Density estimation via cross-validation: Model selection point of view

Alain Celisse

Density estimation *via* cross-validation: Model selection point of view

Alain Celisse

5th November 2008

Abstract

The problem of density estimation by cross-validation is addressed in the model selection framework. Extensively used in practice, cross-validation remains poorly understood, especially in the non-asymptotic setting. More precisely, our analysis mainly focuses on a cross-validation based algorithm named leave-p-out. Our better understanding of the leave-p-out with respect to the cardinality p of the test set yields more insight into other cross-validation based algorithms.
From a general point of view, cross-validation is devoted to estimate the risk of an estimator. Usually due to a prohibitive computational complexity, the leave-p-out is taken for intractable. However, we turned it into a fully effective procedure thanks to closed-form formulas for the risk estimator of a wide range of widespread estimators.
Embedding leave-p-out in the model selection setting enables a new interpretation of this algorithm in terms of a penalized criterion, with a random penalty. Furthermore, the amount of overpenalization it provides turns out to increase with p.
A theoretical assessment of the leave-p-out performance is provided thanks to two oracle inequalities applying respectively to either bounded densities or square integrable ones.
With different sieves such as piecewise constant functions or trigonometric and dyadic polynomials, the leave-p-out based strategy exhibits some adaptivity properties in the minimax sense with respect to Hölder as well as Besov spaces.

Keywords: Density estimation, cross-validation, model selection, leave-p-out, random penalty, oracle inequality, projection estimators, adaptivity in the minimax sense, Hölder, Besov.

1 Introduction

Model selection via penalization has been introduced by the seminal works of Mallows and Akaike with respectively C_p [27] and AIC [1], and also by Schwarz [33] who proposed the BIC criterion. AIC and BIC have an asymptotic flavour, which makes their performance depend on the model collection in hand as well as on the sample size [2].
More recently, Birgé and Massart [6, 7, 8] have developed a non-asymptotic approach, inspired from the pioneering work of Barron and Cover [4]. It aims at choosing a model among a countable family $\{S_m\}_{m\in\mathcal{M}_n}$ where $\mathcal{M}_n$ is allowed to depend on the sample size n. From this point of view, an estimator $\widehat{s}_m$ is associated with each model S_m, and a penalty-based procedure is designed and then minimized to provide a final estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$. The goal of this approach is *efficiency*, that is the risk of $\widetilde{s}$ is as small as that of the minimizer over all the estimators in the collection. Actually, this cannot be reached and the quality assessment of the procedure is made through an oracle inequality

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] \leq C \inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[\left\|s - \widehat{s}_m\right\|^2\right] + R(m,n)\right\},$$

where $C \geq 1$ is a constant independent from the density s, while $R(m,n)$ denotes a remainder term with respect to $\mathbb{E}\left[\left\|s - \widehat{s}_m\right\|^2\right]$. Such an inequality quantifies how close the risk of $\widetilde{s}$ is to the smallest one among those of the $\widehat{s}_m$s (up to a constant C and a remainder term). Thus the closer C to 1 and $R(m,n)$ to 0, the better the procedure.
In the density estimation framework, Barron *et al.* [3] developed a general approach based on deterministic

penalties, leading to an oracle inequality involving Kullback-Leibler divergence and Hellinger distance. This result has been adapted to the particular case of histograms by Castellan [15, 16] and further studied in Birgé and Rozenholc [9]. With the quadratic risk, the penalties proposed by Birgé and Massart [6, 3] also enjoy some optimality properties when applied to projection estimators. They establish that the resulting estimator exhibits some adaptivity in the minimax sense with respect to Besov spaces for several appropriate functional bases [6].

Unlike the aforementioned approaches relying on some deterministic penalties, we here address the problem of density estimation *via* cross-validation (CV). "Cross-validation" refers to a family of resampling-based algorithms, resulting from a heuristic argument. The first cross-validation algorithms have been respectively introduced in a regression context by Stone [37, 38] for the leave-one-out (Loo) and Geisser [21, 22] for the V-fold cross-validation (VFCV), and by Stone [36] in the density estimation framework.
Since these algorithms can be computationally demanding or even intractable, Rudemo [31] and Bowman [11] provided some closed-form expressions for the Loo estimator of the risk of histograms or kernel estimators. These results have been recently generalized by Celisse and Robin [17] to the leave-p-out cross-validation (Lpo).
Most of theoretical results about the effectiveness of CV algorithms are asymptotic and mainly concern the regression framework. For a fixed model, Burman [12, 13] expands several CV-based estimators of the risk of $\widehat{s}_m$ and concludes that Loo is the best one in terms of bias and variance. Besides several comparisons are pursued between Loo and various penalized criteria in Li [25] and Zhang [41] in view of *asymptotic efficiency*, while [34, 40] essentially focus on the *identification* or *consistency* point of view. We refer the interested reader to Shao [35] for an extensive review about asymptotic optimality properties in terms of efficiency and identification of some penalized criteria as well as some CV algorithms.
As for non-asymptotic results in the density setting, Birgé and Massart [6] obtained an oracle inequality that may be applied to the Loo procedure. However to the best of our knowledge, no result of this type has already been proved for the Lpo algorithm in the density estimation setup.

In the literature about model selection via penalization, an important notion is that of *complexity of the collection of models* $\{S_m\}_{m\in\mathcal{M}_n}$ [2], named in the sequel either *collection complexity* or *richness*. This notion already arises with discrete models in the Minimum Description Length of Rissanen [30] as well as in the work of Barron and Cover [4] about minimum complexity. It is further generalized by Barron *et al.* [3] as well as in Birgé and Massart [6, 7, 8] to the case of continuous models.
This notion of richness refers to the structure of the collection of models we consider. In the same way as the complexity of a model (*model complexity*) may be characterized by the dimension with finite dimensional vector spaces for instance, Barron *et al.* [3] quantify the *collection complexity* by the number of models with the same dimension. In this setting, two broad situations are usually distinguished: the polynomial and the exponential complexity frameworks [8]. Baraud *et al.* [2] and Sauvé [32] introduce a complexity index, which enables to distinguish different complexity levels in the exponential setup for instance.

The main concern of this paper is to provide a new understanding of cross-validation algorithms in terms of penalized criteria. We therefore perform the analysis of the Lpo algorithm in density estimation with the model selection framework. Our interest lies first in closed-form expressions derived for a wide range of widespread estimators, which turns the Lpo into a fully effective algorithm, and also in non-asymptotic results such as oracle inequalities. Adaptivity in the minimax sense with respect to some functional spaces is considered as well, for appropriate collection of models.

The paper is organized as follows. The following section is devoted to the description of the Lpo algorithm. Some closed-form expressions are derived for the resulting risk estimator with the broad class of projection estimators. Unlike the usual situation, the Lpo risk estimator is *no longer intractable*. Furthermore, closed-form expressions are also derived for both the bias and variance of this risk estimator. Besides, new highlight is given to cross-validation thanks to the relationship between the Lpo estimator and penalized criteria. The subsequent conclusion is that Lpo systematically leads to overpenalization, by an amount growing with p, the cardinality of the test set. The main concern in Section 3 is to derive

two oracle inequalities, which theoretically state the optimality of the Lpo-based procedure. The first inequality holds with bounded densities s, whereas the second one applies to densities in L^2, at the price of an additional assumption. Adaptivity in the minimax sense is at the core of Section 4, which aims at deriving such results with respect to Hölder and Besov spaces, for appropriate collections of models. While histograms are used with Hölder balls, dyadic as well as trigonometric polynomials enable to get such results with Besov balls. Finally, Section 5 collects most of the proofs of this paper.

2 Leave-p-out cross-validation

In this work, we address the problem of density estimation *via* cross-validation (CV) in the model selection framework. Resampling-based strategies such as CV are usually time-consuming and even sometimes computationally intractable. The interest of the forthcoming approach is to derive closed-form expressions for the CV-based estimator of the risk of projection estimators, which are widespread in the density estimation community ([31, 19, 6, 3]).

