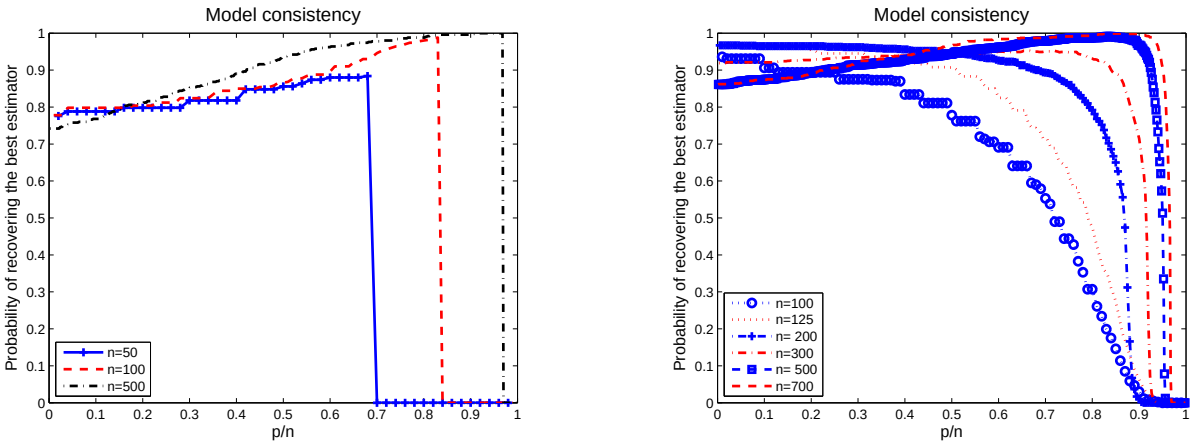
The proof is similar to that of Theorem 3.3 and has been postponed to Section S.6 (supplementary material). The constraints on p strongly depend on the rate of convergence of $\widehat{s}_{\bar{m}}$ (nonparametric here). When $S_{\bar{m}}$ is a small nonparametric model ($D_{\bar{m}} \leq (\log n)^4$), (25) is very similar to (24) in the parametric setting. In particular $n\mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right] \to +\infty$ as n tends to $+\infty$ implies $n(1-p/n) \to +\infty$ as well. For large nonparametric models ($D_{\bar{m}} > (\log n)^4$), the constraints on p are related to the ratio $\| s - s_{\bar{m}} \|^2 / \mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right]$. For instance when estimating $s \in \mathcal{H}(L, \alpha)$ by regular histograms, this ratio remains bounded while $n\mathbb{E}\left[\, \| s_{\bar{m}} - \widehat{s}_{\bar{m}} \|^2 \,\right]$ grows polynomially in n. Then p/n has to converge to 0 as n increases, but not too fast. In particular Loo ($p = 1$) is suboptimal in that setting (see Figure 4 panel (b)). Note that Theorem 3.4 has the same flavor as Corollary 1 in (Yang, 2007), except the density estimation setting allows to relate p to the features of the best estimator more closely.

3.2.2. *Simulation experiments.* Simulation experiments have been performed in the settings of Theorems 3.3 and 3.4, respectively when s belongs to (resp. does not belong to) the model collection. We used a polynomial model collection made of regular piecewise constant functions described in Section 2.1.2 for which Assumptions (**P-NP**) and (**BeMo**) are fulfilled with $\mu = 5.10^{-1}$. In each setting $N = 1\,000$ samples have been drawn. Results are given in Figures 3 and 4 where $\mathbb{P}\left[\, \widehat{m} = \bar{m} \,\right]$ is displayed with respect to the ratio p/n. Let us also mention that Lemma S.6 clearly shows (**LoEx**) holds true for all densities defined in the following as long as $D_m \geq 2$ for every model in the collection.

When s *belongs to the model collection* (Figure 3), the following densities have been used.

1. $s_1(t) = \frac{6}{8}\mathbb{1}_{[0,1/2]}(t) + \frac{10}{8}\mathbb{1}_{[1/2,1]}(t), \quad t \in [0,1]$ (panel (a)),
2. $s_2(t) = \frac{135}{112}\mathbb{1}_{[0,1/3]}(t) + \frac{135}{56}\mathbb{1}_{[1/3,1/2]}(t) + \frac{1}{4}\mathbb{1}_{[1/2,5/7]}(t) + \frac{1}{2}\mathbb{1}_{[5/7,1]}(t), \quad t \in [0,1]$ (panel (b)).

As predicted by Theorem 3.3, CV reaches model selection consistency for recovering the best parametric estimator $\widehat{s}_{\bar{m}}$ on condition p/n increases to 1 as n grows to $+\infty$. Comparing (a) and (b), the convergence rate is slower in (b). Unlike (a) where $\bar{m}$ remains almost unchanged as n increases, the best parametric estimator in (b) changes with n as allowed by (21). Therefore the slower convergence rate in (b) results from the higher dimension of the space of piecewise constant functions s_2 belongs to.



(a) (b)

FIG 3. $p/n \mapsto \mathbb{P}[\widehat{m} = \bar{m}]$ *for density* s_1 *(panel* (a)*) and* s_2 *(panel* (b)*).* $N = 1\,000$ *samples have been drawn.*

When s *does not belong to the model collection* (Figure 4), densities with different smoothness assumptions have been considered.