2.1 Statistical framework

Let us start describe the framework and introduce some notations which are repeatedly used throughout the paper.

2.1.1 Notations

In the sequel, we assume that $X_1, \dots, X_n \in [0,1]$ are independent and identically distributed random variables drawn from a probability distribution P of density $s \in L^2$ with respect to Lebesgue's measure. As a distance between s and any function u, we use the quadratic *loss* denoted by $\ell(\cdot,\cdot)$ such that

$$\ell(s,u) := \|s-u\|^2 .$$

Since this quantity depends on s , we introduce the associated *contrast* function

$$\gamma : \ (u,x) \mapsto \gamma(u,x) := \|u\|^2 - 2u(x).$$

This contrast is related to the loss function *via*

$$\ell(s,u) = P\gamma(u) - P\gamma(s), \qquad \text{where} \qquad P\gamma(u) = \mathbb{E}[\gamma(u,X)], \quad X \sim P,$$

for any function u.
The quality assessment of an estimator $\widehat{s} = \widehat{s}(X_1,\dots,X_n)$ of s is made through the corresponding *quadratic risk*

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

As an estimator of the above risk (up to a constant term), we use the *empirical risk*

$$\gamma_n(u) := P_n\gamma(u) = \frac{1}{n}\sum_{i=1}^{n}\gamma(u,X_i),$$

where $P_n = 1/n\sum_{i=1}^n \delta_{X_i}$ denotes the empirical measure.

Let $\{S_m\}_{m\in\mathcal{M}_n}$ denote a countable family of *models*, which are finite dimensional linear spaces. In each model S_m, we look for an estimator $\widehat{s}_m$ of s, defined as the empirical risk minimizer over S_m

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

2.1.2 Projection estimators

Let us now describe the range of estimators to which this work applies.
Set $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ a family of vectors in $L^2([0,1])$, where Λ_n denotes a countable set of indices. For each $m \in \mathcal{M}_n$, let $\Lambda(m)$ be a subset of Λ_n such that $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ is an orthonormal basis of S_m, of dimension D_m.

Thus, the orthogonal projection of s onto S_m is denoted by s_m and is equal to

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

In this setting, it turns out that the empirical risk minimizer corresponding to model S_m is a projection estimator since

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

Examples [18]

- Histograms:
  If we use $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ such that $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ where $\{I_\lambda\}_{\lambda\in\Lambda(m)}$ denotes a partition of $[0,1]$ in $\mathrm{Card}(\Lambda(m)) = D_m$ intervals and $|I_\lambda|$ is the Lebesgue measure of I_λ, the least-squares estimator coincides with the histogram

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

- Trigonometric polynomials:
  Let $\{\varphi_\lambda\}_{\lambda\in\mathbb{Z}}$ the orthonormal basis of $L^2([0,1])$ such that $t \mapsto \varphi_\lambda(t) = e^{2\pi i\lambda t}$. For any finite $\Lambda(m) \subset \mathbb{Z}$, the trigonometric polynomial defined by

  $$t \mapsto \widehat{s}_m(t) = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\ e^{2\pi i\lambda t}$$

  is a projection estimator.

- Wavelet basis:
  Let us consider an orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ of $L^2([0,1])$ made of compact supported wavelets, where $\Lambda_n = \left\{(j,k) \mid j \in \mathbb{N}^* \text{ and } 1 \le k \le 2^j\right\}$. The Haar system is a good example. For any subset $\Lambda(m)$ of Λ_n, the empirical risk minimizer associated with $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ is

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

  which is a projection estimator as well.

2.2 Cross-validation

In this work, we perform density estimation in the model selection framework *via* cross-validation instead of the usual deterministic penalized criteria. To this end, we first recall the rationale behind cross-validation and also detail the specific leave-p-out algorithm.

2.2.1 Cross-validation rationale

Cross-validation (CV) is a heuristics relying on some particular *subsamples* of the original observations. Various CV algorithms exist, which differ one another by the way subsamples are built, that is by their respective subsampling scheme. In order to further detail the rationale behind CV, we introduce some additional notations.
Let us denote by $X_{1,n} = X_1, \ldots, X_n$ the n sample drawn from P, and by P_n its empirical distribution. Each CV-based algorithm can be defined by a vector of weights $W_n = (W_{n,1}, \ldots, W_{n,n})$ made of random variables, *independent from* $X_{1,n}$. The subsamples are then defined from the original data through these weights and denoted by $X^W_{1,n} = X^W_1, \ldots, X^W_n$, with the empirical distribution $P^W_n = 1/n \sum_{i=1}^n W_{n,i} X_i$. Several examples of such subsampling schemes will be detailed in Section 2.2.2.
In the sequel, $\widehat{s} = \widehat{s}(X_{1,n})$ denotes an estimator of s computed from $X_{1,n}$.

The independence requirement
CV has essentially been designed to estimate the risk of any estimator $\widehat{s}$ of s. Actually, it intends to estimate $r = \mathbb{E}\left[P\gamma(\widehat{s}) \right]$, which can be expressed as

$$r = \mathbb{E}_{X_{1,n}} \left[\mathbb{E}_X \left[\gamma\left(\widehat{s}(X_{1,n}, X) \right) \right] \right],$$

where $\mathbb{E}_X$ and $\mathbb{E}_{X_{1,n}}$ represent expectations with respect to X and respectively $X_{1,n}$. We would like to highlight that X and $X_{1,n}$ are *independent*. Thus, there are two levels of randomness in the above expression. This is the cornerstone of the CV strategy, which heavily exploits these two randomness levels in splitting the data into a *training set* and a *test set*. Roughly speaking, the idea is simply to use the data in the training set to build the estimator, while the test set is devoted to the assessment of the estimator performance. Provided $X_{1,n}$ is made of independent data, training and test sets are independent by construction as well.
Remark: Independence is actually the main requirement of CV and even in the non *i.i.d.* case studied by Burman *et al.* [14], authors exploit the order of the dependence structure in removing some data from the sample so that they recover independence.

Strategy
Let us denote by W_n a binary vector corresponding to a particular CV scheme and associated with observations in the training set (examples will be provided in Section 2.2.2), while $\overline{W}_n$ denotes its natural counterpart representing the training set data.
The first step of the CV heuristics consists in "approximating" the random variable $\mathbb{E}_X \left[\gamma\left(\widehat{s}(X_{1,n}), X \right) \right]$, which depends on $X_{1,n}$, thanks to the subsamples. Let us assume that this dependence is made through the empirical distribution P_n. Then, we have

$$\widehat{r}(P_n, P) = \mathbb{E}_X \left[\gamma\left(\widehat{s}(X_{1,n}), X \right) \right].$$

The underlying idea in CV is to replace (P_n, P) by $\left(P^{\overline{W}}_n, P^W_n \right)$:

$$\widehat{r}(P_n, P) \approx \widehat{r}\left(P^{\overline{W}}_n, P^W_n \right),$$

where $P^{\overline{W}}_n$ denotes the empirical distribution of the data in the training set, while P^W_n states for that of the data in the test set.
Whereas the above step remains the same for all the CV-based strategies, the second one differs from a CV algorithm to another and depends on the way the subsamples have been generated. This subsampling scheme is therefore determined by the distribution of the weight vector W_n. Thus integrating over the weights W_n, it comes that

$$r = \mathbb{E}_{X_{1,n}} \left[\widehat{r}(P_n, P) \right] \approx \mathbb{E}_W \left[\widehat{r}\left(P^{\overline{W}}_n, P^W_n \right) \right],$$

where the expectation $\mathbb{E}_W$ is taken with respect to the vector W_n. The right-hand side is actually a random variable, which is the CV estimator of the risk of $\widehat{s}$.
Remark: At each step of the process, $X^W_{1,n}$ and $X^{\overline{W}}_{1,n}$ remain independent.

2.2.2 Leave-p-out and other cross-validation algorithms

The CV botany

In model selection, another interest of resampling methods, especially of CV, is their ability to work with any estimator [40] in a wide range of frameworks [10], unlike (deterministic) penalized criteria, which require a preliminary study of this estimator to design the appropriate penalty. Indeed if we think about AIC-like penalties [3, 16] for instance, there is no immediate warranty for them to be suited to any other estimator than the empirical contrast minimizer. However, the price CV has to pay for such a generality level is essentially the computation cost, which may be very high.

These two remarks as well as the high technicalities of the proofs all motivate the numerous variants of CV algorithms (see also [20] for an extensive review about CV):

- From a historical viewpoint, the *leave-one-out* (Loo) was the first CV scheme that appeared in a quite formalized version in Mosteller and Tukey [29], and then in [37]. It consists in successively removing each observation from the original data and computing the estimator from the $n-1$ remaining ones. The performance of the resulting estimator is then assessed thanks to the removed point. The final Loo risk estimator is defined as the average over the n possible test sets. In order to stick to the resampling formalism, Loo corresponds to the choice of a random vector W_n, such that $W_{n,j} \in \{0, n\}$, $\mathbb{P}(W_{n,j} > 0) = 1/n$ for any j, and $\sum_{j=1}^{n} W_{n,j}/n = 1$. The Loo risk estimator is expressed as

$$\widehat{R}_1(A) = \frac{1}{n} \sum_{i=1}^{n} \gamma(A(X_{1,n}^{(i)}), X_i),$$

  where $X_{1,n}^{(i)}$ represents $X_{1,n}$ from which X_i has been removed and A denotes an estimation algorithm, that is an application that takes as input some data and outputs an estimator. In a nutshell, $A(X_{1,n})$ is the estimator provided by algorithm A, computed from $X_{1,n}$. A typical example for A is the ERM algorithm, *e.g.* the empirical risk minimization.

- The *leave-p-out* (Lpo), with $p \in \{1, \dots, n-1\}$, may be seen as a generalization of the Loo to which it amounts when $p = 1$. It appears in a general setting in Burman [12], and in a linear regression setup in [34, 41]. In density estimation, Celisse and Robin [17] derive a closed-form expression for the Lpo estimator in the estimation of the risk of histograms. It consists in the same procedure as that of the Loo, except that at each of the $\binom{n}{p}$ rounds we remove p observations (instead of only one). The corresponding weights are defined by $W_{n,i} \in \{0, n/p\}$ for any i, $\sum_{i=1}^{n} p/n W_{n,i} = p$ and the probability of any such vector is $\binom{n}{p}^{-1}$. Thus with the same notations as before, the Lpo risk estimator (also named Lpo estimator or Lpo risk) is finally

$$\widehat{R}_p(A) = \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_p} \left[\frac{1}{p} \sum_{i \in e} \gamma\left(A(X_{1,n}^{\overline{e}}), X_i\right) \right],$$

  where $\mathcal{E}_p = \{(i_1, \dots, i_p) \,|\, i_1 < \dots < i_p,\ i_j \in \{1, \dots, n\}\}$, $e \in \mathcal{E}_p$ and $\overline{e} = \{1, \dots, n\} \setminus e$.

- Due to the high computational burden of the previous procedures, Geisser [21, 22] introduces an alternative algorithm named *V-fold cross-validation* (VFCV). The VFCV has been studied in Burman [12, 13] who suggests a correction in order to remove some bias. It relies on a preliminary (random or not) choice of a partition of the data in V subsets of approximately equal size n/V. Each subset is successively left out, and the $V-1$ remaining ones are used to compute the estimator, while the last one is dedicated to its performance assessment. The V-fold risk estimator is the average of the V resulting estimators. For a given random partition of the data, the above description results in V weight vectors W_n of respective probability $1/V$, satisfying $W_{n,i} \in \{0, V\}$ for any i, and $\sum_{i=1}^{n} W_{n,i} = n/V$. This leads us to the following VFCV estimator:

$$\widehat{R}_{\mathrm{VFCV},V}(A) = \frac{1}{V} \sum_{v=1}^{V} \left[\frac{V}{n} \sum_{i \in e_v} \gamma\left(A(X_{1,n}^{\overline{e_v}}), X_i\right) \right],$$

  where e_v denotes the n/V indices of the V-th subset.

- *The Hold(-p)-out* (Hpo), with $p \in \{1, \ldots, n-1\}$, is one of the simplest (to analyze) CV procedures, consisting in randomly partitioning the data in a training and a test sets. But unlike the preceding procedures, the estimator computation and its assessment are only performed once. Since there is no averaging on several resamples, this simple procedure has been often studied (see [5, 10] in classification, [26, 39] in regression). For a randomly chosen $e \in \mathcal{E}_p$, its simple expression is

$$\widehat{R}_{\mathrm{Hpo},p}(A) = \frac{1}{p} \sum_{i \in e} \gamma\left(A(X^{\overline{e}}_{1,n}), X_i\right).$$

Lpo and VFCV
Nowadays, the always increasing amount of data results in very large sample sizes (several thousands and more). Since the Loo [37] requires the computation of one estimator for each successively removed observation, it may be too computationally demanding. To overcome this problem, Geisser [22] proposed the VFCV algorithm, which only requires the computation of V estimators (as many as we have subsets of data). Thus provided $V \ll n$, it is less expensive to use VFCV than Loo. However, the former relies on a preliminary random partitioning of the data in V subsets. This additional randomness induces some unwanted variability [17]. A similar remark applies to Hpo, since the common intuition about it is that choosing only a subset of the data may be misleading if unfortunately these data are not fully representative of the underlying phenomenon.

Keeping this additional randomness issue in mind, Lpo [34] may appear as the "gold standard". Indeed, it does not introduce any additional variability, since all the $\binom{n}{p}$ resamples are taken into account. To go further, VFCV may be understood as an approximation of the "ideal" Lpo, up to some fluctuations due to the additional randomness the former introduces. Note that other attempts to approximate the Lpo have been proposed such that the repeated learning-testing method [12] for instance.
Nevertheless, the price to pay for such an "optimality" is once more the computational issue. The Lpo computation requires to explore $\binom{n}{p}$ resamples, which is intractable even for relatively small n and p. Fortunately in some quite general settings, closed-form expressions can be derived for the Lpo estimator [17].

In the following, we focus on the study of the Lpo and first provide some closed-form formulas, which make this algorithm *fully tractable*. Besides, this resampling scheme provides some more insight in the general behaviour of CV-based algorithms, for which some further work should be done towards a deeper understanding.

2.3 Closed-form expressions

The purpose of this section is to provide closed-form expressions for the Lpo risk estimator, which can be computed very efficiently.

2.3.1 Leave-p-out risk estimator

Before providing formulas, we need the following lemma yielding the key quantities.

Lemma 2.1. *Let $\widehat{s}_m(X^{\overline{e}}_{1,n})$ denote a generic projection estimator based on model S_m and computed from the training data $X^{\overline{e}}_{1,n}$. Then,*

$$\sum_{e \in \mathcal{E}_p} \left\|\widehat{s}_m(X^{\overline{e}}_{1,n})\right\|_2^2 = \frac{1}{(n-p)^2} \left[\binom{n-1}{p} \sum_{k=1}^{n} \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2(X_k) + \binom{n-2}{p} \sum_{k \neq \ell} \sum_{\lambda \in \Lambda(m)} \varphi_\lambda(X_k)\varphi_\lambda(X_\ell) \right], \tag{1}$$

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

The proof of Lemma 2.1 is deferred to Section 5.

From the previous result, we deduce the following general expression for the Lpo-based estimator of the risk with projection estimators.