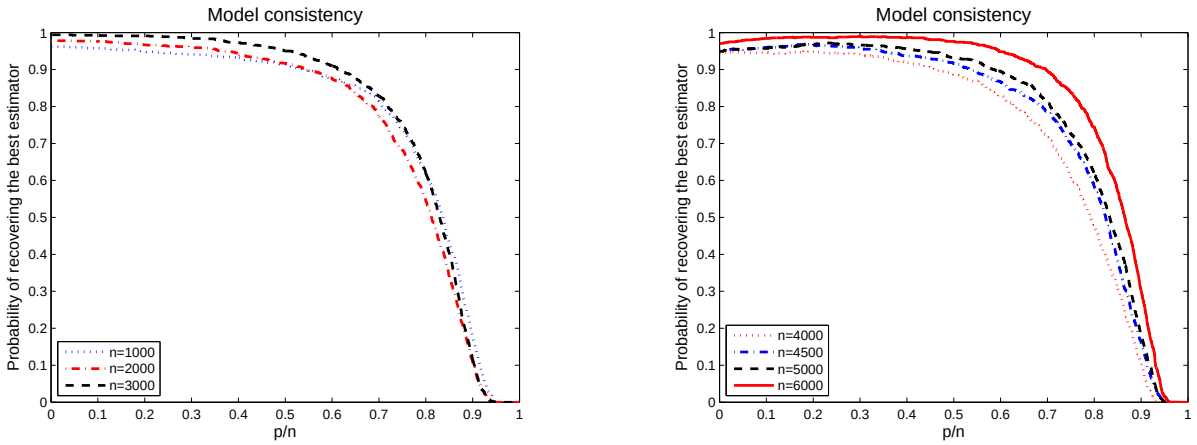
1. $s_3(x) = \beta(10, 7; x)$, for every $t \in [0,1]$ (panel (a)),
2. $s_4(x) = \frac{6}{5}x^{1/5}$, for every $t \in [0,1]$ (panel (b)).

The converse situation arises since CV reaches model selection consistency as long as p/n decreases to 0 as n tends to $+\infty$. Consistently with Theorem 3.4 this rate strongly depends on the risk of the best estimator, that is of the smoothness of the target. While model selection consistency is illustrated by both panels (a) and (b), it is faster for the smoothest density s_3 than for s_4. In Figure 4 panel (b), the highest probability is achieved for $p/n \approx 0.18$ with $n = 6\,000$.

4. Discussion. From the present analysis of CV procedures in the density estimation framework, we were able to prove the optimality of leave-one-out cross-validation for risk estimation, which is consistent with earlier results in the regression setting Burman (1989).

However 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).

Estimation. When the best model has dimension growing with n (faster than $(\log n)^a$ for some $a > 0$) and the model collection has a polynomial complexity (**Pol**), Theorem 3.1 proves any



(a) (b)

FIG 4. $p/n \mapsto \mathbb{P}[\widehat{m} = \bar{m}]$ *for density* s_3 *(panel* (a)*) and* s_4 *(panel* (b)*).* $N = 1\,000$ *samples have been drawn.*

p such that $p/n \to 0$ leads to an asymptotically optimal model selection procedure. This is consistent with the asymptotic equivalence between Lpo (as long as $p/n \to 0$) and Mallows'C_p previously settled in the regression setting (Shao, 1997).

From a non asymptotic point of view, Corollary 3.1 suggests choosing $p > 1$ (for finite sample) could balance the overfitting phenomenon arising from selecting a model from a large collection. This overfitting phenomenon is already well known with penalized criteria such as Mallow's ones, inducing the need for heavier constants in front of the penalty Arlot and Massart (2009). Therefore increasing p amounts to penalize more strongly complex models (with large dimension).

Identification. As settled by Yang (2007) for regression, Section 3.2 highlights the optimal p depends on the rate of convergence of the estimator one tries to recover (and on the structure of the model collection).

When the target estimator has a parametric rate with a polynomial collection, Theorem 3.3 proves $p/n \to 1$ leads to model selection consistency. This fact has been already noticed by Shao (1993) in the regression setting who proved leave-one-out is not model selection consistent. Remembering the asymptotic equivalence between Lpo and BIC-like criteria (Shao, 1997) established with the linear regression model, this confirms the somewhat paradoxical requirement (Yang, 2006) to devote most of available data ($p > n/2$) to the test set when trying to recover a parametric estimator.

Drawing such a simple conclusion is harder when the best estimator has a nonparametric rate as detailed by Theorem 3.4. If the best estimator has a rate close to parametric, then $p/n \to 1$ provides model selection consistency. Conversely if the rate is slower (for instance polynomial of order n^{-a}, for some $a > 0$), then requiring $p/n \to 0$ enables to recover the target estimator. Relating that way the optimal p to the rate of convergence of the best estimator has been already done in the regression context by (Yang, 2007, see Corollary 1).

Note that when the best estimator is nonparametric (for instance with polynomial rate),

Theorem 3.1 and Theorem 3.4 imply $p/n \to 0$ leads to both efficiency and respectively model selection consistency. However there is no contradiction with the earlier paper by Yang (2005) where it was proved no model selection criterion can share both efficiency and model selection consistency in a parametric setting. For instance Li (1987) has established model selection consistency for leave-one-out with nonparametric estimators in regression.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal cross-validation in density estimation with the L^2-loss": Technical proofs and details
(doi: COMPLETED BY THE TYPESETTER; .pdf). Owing to space constraints, we have moved technical proofs to a supplementary document corresponding to Section S.1 and following ones.

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. 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.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):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. and Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368.

de la Peña, V. H. and Giné, E. (1999). *Decoupling: From Dependence to independence*. Springer-Verlag, New York.

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. Educ. 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.

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

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

Yang, Y. (2006). Comparing Learning Methods for Classification. *Statistica Sinica*, 16:635–657.

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

Yang, Y. and Barron, A. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44:117–133.

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

APPENDIX A: ESTIMATION POINT OF VIEW

A.1. Proofs of main results.

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

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

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

Then, combining Propositions A.1 and A.2 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

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

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 S.1 and S.2 on the events Ω_{left} and Ω_{right}. The important point in this proof is given by Lemmas S.4 and S.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 S.9 and Corollaries S.1 and S.2 with $L_m = 0 = r_n(m)$ to get

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

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

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

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 S.4 and S.5 show $\min\left\{D_{m^*}, D_{\widehat{m}(p)}\right\} \geq (\log n)^4$. Then, it is enough to apply Propositions S.4 and S.3 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 $1 - 6/n^2$.

□

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 decreasing 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.

□

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

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2\right] = \mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}}\right] + \mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}(p)}\right\|^2 \mathbb{1}_{\widetilde{\Omega}^c}\right].$$

Then Theorem 3.1 applied to the first expectation in the right-hand side leads to

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

Applying (**Bias**), one gets

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

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

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

□

A.2. Bounding remainder terms.