Proposition 2.1. *For any $m \in \mathcal{M}_n$, let $\widehat{s}_m$ denote the projection estimator onto the model S_m, spanned by the orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. Then for any $p \in \{1,\dots,n-1\}$,*

$$\widehat{R}_p(m) = \frac{1}{n(n-p)} \sum_{\lambda\in\Lambda(m)} \left[\sum_j \varphi_\lambda^2(X_j) - \frac{n-p+1}{n-1} \sum_{j\neq k} \varphi_\lambda(X_j)\varphi_\lambda(X_k) \right]. \tag{3}$$

REMARK: The computation cost of (3) reduces to that of $\sum_j \varphi_\lambda^2(X_j)$ as well as $\left(\sum_j \varphi_\lambda(X_j)\right)^2$, which is noticeably cheap.

Proof. In the density estimation framework, the contrast associated with the L^2-loss is $\gamma(t,X) = \|t\|^2 - 2t(X)$. Subsequently, the Lpo estimator is

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

Besides, the general projection estimator is

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

The simple application of (1) and (2) provides the expected conclusion. □

Examples
We are now in position to specify the expression of the Lpo risk estimator in Proposition 2.1 with particular projection estimators.

1. Histograms:

   Corollary 2.1. *With the same notations as before, assume that $\widehat{s}_m$ denotes the histogram estimator built from the partition $I(m) = (I_1,\dots,I_{D_m})$ of $[0,1]$ in D_m intervals of respective length $|I_\lambda|$. Then for $p \in \{1,\dots,n-1\}$,*

   $$\widehat{R}_p(m) = \frac{1}{(n-1)(n-p)} \sum_{\lambda=1}^{D_m} \frac{1}{|I_\lambda|} \left[(2n-p)\frac{n_\lambda}{n} - n(n-p+1)\left(\frac{n_\lambda}{n}\right)^2 \right], \tag{4}$$

   where $n_\lambda = \sharp\{i \mid X_i \in I_\lambda\}$.

   Proof. (4) comes simply from the application of (3) with $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$. □

2. Trigonometric polynomials:

   Corollary 2.2. *Let φ_λ denote either $t \mapsto \cos(2\pi kt)$, if $\lambda \in 2\mathbb{N}$ or $t \mapsto \sin(2\pi kt)$, if $\lambda \in 2\mathbb{N}+1$. Let us further assume that $\Lambda(m) = \{0,\dots,2K\}$ for an integer $K > 0$. Then,*

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

3. Haar basis:

Corollary 2.3. *Set* $\varphi:\ t\mapsto \mathbb{1}_{[0,1]}$ *and* $\varphi_{j,k}(t)=2^{j/2}\varphi(2^j\cdot -k)$, *where* $j\in\mathbb{N}$ *and* $0\le k\le 2^j-1$. *For any* $m\in\mathcal{M}_n$, *let us define* $\Lambda(m)\subset\left\{(j,k)\mid j\in\mathbb{N},\ 0\le k\le 2^j-1\right\}$. *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 $n_{j,k}=\mathrm{Card}\left(\left\{i\mid X_i\in[k/2^j,(k+1)/2^j]\right\}\right)$.

2.3.2 Moment calculations

We first deal with general projection estimators for which we provide explicit expectation and variance.

Proposition 2.2. *With the same notations as in Proposition 2.1, we have for any* $1\le p\le n-1$,

$$\begin{aligned}
\mathbb{E}\widehat{R}_p(m) &= \frac{1}{n-p}\sum_{\lambda\in\Lambda(m)}\left[\mathbb{E}\varphi_\lambda^2(X)-(\mathbb{E}\varphi_\lambda(X))^2\right]-\sum_{\lambda\in\Lambda(m)}(\mathbb{E}\varphi_\lambda(X))^2,\\
\mathrm{Var}\left[\widehat{R}_p(m)\right] &= (n(n-1)(n-p))^{-2}\left[2\beta^2t_1\sum_\lambda\left(P\varphi_\lambda^2\right)^2+4\alpha\beta t_1\sum_\lambda P\varphi_\lambda^3P\varphi_\lambda+n\alpha^2\mathbb{E}\left(\sum_\lambda\varphi_\lambda^2\right)^2-\right.\\
&\quad n\alpha^2\left(\sum_\lambda P\varphi_\lambda^2\right)^2+2\beta^2t_1\sum_{\lambda\neq\lambda'}(P\varphi_\lambda\varphi_{\lambda'})^2+4\beta^2t_2\mathbb{E}\left(\sum_\lambda\varphi_\lambda P\varphi_\lambda\right)^2+\\
&\quad\left.(-4n+6)t_1\beta^2\left(\sum_\lambda(P\varphi_\lambda)^2\right)^2+4\alpha\beta t_1\sum_{\lambda\neq\lambda'}P\varphi_\lambda^2\varphi_{\lambda'}P\varphi_{\lambda'}-4t_1\alpha\beta\sum_\lambda P\varphi_\lambda^2\sum_{\lambda'}(P\varphi_{\lambda'})^2\right],
\end{aligned}$$

where $P\varphi_\lambda=\mathbb{E}\varphi_\lambda(X)$, *and*

$$\begin{aligned}
\alpha&=n-1 & \beta&=n-p+1,\\
t_1&=n(n-1) & t_2&=t_1(n-2),
\end{aligned}$$

The technical proof is given in Section 5. Note that these formulas may be derived provided $P\left|\varphi_\lambda\right|^3<+\infty$ for any $\lambda\in\Lambda(m)$, which is satisfied if s is assumed to be bounded and $\int\left|\varphi_\lambda\right|^3<+\infty$ (φ_λ continuous and compact supported for instance).
The bias of the Lpo risk estimator may be a more interesting quantity to work with. From Proposition 2.2, we derive its expression.

Corollary 2.4. *For any projection estimator, the bias of the Lpo estimator is equal to*

$$\begin{aligned}
\mathbb{B}\left[\widehat{R}_p(m)\right]:=\mathbb{E}\widehat{R}_p(m)-R_n(m) &= \frac{p}{n(n-p)}\sum_{\lambda\in m}\left[\mathbb{E}\varphi_\lambda^2(X)-(\mathbb{E}\varphi_\lambda(X))^2\right],\\
&= \frac{p}{n(n-p)}\sum_{\lambda\in\Lambda(m)}\mathrm{Var}\left[\varphi_\lambda(X)\right]\ge 0,
\end{aligned}$$

where $R_n(m)=\mathbb{E}\left[\left\|\widehat{s}_m\right\|^2-2\int_{[0,1]}s\,\widehat{s}_m\right]$.

Illustration
If we apply Proposition 2.2 to histogram estimators, we obtain the following expressions for the expectation and the variance of the Lpo risk estimator:

Corollary 2.5. *For any $\lambda \in \Lambda(m)$, set $\alpha_\lambda = \mathbb{P}(X_i \in I_\lambda)$. Then,*

$$\begin{aligned}
\mathbb{E}\left[\widehat{R}_p(m)\right] &= \frac{1}{n-p}\sum_{\lambda\in m}\frac{1}{\omega_\lambda}\alpha_\lambda\left(1-\alpha_\lambda\right) - \sum_{\lambda\in m}\frac{1}{\omega_\lambda}\alpha_\lambda^2, \\
\mathrm{Var}\left[\widehat{R}_p(m)\right] &= \frac{p^2 q_2(n,\alpha,\omega) + p\, q_1(n,\alpha,\omega) + q_0(n,\alpha,\omega)}{[n(n-1)(n-p)]^2},
\end{aligned}$$

where

$$\begin{aligned}
\forall(i,j) \quad &\in \quad \{1,\ldots,3\}\times\{1,2\}, \quad s_{i,j} = \sum_{k=1}^{D}\alpha_k^i/\omega_k^j, \\
q_2(n,\alpha,\omega) &= n(n-1)\left[2s_{2,2} + 4s_{3,2}(n-2) + s_{2,1}^2(-4n+6)\right], \\
q_1(n,\alpha,\omega) &= n(n-1)\left[-8s_{2,2} - 8s_{3,2}(n-2)(n+1) - 4s_{1,1}s_{2,1}(n-1) - \right. \\
&\quad \left. 2s_{2,1}^2(-4n^2+2n+6)\right], \\
q_0(n,\alpha,\omega) &= n(n-1)\left[s_{1,2}(n-1) - 2s_{2,2}(n^2-2n-3) + \right. \\
&\quad 4s_{3,2}(n-2)(n+1)^2 - s_{1,1}^2(n-1) + \\
&\quad \left. 4s_{1,1}s_{2,1}(n^2-1) + s_{2,1}^2(-4n+6)(n+1)^2\right].
\end{aligned}$$