PROPOSITION A.1 (Bound on $\nu_n(\phi_m - \phi_{m'})$).
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* $\mathbb{P}[\Omega_{\mathrm{rem},1}] \geq 1 - 2\sum_{m\in\mathcal{M}} e^{-x_m}$, *on which for every* $m, m' \in \mathcal{M}_n$

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

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

PROOF OF PROPOSITION A.1. A straightforward use of (42) leads to the desired conclusion. □

PROPOSITION A.2 (Bound on $\nu_n(s_m - s_{m'})$). *Let us assume* (**Pol**),(**SqI**), (**RegD**), (**LoEx**), *and* (**OrSp**) *hold true. Then, there exists a sequence* $(\delta_n)_{\mathbb{N}}$ *such that for every* $m, m' \in \mathcal{M}$,

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

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* $m, m' \in \mathcal{M}$

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

PROOF OF PROPOSITION A.2. For every $\eta > 0$,

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

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

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

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

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

(27)

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

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

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

Then,

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

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

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

Then,

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

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

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

it comes

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

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

□

APPENDIX B: IDENTIFICATION POINT OF VIEW

B.1. Main proof.

PROOF OF THEOREM 3.3. The general purpose is to prove there exist an event Ω_n with $\mathbb{P}(\Omega_n) \to 1$ as n tends to $+\infty$ and a positive integer N such that on Ω_n, for every $n \geq N$, every $m \neq \bar{m}$ satisfies

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq -u_n(m)(1 + o(1)), \tag{28}$$

where $u_n(m) > 0$ denotes a real number for every n and m, and $o(1)$ does not depend on m. In particular this implies

$$\mathbb{P}(\widehat{m} = \bar{m}) = \mathbb{P}\left(\forall m \neq \bar{m}, \quad \widehat{R}_p(\bar{m}) - \widehat{R}_p(m) < 0\right) \xrightarrow[n\to+\infty]{} 1, \tag{29}$$

which would conclude the proof.

Let us consider the event $\Omega_{\text{left}} \cap \Omega_{\text{right}}$ in Proposition S.7 with $\beta_1 = \beta_2 = 1/n^2$, and the events $\Omega_{\text{rem},1}$ (Proposition A.1) and $\Omega_{\text{rem},3}$ (Proposition S.6). Then with $\Omega_n = \Omega_{\text{left}} \cap \Omega_{\text{right}} \cap \Omega_{\text{rem},1} \cap \Omega_{\text{rem},3}$ and $\mathbb{P}[\Omega_n^c] \leq 8/n^2$, showing (29) amounts to prove

$$\mathbb{P}\left(\Omega_n \cap \left\{\forall m \neq \bar{m}, \quad \widehat{R}_p(\bar{m}) - \widehat{R}_p(m) < 0\right\}\right) \xrightarrow[n\to+\infty]{} 1.$$

Let us now focus on the event Ω_n. The two main steps correspond to distinguishing between *parametric* and *nonparametric* models S_m (see (**P-NP**)). For every m, let us define $B(m) = \|s - s_m\|^2$ and $V(m) = \mathbb{E}\left[\|s_m - \widehat{s}_m\|^2\right]$, where $s_m = \text{Argmin}_{t\in S_m} \|s - t\|^2$. From line to line, the value of δ_n may change, but it always denotes a sequence decreasing to 0 and such that $n\delta_n \to +\infty$ as n grows.

If $\widehat{s}_m$ has a parametric rate.

- If $s \in S_m$:
  Let us first notice $s_m = s = s_{\bar{m}}$, which implies $R_n(m) = V(m)$ and $R_n(\bar{m}) = V(\bar{m})$. Then, Proposition S.2, and Propositions A.1 and S.7 lead to

$$\left|\left[\widehat{R}_p(\bar{m}) - \widehat{R}_p(m)\right] - \frac{n}{n-p}\left[R_n(\bar{m}) - R_n(m)\right]\right| \leq \left(36L_n^2 + 3\delta_n + \delta_n\frac{n}{n-p}\right)\left[R_n(m) + R_n(\bar{m})\right]$$
$$= o\left(\frac{n}{n-p}\right)\left[R_n(m) + R_n(\bar{m})\right],$$

  by requiring $L_n^2 = o\left((1 - p/n)^{-1}\right)$, which provides (28) by use of (**BeMo**). Note that in the previous inequality, $r_n(\bar{m})$ and $r_n(m)$ (from Proposition S.7) have been omitted since they are negligible with respect to the other terms.
- If $s \notin S_m$:

Similarly, Proposition S.2, Proposition A.1, and Propositions S.6 and S.7 lead to

$$\begin{aligned}&\left|\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]-\left[B(\bar{m})-B(m)\right]-\frac{n}{n-p}\left[V(\bar{m})-V(m)\right]\right|\\&\leq 6\delta_n\left[B(\bar{m})+B(m)\right]+\left(36L_n^2+3\delta_n\frac{\Phi}{\xi}+\delta_n\left(3+\frac{n}{n-p}\right)\right)\left[V(m)+V(\bar{m})\right]\\&\quad+3\delta_n\|s\|\sqrt{\frac{\Phi}{\xi}}\sqrt{V(\bar{m})+V(m)}+r_n(\bar{m}).\end{aligned}$$

With $s\in\cup_{m'}S_{m'}$ and $s\notin S_m$, it comes $B(\bar{m})=0$ and $B(m)\geq\tau>0$ by (**P-NP**). Since both $\widehat{s}_m$ and $\widehat{s}_{\bar{m}}$ have parametric rates, requiring $n(1-p/n)\to+\infty$ as n grows implies (28), that is

$$\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]\leq -B(m)\left(1+o(1)\right).$$

If $\widehat{s}_m$ has a nonparametric rate.