2.4 Random penalty

Ideal and Lpo penalties

From a collection of models $\{S_m\}_{m\in\mathcal{M}_n}$, the purpose of model selection is to design a procedure which provides us with the "best" candidate model. For instance, this choice is made by minimization of a penalized criterion $\mathrm{crit}(\cdot)$ [3], defined by

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m), \tag{5}$$

where $P_n\gamma\left(\widehat{s}_m\right)$ is the empirical risk of an estimator $\widehat{s}_m$. $\mathrm{pen}(\cdot): \mathcal{M}_n \to \mathbb{R}_+$ denotes the penalty term, which takes into account the complexity of the model S_m.

Ideally, the optimal criterion we would like to minimize over $\mathcal{M}_n$ is the random quantity

$$\mathrm{crit}_{id}(m) = P\gamma(\widehat{s}_m) := \mathbb{E}\gamma\left(\widehat{s}_m, X\right) \tag{6}$$

where the expectation is taken with respect to $X \sim P$, which is independent from the original data. crit_{id} quantifies the mean error incurred by the estimator $\widehat{s}_m$ computed from the observations in hand, at a new point X.

The link between these two criteria (5) and (6) can be made by rewriting

$$\mathrm{crit}_{id}(m) = P_n\gamma(\widehat{s}_m) + \left[P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m)\right],$$

so that we introduce the *ideal penalty*

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_{id}(m) := P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m).$$

On the other hand following the CV strategy, we perform model selection by minimizing the Lpo risk estimator over $\mathcal{M}_n$. Thus for a given $1 \le p \le n-1$, the candidate $\widehat{m}$ satisfies

$$\widehat{m} = \mathrm{Argmin}_{m\in\mathcal{M}_n}\widehat{R}_p(m).$$

The existence of a strong relationship between penalized criteria and CV is strongly supported by the large amount of literature about the (asymptotic) comparison of these two aspects [38, 25, 41]. Therefore, we try to embed the CV strategy into the wider scope of penalized criterion minimization:

$$\widehat{m} = \mathrm{Argmin}_{m\in\mathcal{M}_n}\left\{P_n\gamma(\widehat{s}_m) + \left[\widehat{R}_p(m) - P_n\gamma(\widehat{s}_m)\right]\right\}.$$

The quantity in square brackets is a random penalty that we call *Lpo penalty*:

$$\forall m \in \mathcal{M}_n, \qquad \mathrm{pen}_p(m) := \widehat{R}_p(m) - P_n\gamma(\widehat{s}_m).$$

Note that a somewhat related approach, applied to Loo, can be found in Birgé and Massart [6].

Lpo overpenalization
Thanks to this parallel between CV and penalized criteria, we intend to get more insight in the behaviour of CV techniques, for instance with respect to the parameter p. To this end, we pursue comparison between pen_{id} and pen_p, so that we characterize some features in the behaviour of pen_p with respect to p. This comparison is carried out through the expectations of these penalties, which are both random variables.

The main concern of the following result is to assess the behaviour (in expectation) of the Lpo penalty with respect to the ideal one. This question is addressed with general projection estimators.
We start with a preliminary lemma:

Lemma 2.2. *With the same notations as before with any projection estimator $\widehat{s}_m$ onto S_m, we obtain*

$$\begin{aligned}
\mathbb{E}\left[\,\mathrm{pen}_{id}(m)\,\right] &= \frac{2}{n}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right), \\
\mathbb{E}\left[\,\mathrm{pen}_{p}(m)\,\right] &= \frac{2n-p}{n(n-p)}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right).
\end{aligned}$$

We now state the main assertion about the Lpo penalty associated with projection estimators:

Proposition 2.3. *For any $m \in \mathcal{M}_n$, let $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ denote an orthonormal basis of S_m and $\widehat{s}_m$, the projection estimator onto S_m. Then, we get*

$$\forall m \in \mathcal{M}_n,\ 1 \le p \le n-1, \qquad \mathbb{E}\left[\,\mathrm{pen}_p(m) - \mathrm{pen}_{id}(m)\,\right] = \frac{p}{n(n-p)}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\varphi_\lambda(X)\right) \ge 0.$$

Since this quantity remains nonnegative whatever p, we conclude that the Lpo penalty always overpenalizes, which remains true for any orthonormal basis. Moreover, the amount of overpenalization increases with p. Thus, the Loo provides the weakest overpenalization of order $\mathcal{O}\left(1/n^2\right)$, whereas the Lpo with $p \simeq n/2$ (which is similar to the 2-fold CV) corresponds to an overpenalization of the same order as the expectation of the ideal penalty, that is $\mathcal{O}\left(1/n\right)$.

3 Oracle inequalities

In the following, we assess the quality of the Lpo-based model selection procedure through the statement of oracle inequalities. These results are settled in the polynomial complexity framework and hold for any projection estimator. To our knowledge, it is the first non-asymptotic results about the performance of the Lpo algorithm in this framework.
Remark: We point out that unlike the usual approach in model selection via penalization, our purpose is not to design a penalty function since the Lpo estimator itself can be understood as a penalized criterion (Section 2.4).

3.1 Preliminaries

Our main results rely on several assumptions that we now detail and discuss.
Set $X \sim s$ and for any index m,

$$\phi_m = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 \qquad \text{and} \qquad V_m = \mathbb{E}\,\phi_m(X).$$

Then, we define the following assumptions

$$(Reg) \qquad \exists \Phi > 0 / \quad \sup_{m \in \mathcal{M}_n} \|\phi_m\|_\infty \leq \Phi \, n / \left(\log n\right)^2,$$

$$(Reg2) \qquad \exists \Phi > 0 / \quad \sup_{m \in \mathcal{M}_n} \left\{ \sup_{(a_\lambda)_\lambda, |a|_\infty = 1} \left\| \sum_\lambda a_\lambda \varphi_\lambda \right\|_\infty \right\} \leq \sqrt{\Phi \, n / \left(\log n\right)^2},$$

$$(Ad) \qquad \exists \xi > 0 / \; \forall m \in \mathcal{M}_n \text{ with } D_m \geq 2, \quad n \mathbb{E}\left[\|s_m - \widehat{s}_m\|_2^2 \right] \geq \xi D_m,$$

$$(Pol) \qquad \exists \delta > 0 / \; \forall D \geq 1, \; |\{m \in \mathcal{M}_n \mid D_m = D\}| \leq D^\delta.$$

Since $\phi_m = \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2$, $\|\phi_m\|_\infty$ may be understood as a regularity measure of the basis $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$. Thus, (Reg) relates the regularity of the considered basis to the amount of data. This assumption has already been used by Castellan [16] for instance. Let us assume we use histogram estimators based on a partition $\{I_1, \ldots, I_{D_m}\}$ of $[0,1]$ in D_m intervals, and that $\varphi_\lambda = \mathbb{1}_{I_\lambda} / \sqrt{|I_\lambda|}$, where $|I_\lambda|$ is the length of I_λ. Then, (Reg) gives a lower bound on the minimal length of any interval I_λ of the partition with respect to the number of observations. In other words, we cannot consider partitions made of intervals with less than $n/(\log n)^2$ observations.