- If $s\in S_m$
  Proposition S.2, $s_m=s=s_{\bar{m}}$, and Propositions S.6 and S.7 combined with $L_n^2=o\left(n/(n-p)\right)$ provide

  $$\left|\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]-\frac{n}{n-p}\left[V(\bar{m})-V(m)\right]\right|\leq o\left(\frac{n}{n-p}\right)V(m)+o\left(\frac{n}{n-p}\right)V(\bar{m}),$$

  where $o\left(n/(n-p)\right)$ does not depend on m. Since S_m is nonparametric, (**P-NP**) gives

  $$\frac{V(\bar{m})}{V(m)}\leq\frac{\pi}{(\log n)^2}\left(n/(\log n)^2\inf_m V(m)\right)^{-1}\xrightarrow[n\to+\infty]{}0,$$

  which implies (28) with

  $$(n-p)\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]\leq -nV(m)\left(1+o(1)\right).$$

- If $s\notin S_m$
  Both $L_n^2=o\left(n/(n-p)\right)$ and the same argument as above lead to

  $$\left|\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]+B(m)+\frac{n}{n-p}V(m)\left[1+o(1)\right]\right|\leq V(m)\;o\left(\frac{n}{n-p}\right)+B(m)\;o\left(1\right).$$

  Then, (28) holds true with

  $$\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]\leq -B(m)(1+o(1))-\frac{n}{n-p}V(m)(1+o(1)).$$

Then there exists an integer N such that for $n\geq N$, on the event Ω_n, (28) holds true for every $m\in\mathcal{M}_n$, which concludes the proof.

□

SUPPLEMENTARY MATERIAL TO "OPTIMAL CROSS-VALIDATION IN DENSITY ESTIMATION WITH THE L^2-LOSS"

S.1. Closed-form expressions and proofs.

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 S.1.

□

PROOF OF PROPOSITION 2.1. From definitions of the contrast (Eq. (2)) and the Lpo estimator Eq. (10), it comes

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

Then, Lemma 2.1 provides the expected conclusion. □

PROOF OF COROLLARY 2.5. Combining Proposition 2.2, Lemmas S.2 and S.3, and Proposition S.1, it comes

$$\begin{aligned} a_n &= 4n\beta + O(1),\\ b_n &= 8n\beta - 4n\gamma + O(1),\\ c_n &= 4n\beta - 4n\gamma + n\delta + O(1), \end{aligned}$$

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 conclusion with $A = 4\beta$, $B = 8\beta - 4\gamma$, and $C = 4\beta - 4\gamma + \delta$. □

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

$$\sum_{e \in \mathcal{E}_p} \mathbb{1}_{(j \in (e))} = \binom{n-1}{p} \quad \text{and} \quad \sum_{e \in \mathcal{E}_p} \mathbb{1}_{(j \in (e))} \mathbb{1}_{(k \in (e))} = \binom{n-2}{p-1},$$

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

LEMMA S.2. *With the same notation as Proposition 2.2, it comes*

$$a_n = n^2 \mathrm{Var}\left[\sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2\right] + \mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2\right] - 2n \mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_{\lambda \in \Lambda(m)} (P_n \varphi_\lambda)^2\right],$$

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

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

LEMMA S.3. *With the notation of Section 2.3, simple algebra leads to*

$$\mathrm{Var}\left[\sum_\lambda P_n \varphi_\lambda^2\right] = \frac{1}{n}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right],$$

$$\mathrm{Cov}\left[\sum_\lambda P_n \varphi_\lambda^2, \sum_\lambda (P_n \varphi_\lambda)^2\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_\lambda \varphi_\lambda^2(X_1)\right] + 2\frac{n-1}{n^2}\mathrm{Cov}\left[\sum_\lambda \varphi_\lambda^2(X_1), \sum_\lambda \varphi_\lambda(X_1)\varphi_\lambda(X_2)\right]$$

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

PROPOSITION S.1. *With the notation of Lemma S.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 S.1. Using Lemmas S.2 and S.3, it comes

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

In the same way,

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

Finally,

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

□

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

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

where

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

S.2. Deviations of $\sqrt{n}Z_m$.

S.2.1. *Right deviation.*

PROPOSITION S.3 (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* $b_m = \sup_{t\in S_m}\|t\|_\infty$. *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* $m \in \mathcal{M}$,

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

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* $C \geq \left(\sqrt{\xi}\right)^{-1}\log\left(5n^{a_{\mathcal{M}}}\beta_1^{-1}\right)$.

PROOF OF PROPOSITION S.3.
Let us use Eq. (43) and upper bound the deviation terms. Assuming (**SqI**) and (**RegD**), Lemma S.7 leads to

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

Furthermore, (**LoEx**) entails

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

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, let us remark that

$$\sum_{D_m \geq (\log n)^4} n^{-(a_{\mathcal{M}}+2)} \leq \frac{1}{n^2},$$

and that (**Pol**) and (**LoEx**) provide

$$\sum_{D_m < (\log n)^4} e^{-C\sqrt{n}E(Z_m)} \leq \sum_{d=1}^{(\log n)^4} \left|\{m \in \mathcal{M}_n \mid D_m = d\}\right| e^{-C\sqrt{\xi}\sqrt{d}} \leq n^{a_{\mathcal{M}}} \sum_{d=1}^{(\log n)^4} e^{-C\sqrt{\xi}\sqrt{d}},$$

which, by Lemma S.12, leads to

$$\sum_{D_m < (\log n)^4} e^{-C\sqrt{n}E(Z_m)} \leq n^{a_{\mathcal{M}}} \frac{5}{1 \wedge C^2\xi} e^{-C\sqrt{\xi}}.$$

Then,

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \frac{1}{n^2} + n^{a_{\mathcal{M}}} \frac{5}{1\wedge C^2\xi} e^{-C\sqrt{\xi}}.$$

□

COROLLARY S.1. *For $Z_m = \sup_{t\in S_m} \nu_n(t)$, set $L_m = \sqrt{4\sqrt{\frac{\Phi}{\xi}} C \|s\| \mathbb{1}_{(\sqrt{D_m}<(\log n)^2)}}$. Then on the event Ω_{right} defined in Proposition S.3,*