$(Reg2)$ is another regularity assumption about $\{\varphi_\lambda\}_{\lambda \in \Lambda(m)}$. In the specific case of a basis defined from a partition of $[0,1]$ (like histograms or piecewise polynomials), (Reg) implies $(Reg2)$. However, this does no longer hold with general bases of functions. Besides, the constant Φ is not necessarily the same in both (Reg) and $(Reg2)$. However replacing one of them by the maximum value provides the same constant, which is assumed in the following. A similar requirement to $(Reg2)$ can be found in Massart [28].

Developing $\mathbb{E}\|s_m - \widehat{s}_m\|^2$, we observe that

$$\begin{aligned} \mathbb{E}\|s_m - \widehat{s}_m\|^2 &= \sum_{\lambda \in \Lambda(m)} \mathbb{E}\left[\nu_n^2(\varphi_\lambda) \right], \\ &= \sum_{\lambda \in \Lambda(m)} \frac{1}{n} \mathrm{Var}\left[\varphi_\lambda(X) \right]. \end{aligned}$$

For instance if we use histograms, $\mathrm{Var}\left[\varphi_\lambda(X)\right]$ vanishes if and only if the support of s is included in I_λ. (Ad) therefore requires that for any m, there are always "enough" informative basis vectors, if an informative vector is a vector such that $\mathrm{Var}\left[\varphi_\lambda(X)\right] \neq 0$. With histograms, it means that we choose bases with mainly more and more vectors where $s \neq 0$. For instance, (Ad) holds with histograms if $s \geq \rho > 0$ on $[0,1]$. This assumption can also be found in [28].

A model collection is said to have a polynomial complexity if (Pol) holds. At most, the cardinality of the set of models with dimension D is polynomial in D. Such an assumption is satisfied with nested models for instance ([6]). It straightforwardly implies that $\mathrm{Card}\left(\mathcal{M}_n\right) \leq n^{\delta+1}$.

3.2 Main results

In the present section, we provide two oracle inequalities, which warranty the ability of the Lpo-based procedure to select an effective density estimator. The first result applies to bounded densities, while the second one concerns the more general case of square integrable densities, at the price of an additional assumption. In the following, we will provide several examples of widely used bases for which the latter assumption is satisfied.

Bounded density

Theorem 3.1. *Let s denote a bonded density on [0,1] and $X_1, \dots, X_n$ be n i.i.d. random variables drawn from s. Set $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ a finite family of bounded functions on $[0,1]$ such that for any $m \in \mathcal{M}_n$, S_m denotes the vector space of dimension D_m, spanned by the orthonormal family $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. Let us assume that (Reg) , $(Reg2)$, (Ad) and (Pol) hold.*
For $n \geq 29$, set $0 < \epsilon < 1$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} < 1\,, \tag{7}$$

where $\zeta(\epsilon) = \left[1-(1+\epsilon)^{-8}\right]$. Then for any $1 \leq p \leq n-1$ satisfying

$$(Ran) \qquad \frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n}\frac{1+\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \alpha \leq \frac{p}{n} \leq 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} - \beta$$

with $0 < \alpha, \beta < 1$, we have

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq \Gamma(\epsilon,\alpha,\beta) \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] + \frac{\kappa(\epsilon,s,\Phi,\alpha,\beta,\delta)}{n},$$

where $\Gamma(\epsilon,\alpha,\beta) \geq 1$ is a constant (with respect to n) independent from s and $\kappa(\epsilon,s,\Phi,\alpha,\beta,\delta) \geq 0$ is another constant.

The proof of this result is deferred to Section 5.
REMARKS:

- (Ran) is a *sufficient condition* for the oracle inequality to hold. In this assumption, α and β can be chosen as small as we want, but cannot vanish.
- The existence of ϵ satisfying the inequality (7) stems from a technical lemma given in the proof of Theorem 3.1.
- As it is made clear from the proof of the aforementioned technical lemma, the choice of ϵ is constrained. For instance, ϵ cannot be too much close to 0. This explains why the nonintuitive bounds in (Ran) cannot be easily simplified. Furthermore, this enlightens that "small values" of p could be excluded from the range of values described in (Ran), to which the oracle inequality applies.
- The independence of $\Gamma(\epsilon,\alpha,\beta)$ from s is essential in our framework since we have in mind the use of this result to derive some adaptivity in the minimax sense properties.

Square-integrable density

The second result is derived following the same idea as the previous one, thanks to an additional mild assumption on the considered bases. This requirement turns out to be non restrictive at all, since it is met by a broad class of orthonormal bases.

Theorem 3.2. *Let s denote a density in $L^2([0,1])$ and $X_1, \dots, X_n$ be n i.i.d. random variables drawn from s. We set $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$ a finite family of bounded functions on $[0,1]$ such that for any $m \in \mathcal{M}_n$, S_m denotes the vector space of dimension D_m, spanned by the orthonormal family $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. Let us assume that (Reg) , $(Reg2)$, (Ad) and (Pol) hold, and moreover that*

$$(Reg3) \qquad \exists \Phi > 0 / \ \forall m \in \mathcal{M}_n, \qquad \|\phi_m\|_\infty \leq \Phi D_m.$$

For $n \geq 29$, set $0 < \epsilon < 1$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} < 1\,,$$

where $\zeta(\epsilon) = \left[1-(1+\epsilon)^{-8}\right]$. Then for any $1 \leq p \leq n-1$ satisfying

$$(Ran) \qquad \frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n}\frac{1+\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \alpha \leq \frac{p}{n} \leq 1 - \frac{2}{\zeta(\epsilon)(n-1)-2} - \beta$$

with $0 < \alpha, \beta < 1$*, we have*

$$\mathbb{E}\left[\left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] \leq \Gamma(\epsilon, \alpha, \beta) \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\left\|s - \widehat{s}_m\right\|^2\right] + \frac{\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta)}{n},$$

where $\Gamma(\epsilon, \alpha, \beta) \geq 1$ *is a constant (with respect to* n*) independent from* s *and* $\kappa(\epsilon, s, \Phi, \alpha, \beta, \delta) \geq 0$ *is another constant.*

For the sake of clarity, the proof is also deferred to Section 5. Moreover, it is very similar to that of Theorem 3.1, so that we only detail the main differences along the proof.

Remark: Assumption $(Reg3)$ is quite different from (Reg). Whereas the latter relates the "regularity" of any basis to the number of observations uniformly over $\mathcal{M}_n$, $(Reg3)$ rather controls $\|\phi_m\|_\infty$ for each model with respect to its dimension. Every models with the same dimension must be somehow alike in that their associated sup-norm $\|\phi_m\|_\infty$ remain upper bounded by ΦD_m.
This assumption can be also found in Birgé and Massart [6].

Examples
We now illustrate the high level of generality of assumption $(Reg3)$ thanks to several examples of widespread functional bases to which $(Reg3)$ applies.

- It is easy to check that $(Reg3)$ applies to *regular* histograms with $\Phi = 1$ (Section 4.3).
- A typical example of basis satisfying $(Reg3)$ is the trigonometric basis. For an integer m, let $\Lambda(m) = \{0, \ldots, 2m\}$ denote a set of indices where $\varphi_0 = \mathbb{1}_{[0,1]}$, $\varphi_\lambda(t) = \sqrt{2}\sin(2k\pi t)$ if $\lambda = 2k - 1$ and $\varphi_\lambda(t) = \sqrt{2}\cos(2k\pi t)$ if $\lambda = 2k$.
  Then,

$$\forall t \in [0,1], \qquad \sum_{\lambda \in \Lambda(m)} \varphi_\lambda^2(t) = 1 + 2\sum_{k=1}^{m} \left(\cos^2(2k\pi t) + \sin^2(2k\pi t)\right), \\ = 2m + 1.$$

  Since $D_m = 2m + 1$, it comes that $\|\phi_m\|_\infty = D_m$ and $(Reg3)$ holds with $\Phi = 1$.
- Barron *et al.* [3] (Lemma 7.13) proved that with piecewise polynomials on a regular partition of $[0,1]$ with degree not larger than r on each element of this partition,

$$\|\phi_m\|_\infty \leq (r+1)(2r+1)D_m.$$

  The resulting constant $\Phi = (r+1)(2r+1)$ is subsequently independent from m.
- Haar basis: For any positive integer j, we introduce $\Lambda(j) = \left\{(j,k) \mid 0 \leq k \leq 2^j - 1\right\}$. Furthermore, set $\varphi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1]}$ and for any $\lambda = (j,k)$, let us define $\varphi_{j,k}(t) = 2^{j/2}\varphi\left(2^j t - k + 1\right)$ on $[0,1]$. For a positive integer $m \in \mathcal{M}_n$, let us consider S_m as the linear space spanned by $\{\varphi_\lambda\}_{\lambda \in \cup_{j \leq m} \Lambda(j)}$. Then, it can be seen that

$$\|\phi_m\|_\infty = D_m$$

  since for each j, there is only one $0 \leq k \leq 2^j - 1$, which contributes to the sum in ϕ_m.
  For more general wavelet bases, an upper bound, uniform with respect to m, can be established [6].