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

S.2.2. *Left deviation.*

PROPOSITION S.4 (Left deviation of $\sqrt{n}Z_m$). *Let us assume* (**Pol**), (**SqI**), (**RegD**), *and* (**LoEx**) *hold true, and set $Z_m = \sup_{t\in S_m} \nu_n(t)$, $\sigma_m^2 = \sup_{t\in S_m} \operatorname{Var}[t(X_1)]$ 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 $m \in \mathcal{M}$,*

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

on an event Ω_{left} with $\mathbb{P}[\Omega_{\text{left}}] \geq 1-1/n^2-\beta_2$ for any $\beta_2 \in (0,1)$ and $C \geq \left(\sqrt{\xi}\right)^{-1} \log\left(5n^{a_{\mathcal{M}}}\beta_2^{-1}\right)$

PROOF OF PROPOSITION S.4. Similar to that of Proposition S.3 with the use of Eq. (44) and the additional Proposition S.5 which provides an upper bound of $\mathbb{E}(Z_m)^2$ depending on $\mathbb{E}(Z_m^2)$. □

PROPOSITION S.5 (Upper bound on $\operatorname{Var}(Z)$). *Let $X_1,\dots,X_n$ 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 $\sup_{t\in S}\|t\|_\infty \leq b$, $\sup_{t\in S}\operatorname{Var}(t(X_1)) = \sigma^2$, and set $Z = \sup_{t\in S}\nu_n(t)$. Then,*

$$\operatorname{Var}(Z) \leq \frac{2\sigma^2 + 32b\mathbb{E}(Z)}{n}. \tag{30}$$

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*

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

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

PROOF OF PROPOSITION S.5. Assumptions (**SqI**), (**RegD**), and (**LoEx**) provide

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

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

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

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

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

   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 $\delta_n = \max\{\delta_{1,n}, \delta_{2,n}\}$ decreasing to 0 with $n\delta_n \to +\infty$ as $n \to +\infty$, such that

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

□

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

- *if $\delta_n + L_m \leq 1$,*

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

- *otherwise,*

$$Z_m^2 - \mathbb{E}\left(Z_m^2\right) \geq -\mathbb{E}\left(Z_m^2\right).$$

PROOF OF COROLLARY S.2.

- If $\delta_n + L_m \leq 1$, Propositions S.4 and S.5 provide

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

  and Lemma S.10 enables to conclude.
- Otherwise from Proposition S.4, one easily derives that $Z_m \geq 0$ on the event Ω_{left}.

□

S.3. Dimension behavior with respect to n.

LEMMA S.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 S.1 and Corollary S.2, it comes*

$$D_{m^*} \geq (\log n)^4, \tag{31}$$

for large enough values of n.

PROOF OF LEMMA S.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 S.1 and S.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 S.5 (Chosen model dimension). *Let us assume* (**Bias**), (**Rich**), (**LoEx**), *and* (**RegD**) *hold true. Then with the notation of Lemma S.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.1 and Proposition A.2, it comes*

$$D_{\widehat{m}} \geq (\log n)^4, \tag{32}$$

for large enough values of n.

PROOF OF LEMMA S.4. For any model m such that $\widehat{R}_p(m) \leq \widehat{R}_p(m_0)$, Proposition A.1, Proposition A.2, and Proposition S.2 lead to

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

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

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

Second, using (**Bias**) provides

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

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

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

□

S.4. Estimation: Technical results.

LEMMA S.6. *Let s denote a density in $L^2([0,1])$ and $0 < \eta, \ell < 1$ be constants such that*

$$|\{x \in [0,1] \mid s(x) \geq \eta\}| \geq \ell.$$

*Then using regular histograms such that $\ell > 1/D_m$, (**LoEx**) is fulfilled with $\xi = \frac{1}{32}\eta(1-\eta)\ell$.*

PROOF OF LEMMA S.6. First for $0 < \eta < 1$, let us define

$$L_\eta = \{x \in [0,1] \mid s(x) \geq \eta\}$$

such that $|L_\eta| \geq \ell$. Then $\ell > 1/D_m$ implies $P\varphi_\lambda = p_\lambda/\sqrt{D_m} \neq 1/\sqrt{D_m}$ for every λ and then

$$\sum_\lambda \mathrm{Var}\,[\,\varphi_\lambda(X)\,] = D_m \sum_\lambda p_\lambda(1-p_\lambda) \geq D_m \sum_{I_\lambda \cap L_\eta \neq \emptyset} \eta\,|I_\lambda \cap L_\eta|\,(1 - \eta\,|I_\lambda \cap L_\eta|).$$

Therefore, it results

$$\begin{aligned}
\sum_\lambda \mathrm{Var}\,[\,\varphi_\lambda(X)\,] &\geq D_m\left(\sum_{I_\lambda \cap L_\eta \neq \emptyset} \eta\,|I_\lambda \cap L_\eta| - \sum_{I_\lambda \cap L_\eta \neq \emptyset} \eta^2\,|I_\lambda \cap L_\eta|^2\right) \\
&\geq D_m\eta(1-\eta) \sum_{I_\lambda \cap L_\eta \neq \emptyset} |I_\lambda \cap L_\eta| = D_m\eta(1-\eta)\,|L_\eta| \\
&\geq D_m\eta(1-\eta)\ell.
\end{aligned}$$

Second, the Bonami-Becker inequality applied to homogeneous Rademacher chaos of order one (see Boucheron et al., 2013, p.140) combined with Lemma 1.2.6 in de la Peña and Giné (1999) applied to symmetrized statistics lead to

$$\mathbb{E}\left[\,\|s_m - \widehat{s}_m\|\,\right] \geq \frac{1}{4\sqrt{2}}\sqrt{\mathbb{E}\left[\,\|s_m - \widehat{s}_m\|^2\,\right]} = \frac{1}{4\sqrt{2}}\sqrt{\frac{1}{n}\sum_{\lambda}\mathrm{Var}\left[\,\varphi_\lambda(X)\,\right]}.$$

The final result comes from gathering the two lower bounds. □

LEMMA S.7. *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 $\sup_{t\in S}\mathrm{Var}\left(t(X_1)\right) = \sigma^2$. Let us assume* (**SqI**). *Then*

$$\sigma^2 \leq \|s\|\, b. \tag{33}$$

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

$$\sigma^2 \leq \|s\|\sqrt{\Phi D}, \tag{34}$$

where D denotes the dimension of the vector space S.

LEMMA S.8. (**Dmax**) *implies*

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

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 S.9. *Let us assume that* $0 \leq \delta_n + L_m < 1/4$ *for every* $m \in \mathcal{M}$. *Then on the event* $\Omega_{\text{left}} \cap \Omega_{\text{right}}$ *(with* Ω_{left} *and* Ω_{right} *defined in Proposition S.4 and Proposition S.3 respectively), for every* $m, m' \in \mathcal{M}$,

$$Z_{m'}^2 \left(1 - 4(\delta_n + L_{m'})\right) \leq \mathbb{E}(Z_{m'}^2), \qquad Z_{m'}^2 - \mathbb{E}(Z_{m'}^2) \leq 4Z_{m'}^2(\delta_n + L_{m'}). \tag{35}$$

and

$$\mathbb{E}(Z_m^2) \leq Z_m^2 \left(1 + 4(\delta_n + L_m)\right) + r_n, \qquad \mathbb{E}(Z_m^2) - Z_m^2 \leq 4Z_m^2(\delta_n + L_m) + r_n, \tag{36}$$

where $r_n(m)$ *is defined in Proposition S.5*

PROOF OF LEMMA S.9.

Proof of (35) From Corollary S.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 (36) Assuming $\delta_n + L_m < 1/4$, Corollary S.2 and Lemma S.11 provide

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

□

LEMMA S.10. *For every* $a, b \in (0, 1)$ *such that* $a < b(1-b)^{-1}$,

$$\frac{1+a}{(1-b)^2} \leq \frac{1}{(1-b)^3}. \tag{37}$$

Moreover if $0 < a = b < 1$, *then* $a < a(1-a)^{-1}$ *and Eq.* (37) *holds true.*

LEMMA S.11. *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 $x \in]1, +\infty[$,

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

LEMMA S.12. *For every $A > 0$ and $K \in \mathbb{N}^*$,*

$$\sum_{d=1}^{K} e^{-A\sqrt{d}} \leq \frac{5}{1 \wedge A^2} e^{-A},$$

where $a \wedge b = \min(a, b)$.

PROOF OF LEMMA S.12.

$$\begin{aligned}
\sum_{d=1}^{K} e^{-A\sqrt{d}} &\leq e^{-A} + 2 \sum_{d=2}^{K} \int_{\sqrt{d-1}}^{\sqrt{d}} t e^{-At} dt \\
&= e^{-A} + 2 \left[\frac{e^{-A} - \sqrt{K} e^{-A\sqrt{K}}}{A} + \frac{e^{-A} - e^{-A\sqrt{K}}}{A^2} \right] \\
&\leq e^{-A} + 2 \left[\frac{e^{-A}}{A} + \frac{e^{-A}}{A^2} \right] \leq \frac{5}{1 \wedge A^2} e^{-A}.
\end{aligned}$$

□

S.5. Adaptivity in the minimax sense.

S.5.1. *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 S.13, 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.

S.5.2. *Technical Lemmas.*

LEMMA S.13. *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,*

$$\frac{c_\ell}{D^\ell} \leq \|s - s_D\|^2 \leq \frac{c_u}{D^u}, \tag{38}$$

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 S.13. First, let us notice (38) 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

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

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 S.14, it leads to

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

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,

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

□

LEMMA S.14. *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,*

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

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

PROOF OF LEMMA S.14. 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 since $f \in \mathcal{H}(L,\alpha)$,

$$\frac{\Delta}{4} = \sup_{x\in J} \left|s(x^+) - s(x)\right| \leq L \sup_{x\in J} \left|x^+ - x\right|^\alpha \leq L\,|J|^\alpha,$$

which implies

$$|J| \geq \left(\frac{\Delta}{4L}\right)^{1/\alpha}.$$

Finally,

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

□

S.6. Identification: main result.

PROOF OF THEOREM 3.4. The strategy is the same as in the proof of Theorem 3.3. With the same notation, let us consider the event $\Omega_n = \Omega_{\text{left}} \cap \Omega_{\text{right}} \cap \Omega_{\text{rem},1} \cap \Omega_{\text{rem},2}$ where $\Omega_{\text{rem},2}$ is given in Proposition A.2, and $\mathbb{P}\left[\,\Omega_n^c\,\right] \leq 8/n^2$.

If $R_n(\bar{m}) \geq \kappa > 0$. The successive use of Proposition S.2, Propositions A.1 and A.2 (where (**OrSp**) is replaced by (**P-NP**)), and Proposition S.7 lead to

$$\begin{aligned}
&\left| \left[\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \right] - \left[B(\bar{m}) - B(m) \right] - \frac{n}{n-p} \left[V(\bar{m}) - V(m) \right] \right| \\
\leq & 3\delta_n \left[R_n(\bar{m}) + R_n(m) \right] + \left(\delta_n \left(3 + \frac{n}{n-p} \right) \right) \left[V(m) + V(\bar{m}) \right] \\
& + \left(36L_n^2 + 9\delta_n \right) V(\bar{m}) + \left(36L_n^2 + 9\delta_n \right) V(m) + 3r_n(\bar{m}) + 3r_n(m).
\end{aligned}$$

Combining that $R_n(\bar{m}) \geq \kappa$, (**BeMo**), and that $p/(n-p)V(\bar{m}) = o(1)$, it results for every $m \neq \bar{m}$

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq -\kappa\mu + o(1) - \frac{p}{n-p} V(m)(1 + o(1)),$$

where $o(1)$ does not depend on m by use of (**Dmax**).

Let us now assume $R_n(\bar{m}) = o(1)$, which implies $B(\bar{m}) = o(1)$.