4 Adaptivity

In this section, the idea is to apply theorems of Section 3 to derive several adaptivity results in the minimax sense with respect to Hölder as well as Besov functional spaces.

4.1 Adaptivity in the minimax sense

Let us assume that s belongs to a set of functions $\mathcal{T}(\theta)$ indexed by a parameter $\theta \in \Theta$. Moreover, let us define an estimator $\widehat{s}$ of s.
An estimator $\widehat{s}$ is said to be *adaptive for* θ if, without knowing θ, it "works as well as" any estimator which would exploit this knowledge. In the present work, the effectiveness measure is the L^2-risk and we say that an estimator enjoys such an adaptivity property provided its risk is nearly the same (up to some constants) as the *minimax risk* with respect to $\mathcal{T}(\theta)$:

$$\inf_{\widehat{s}} \sup_{s\in\mathcal{T}(\theta)} \mathbb{E}\left[\|s-\widehat{s}\|^2\right] \leq \sup_{s\in\mathcal{T}(\theta)} \mathbb{E}\left[\|s-\widehat{s}_{\widehat{m}}\|^2\right] \leq C \inf_{\widehat{s}} \sup_{s\in\mathcal{T}(\theta)} \mathbb{E}\left[\|s-\widehat{s}\|^2\right],$$

where the infimum is taken over all possible estimators.
Remark: Very often, $C \geq 1$ depends on the unknown parameters θ, but neither from s nor from n.
Furthermore if this property holds for every parameters θ in a set Θ, then $\widehat{s}$ is said to be *adaptive in the minimax sense with respect to the family* $\{\mathcal{T}(\theta)\}_{\theta\in\Theta}$. We refer to Barron *et al.* [3] for a unified presentation about various notions of adaptivity.

4.2 Description of the collections of models

Since such optimality results depend on the approximation properties of the models we use, we describe three different model collections, each one being defined from a specific family of vectors $\{\varphi_\lambda\}_{\lambda\in\Lambda_n}$.

4.2.1 Piecewise constant functions (Pc)

For a given partition of $[0,1]$ in D regular intervals $(I_\lambda)_{\lambda\in\Lambda(m)}$ of length $1/D$ and $m \in \mathcal{M}_n$, let us define the model

$$S_m = \left\{ t \mid t = \sum_{\lambda\in\Lambda(m)} a_\lambda \varphi_\lambda,\ (\alpha_\lambda)_\lambda \in \mathbb{R} \right\},$$

where $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$ and $|I_\lambda|$ denotes the length of I_λ. S_m is the vector space of dimension $D_m = D$ spanned by the orthonormal family $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$. It is made of all piecewise constant functions defined on the partition $I = (I_1, \ldots, I_{D_m})$.
Thus with each index $m \in \mathcal{M}_n$, we associate the linear space S_m of piecewise constant functions defined on a *regular partition* of $[0,1]$ in D_m intervals of length $1/D_m$. Moreover, let $N_n = \max_{m\in\mathcal{M}_n} D_m$ be the maximal dimension of a model belonging to the collection.

4.2.2 Piecewise dyadic polynomials (Pp)

Set $\mathcal{M}_n = \{0, \ldots, J_n\}$ and for any $m \in \mathcal{M}_n$, S_m denotes the linear space of functions

$$t = \sum_{k=0}^{2^m-1} P_k\, \mathbb{1}_{[k2^{-m},(k+1)2^{-m})},$$

where the P_ks denote polynomials of degree less than r. The dimension of S_m is subsequently defined by

$$D_m = r\,2^m \qquad \text{and} \qquad N_n = \max_{m\in\mathcal{M}_n} D_m = r\,2^{J_n}.$$

Remark: With this collection of models, (Pol) is satisfied since there is at most one model for each dimension.

4.2.3 Trigonometric polynomials (Tp)

Set $\mathcal{M}_n = \{0, \ldots, J_n\}$, where J_n is a positive integer. For any $m \in \mathcal{M}_n$, let $\Lambda(m) = \{0, \ldots, 2m\}$ denote a set of indices such that $\varphi_0(t) = \mathbb{1}_{[0,1]}$, $\varphi_\lambda(t) = \sqrt{2}\sin(2k\pi t)$ if $\lambda = 2k-1$ and $\varphi_\lambda(t) = \sqrt{2}\cos(2k\pi t)$ if $\lambda = 2k$.
Then, the model S_m is the linear space spanned by $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$, of dimension $D_m = 2m+1$. Any $t \in S_m$ can be expressed as

$$\forall x \in [0,1], \qquad t(x) = a_0 + \sum_{k=1}^{m} \left[a_k \sqrt{2}\cos(2\pi k x) + b_k \sqrt{2}\sin(2\pi k x) \right],$$

the a_ks and b_ks belong to $\mathbb{R}$.
Moreover, J_n and N_n are related by the following relationship $N_n = 2J_n + 1$.

4.3 Hölder functional space

Our purpose is to show that the Lpo-based approach enjoys some adaptivity when s belongs to an unknown Hölder space $\mathcal{H}(L,\alpha)$ for $L > 0$ and $\alpha \in (0,1]$. We recall that a function $f : [0,1] \to \mathbb{R}$ belongs to $\mathcal{H}(L,\alpha)$ with $L > 0$ and $0 < \alpha \leq 1$ if

$$\forall x, y \in [0,1], \quad |f(x) - f(y)| \leq L\,|x-y|^\alpha .$$

We refer to De Vore and Lorentz [18] for an extensive study of a wide range of functional spaces.

In order to reach this goal, we approximate s by piecewise constant functions, using the model collection (**Pc**) described in Section 4.2.1.
The histogram estimator built from model S_m is defined by

$$\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} P_n\varphi_\lambda\ \varphi_\lambda = \sum_{\lambda\in\Lambda(m)} \frac{n_\lambda}{n}\frac{\mathbb{1}_{I_\lambda}}{|I_\lambda|},$$

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

Since the adaptivity property results from the oracle inequalities in Section 3, we have to check the different assumptions it relies on.

- With the collection (**Pc**), $m \mapsto D_m$ is a one-to-one mapping from $\mathcal{M}_n$ towards $\mathcal{D} = \{D_m \mid m \in \mathcal{M}_n\}$, which entails that (Pol) is satisfied since our collection is made of only one model for each dimension.

- Since $\varphi_\lambda = \mathbb{1}_{I_\lambda}/\sqrt{|I_\lambda|}$,

$$\begin{aligned}\|\phi_m\|_\infty &= \sum_{t\in[0,1]} \left(\sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2(t) \right), \\ &= \max_{\lambda\in\Lambda(m)} \frac{1}{|I_\lambda|} = D_m.\end{aligned}$$

  Thus, (Reg) amounts to require that

$$\max_m D_m = N_n \leq \Phi\, n/\left(\log n\right)^2,$$

  which means that on average, there are at least about $\left(\log n\right)^2/n$ points in each interval of any partition we consider.