1. If $D(\bar{m}) \leq (\log n)^4$:
   If $\widehat{s}_m$ *has a parametric rate.*
   The successive use of Proposition S.2, Propositions A.1 and A.2 (where (**OrSp**) is replaced by (**P-NP**)), and Proposition S.7 lead to

$$\begin{aligned}
&\left| \left[\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \right] - \left[B(\bar{m}) - B(m) \right] - \frac{n}{n-p} \left[V(\bar{m}) - V(m) \right] \right| \\
\leq & 3\delta_n \left[R_n(\bar{m}) + R_n(m) \right] + \left(\delta_n \left(3 + \frac{n}{n-p} \right) \right) \left[V(m) + V(\bar{m}) \right] \\
& + 36L_n^2 V(\bar{m}) + 36L_n^2 V(m) + 3r_n(\bar{m}) + 3r_n(m). \qquad (39)
\end{aligned}$$

   Since $s \notin S_m$ and $B(m) \geq \tau > 0$ by (**P-NP**), it comes

$$\begin{aligned}
\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq & \left[B(\bar{m}) - B(m) \right] + \frac{n}{n-p} \left[V(\bar{m}) - V(m) \right] + 3\delta_n \left[R_n(\bar{m}) + R_n(m) \right] \\
& + \left(\delta_n \left(3 + \frac{n}{n-p} \right) \right) \left[V(m) + V(\bar{m}) \right] \\
& + 36L_n^2 V(\bar{m}) + 36L_n^2 V(m) + 3r_n(\bar{m}) + 3r_n(m) \\
& \leq -\tau \left[1 + o(1) - \frac{B(\bar{m})}{\tau}(1 + o(1)) - \frac{n}{n-p} \left[\frac{V(\bar{m})(1 + o(1)) - (1 - o(1))V(m)}{\tau} \right] \right] \\
& + 36L_n^2 V(\bar{m}) + 36L_n^2 V(m) + 3r_n(\bar{m}) + 3r_n(m) \\
& \leq -\tau \left[1 + o(1) - \frac{B(\bar{m})}{\tau}(1 + o(1)) - \frac{n}{n-p} \left[\frac{V(\bar{m})(1 + o(1))}{\tau} \right] \right] \\
& + 36L_n^2 V(\bar{m}) + 36L_n^2 V(m) + 3r_n(\bar{m}) + 3r_n(m).
\end{aligned}$$

   Finally since (**Dmax**) implies $L_n^2 V(m) \to 0$ as n tends to $+\infty$, it is enough to require $nV(\bar{m})/(n-p) \to 0$ as n tends to $+\infty$. Then

$$\widehat{R}_p(\bar{m}) - \widehat{R}_p(m) \leq -\tau \left[1 + o(1) \right],$$

where $o(1)$ is independent of m.
If $\widehat{s}_m$ has a nonparametric rate. From the following inequality

$$\begin{aligned}&\left|\left[\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\right]-\left[B(\bar{m})-B(m)\right]-\frac{n}{n-p}\left[V(\bar{m})-V(m)\right]\right|\\ \leq&3\delta_n\left[R_n(\bar{m})+R_n(m)\right]+\left(\delta_n\left(3+\frac{n}{n-p}\right)\right)\left[V(m)+V(\bar{m})\right]\\ &+36L_n^2V(\bar{m})+(36L_n^2+9\delta_n)V(m)+3r_n(\bar{m})+3r_n(m),\end{aligned}$$

let us require $L_n^2=o(n/(n-p))$ to control the $L_n^2V(\bar{m})$ term. Then, (**BeMo**) implies

$$\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\leq-\mu R_n(\bar{m})\left[1+o(1)\right],$$

where $o(1)$ is independent of m.

2. If $D(\bar{m})>(\log n)^4$:
If $\widehat{s}_m$ has a parametric rate. From $s\notin S_m$ combined with $B(m)\geq\tau>0$, it comes

$$\begin{aligned}\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\leq&-\tau\left[1+o(1)-\frac{B(\bar{m})}{\tau}(1+o(1))-\frac{n}{n-p}\left[\frac{V(\bar{m})(1+o(1))}{\tau}\right]\right]\\ &+9\delta_nV(\bar{m})+36L_n^2V(m)+3r_n(\bar{m})+3r_n(m).\end{aligned}$$

Then, $nV(\bar{m})/(n-p)\to0$ as n tends to $+\infty$ implies

$$\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\leq-\tau\left[1+o(1)\right],$$

where $o(1)$ is independent of m.
If $\widehat{s}_m$ has a nonparametric rate.

$$\begin{aligned}\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\leq&-\mu\left(1+o(1)\right)R_n(\bar{m})\\ &+\frac{p}{n-p}\left[\left(1+o(1)\right)V(\bar{m})-\left(1-o(1)\right)V(m)\right]\\ &+(36L_n^2\mathbb{1}_{(D_m\leq(\log n)^4)}+9\delta_n\mathbb{1}_{(D_m>(\log n)^4)}V(m)\\ \leq&-\mu\left(1+o(1)\right)R_n(\bar{m})\\ &+\frac{p}{n-p}\left[\left(1+o(1)\right)V(\bar{m})-\left(1-o(1)\right)V(m)\right]\\ &+36L_n^2\mathbb{1}_{(D_m\leq(\log n)^4)}V(m).\end{aligned}$$

Since requiring $(\log n)^5/\left[nV(\bar{m})\right]=o\left(p/(n-p)\right)$ implies

$$\widehat{R}_p(\bar{m})-\widehat{R}_p(m)\leq-\mu\left(1+o(1)\right)R_n(\bar{m})+\frac{p}{n-p}\left(1+o(1)\right)V(\bar{m}),$$

one concludes by choosing p such that

$$\frac{p}{n-p}=o\left(1\vee\frac{B(\bar{m})}{V(\bar{m})}\right).$$

Note the two constraints

$$(\log n)^5/\left[\, nV(\bar{m}) \,\right] = o\left(\frac{p/n}{1-p/n}\right), \quad \text{and} \quad \frac{p}{n-p} = o\left(1 \vee \frac{B(\bar{m})}{V(\bar{m})}\right)$$

imply $nV(\bar{m})/(n-p) \to 0$ as n tends to $+\infty$ is fulfilled.

Then there exists an integer N such that for $n \geq N$, on the event Ω_n, (28) holds true for every $m \in \mathcal{M}_n$, which concludes the proof.

□

S.7. Identification: Technical results.

PROPOSITION S.6 (Bound on $\nu_n(s_m - s_{\bar{m}})$). *Let us assume* (**Pol**),(**SqI**), (**RegD**), *and* (**LoEx**) *hold true. Then, there exists a sequence* $(\delta_n)_{\mathbb{N}}$ *and an event* $\Omega_{\mathrm{rem},3}$ *with* $\mathbb{P}\left[\, \Omega_{\mathrm{rem},3} \,\right] \geq 1 - 2/n^2$, *on which for every* $m \in \mathcal{M}$,

$$2\left|\nu_n(s_m - s_{\bar{m}})\right| \leq \delta_n \left\|s_m - s_{\bar{m}}\right\|^2 + \delta_n \left\|s\right\| \sqrt{\frac{\Phi}{\xi}} \sqrt{\frac{n\mathbb{E}\left(Z_m^2\right) + n\mathbb{E}\left(Z_{\bar{m}}^2\right)}{n}} + \delta_n \frac{\Phi}{\xi} \frac{n\mathbb{E}\left(Z_m^2\right) + n\mathbb{E}\left(Z_{\bar{m}}^2\right)}{n},$$

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 S.6. Combining Eq. (27), (**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 $\sum_{m \in \mathcal{M}} 2e^{-x_m} = \sum_{m \in \mathcal{M}} 2n^{-(a_{\mathcal{M}}+2)} \leq 2n^{a_{\mathcal{M}}} n^{-(a_{\mathcal{M}}+2)} = 2/n^2$.

□

PROPOSITION S.7. *On the event* $\Omega_{\text{left}} \cap \Omega_{\text{right}}$ *(with* Ω_{left} *and* Ω_{right} *defined in Proposition S.4 and Proposition S.3 respectively), for large enough values of* n,

- *if* S_m *is either parametric or nonparametric with* $D_m \leq (\log n)^4$ *(see* (**P-NP**)*),*

$$\left|Z_m^2 - \mathbb{E}\left(Z_m^2\right)\right| \leq 12 L_n^2 \mathbb{E}\left(Z_m^2\right) + r_n(m), \tag{40}$$

- *otherwise,*

$$\left|Z_m^2 - \mathbb{E}\left(Z_m^2\right)\right| \leq 3\delta_n \mathbb{E}\left(Z_m^2\right), \tag{41}$$

where $a \vee b = \max(a,b)$, $r_n(m)$ *is defined in Proposition S.5 and* $L_n = \sqrt{4\sqrt{\Phi/\xi} C \|s\|} \geq \sqrt{4\sqrt{\Phi/\xi} C \|s\|}$ *with* $C \geq \left(\sqrt{\xi}\right)^{-1} \log\left(5 n^{a \curlywedge} \beta_1^{-1}\right)$ *(see Proposition S.3).*

LEMMA S.15. *On the event* $\Omega_{\text{left}} \cap \Omega_{\text{right}}$ *(with* Ω_{left} *and* Ω_{right} *defined in Proposition S.4 and Proposition S.3 respectively), for every* $m \in \mathcal{M}$,

$$\left|Z_m^2 - \mathbb{E}\left(Z_m^2\right)\right| \leq 3\left[\,(\delta_n + L_m)^2 \vee (\delta_n + L_m)\,\right] \mathbb{E}\left(Z_m^2\right) + r_n(m),$$

where $a \vee b = \max(a,b)$ *and* $r_n(m)$ *is defined in Proposition S.5.*

PROOF OF LEMMA S.15. From Corollary S.1, one derives

$$Z_m^2 - \mathbb{E}\left(Z_m^2\right) \leq \left[\,(1+\delta_n+L_m)^2 - 1\,\right] \mathbb{E}\left(Z_m\right)^2 \leq 3\left[\,(\delta_n+L_m)^2 \vee (\delta_n+L_m)\,\right] \mathbb{E}\left(Z_m^2\right).$$

1. If $\delta_n + L_m \geq 1$,

$$Z_m^2 - \mathbb{E}\left(Z_m^2\right) \leq 3\left(\delta_n + L_m\right)^2 \mathbb{E}\left(Z_m^2\right),$$

and Proposition S.4 leads to

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

2. If $\delta_n + L_m < 1$,

$$Z_m^2 - \mathbb{E}\left(Z_m^2\right) \leq 3\left(\delta_n + L_m\right) \mathbb{E}\left(Z_m^2\right),$$

and Propositions S.4 and S.5 provide

$$\begin{aligned} Z_m^2 - \mathbb{E}\left(Z_m^2\right) &\geq \left(\frac{(1-\delta_n-L_m)^2}{1+\delta_n+L_m} - 1\right) \mathbb{E}\left(Z_m^2\right) - r_n(m) \frac{(1-\delta_n-L_m)^2}{1+\delta_n+L_m} \\ &= (\delta_n + L_m) \frac{-3 + (\delta_n - L_m)}{1+\delta_n+L_m} \mathbb{E}\left(Z_m^2\right) - r_n(m) \frac{(1-\delta_n-L_m)^2}{1+\delta_n+L_m} \\ &\geq -3(\delta_n + L_m) \mathbb{E}\left(Z_m^2\right) - r_n(m). \end{aligned}$$

□

S.8. Key concentration inequalities.

THEOREM S.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 $x > 0$,*

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

THEOREM S.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 $\sup_{t\in S}\|t\|_\infty \leq b$ 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 S.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 $\sup_{t\in S}\|t\|_\infty \leq b$ 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 $x > 0$*

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

LABORATOIRE DE MATHÉMATIQUES PAINLEVÉ,
UMR 8524 CNRS-UNIVERSITÉ LILLE 1, MODAL PROJECT-TEAM INRIA
F-59 655, VILLENEUVE D'ASCQ CEDEX
E-MAIL: celisse@math.univ-lille1.fr