- We therefore assume that (Reg) , (Ad) and (Ran) hold.

As for the problem of density estimation on $[0,1]$ when s belongs to some Hölder space, it is known since the early 80s, thanks o Ibragimov and Khas'minskij [23], that the minimax rate with respect to $\mathcal{H}(L,\alpha)$ for the quadratic risk is of order $L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$, with any $L>0$ and $\alpha>0$.
REMARK: However when the problem is the estimation over $\mathbb{R}$, things turn out to be very different. For instance, the minimax rate now depends on the value of regularity parameter α with respect to the parameter p of the L^p-norm used for the assessment [24].

The following result settles that, applied to the collection of models (**Pc**), the Lpo-based procedure yields an adaptive in the minimax sense estimator of the density on $[0,1]$.

Theorem 4.1. *Let us assume that (Reg) , (Ad) and (Ran) hold and that the collection of models is that one denoted by (**Pc**). Furthermore, assume that the target density $s \in \mathcal{H}(L,\alpha)$ for $L>0$ and $\alpha \in (0,1]$. Then,*

$$\sup_{s\in\mathcal{H}(L,\alpha)} \mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq K_\alpha L^{\frac{2}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{8}$$

for a given constant K_α independent from n and s.

Since the minimax risk is of order $L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$, we deduce from this result that $\widehat{s}_{\widehat{m}}$ is adaptive in the minimax sense with respect to $\{\mathcal{H}(L,\alpha)\}_{L>0,\alpha\in(0,1]}$.

REMARK: As we will see in the proof, this result remains true with any polynomial collection of models satisfying the requirements of Theorem 3.1, and including models with dimension of the order of $L^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}$.

Proof. The idea is simply to use Theorem 3.1 and to derive the upper bound from

$$\mathbb{E}\left[\|s - \widehat{s}_m\|^2 \right] = \|s - s_m\|^2 + \mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right].$$

For the bias term, we have

$$\begin{aligned}
\|s - s_m\|^2 &= \sum_{\lambda\in\Lambda(m)} \frac{1}{|I_\lambda|^2} \int_{I_\lambda} \left(\int_{I_\lambda} [\, s(t) - s(x)\,]\, dx \right)^2 dt, \\
&\leq \sum_{\lambda\in\Lambda(m)} L^2 D_m^2 \int_{I_\lambda} \left(\int_{I_\lambda} |t-x|^\alpha\, dx \right)^2 dt \qquad (s\in\mathcal{H}(L,\alpha)), \\
&\leq C_\alpha\, L^2\, D_m^{-2\alpha} \quad \text{(after integration)},
\end{aligned}$$

where $C_\alpha = 4(\alpha+2)\left[(1+\alpha)^2 (2\alpha+3) \right]^{-1}$.
On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right] &= \frac{V_m - \|s_m\|^2}{n}, \\
&\leq \frac{V_m}{n}, \\
&\leq \frac{\|\phi_m\|_\infty}{n}, \\
&= \frac{\sup_{x\in[0,1]} \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2(x)}{n} = \frac{D_m}{n}.
\end{aligned}$$

Hence under the same assumptions as Theorem 3.1, we get that there exists $C\geq 1$ and $\kappa>0$ such that

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq C\left(C_\alpha \inf_{m\in\mathcal{M}_n} \left\{ L^2\, D_m^{-2\alpha} + \frac{D_m}{n} \right\} \right) + \frac{\kappa}{n}.$$

Now, let us define the sequence $\{D_{m_n}\}_n$ such that for each n,

$$\frac{1}{2} L^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}} \leq D_{m_n} \leq 2L^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}.$$

Then, we derive that it exists $K'_\alpha > 0$ such that

$$\inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2 \right] \leq C_\alpha L^2 D_{m_n}^{-2\alpha} + \frac{D_{m_n}}{n} \leq K'_\alpha L^{\frac{2}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}},$$

hence the expected result. □

4.4 Besov functional spaces

Like the previous one, the present section aims at deriving adaptivity in the minimax sense, except that we essentially focus on Besov spaces (rather than Hölder ones). This goal is reached thanks to results of Section 3 as well.

4.4.1 Overview of Besov spaces

We start by briefly recalling in what Besov spaces and balls consist in. We refer to the book by De Vore and Lorentz [18] for an extensive presentation on this matter.

For $\alpha > 0$ and $0 < p \leq +\infty$, we say that a function f in $L^p([0,1])$ belongs to the Besov space $\mathcal{B}^\alpha_{\infty,p} = \mathcal{B}^\alpha_\infty(L^p([0,1]))$ if $|f|_{\mathcal{B}^\alpha_{\infty,p}} < +\infty$, where

$$|f|_{\mathcal{B}^\alpha_{\infty,p}} := \sup_{t>0} \left\{ t^{-\alpha} \omega_r(f,t)_p \right\}, \qquad r = [\alpha] + 1,$$

with

$$\omega_r(f,t)_p := \sup_{|h| \leq t} \|\Delta^r_h(f,\cdot)\|_p, \qquad \text{and} \qquad \Delta^r_h(f,x) := \sum_{k=1}^{r} \binom{k}{r} (-1)^{r-k} f(x+kh).$$

$|\cdot|_{\mathcal{B}^\alpha_{\infty,p}}$ defines a semi-norm, while the metric is provided by the following Besov norm

$$\|f\|_{\mathcal{B}^\alpha_{\infty,p}} := |f|_{\mathcal{B}^\alpha_{\infty,p}} + \|f\|_p.$$

Moreover for a given real $R > 0$, let us define the Besov ball of radius R by

$$\mathcal{B}^\alpha_{\infty,p}(R) = \left\{ f \in L^p \mid \|f\|_{\mathcal{B}^\alpha_{\infty,p}} \leq R \right\}.$$

As far as we are concerned in the sequel, we restrict ourselves to the particular case where $p = 2$, that is $\mathcal{B}^\alpha_{\infty,2}$ for $\alpha > 0$.

4.4.2 Piecewise and trigonometric polynomials

A desirable property for an estimator of s is the minimaxity over a set of functions with a given smoothness. Since the amount of smoothness is unknown in advance, an "ideal" estimator $\widehat{s}$ should be designed so that it automatically adapts to the unknown smoothness.
The main interest of model selection procedures, for which an oracle inequality can be stated, lies in that the final estimator $\widehat{s}_{\widehat{m}}$ enjoys the called *adaptivity property in the minimax sense.*
In the same way as in Section 4.3, our strategy consists in deriving adaptivity results from the oracle inequalities of Section 3. We point out that adaptivity heavily relies on the involved model collection through its approximation properties.
The following results therefore state adaptivity in the minimax sense for both (**Pp**) and (**Tp**) collections, with respect to respectively different Besov spaces.

Let us start with adaptivity with respect to Besov balls $\mathcal{B}^\alpha_{\infty,2}(R)$ for $0 < \alpha < r$, where r denotes the smallest integer larger than the degree of polynomials in (**Pp**).

Theorem 4.2. *Let us consider the collection of models (**Pp**) made of piecewise polynomials of degree less than r and assume that (Reg), $(Reg3)$, (Ad) and (Ran) hold.*
Then for $R > 0$ and $0 < \alpha < r$,

$$\sup_{s \in \mathcal{B}^{\alpha}_{\infty,2}(R)} \mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq C_\alpha R^{\frac{2}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{9}$$

where C_α denotes a given constant independent from n and s.

Proof. The proof follows the same strategy as that of Theorem 4.1 in that it essentially relies on approximation properties of models in (**Pp**).

If S_m denotes a model of dyadic piecewise polynomials of degree less than r on each one of the 2^m regular dyadic intervals, the result in page 359 of DeVore and Lorentz [18] states that provided $r > \alpha$,

$$\inf_{u \in S_m} \|s - u\|^2 \leq K_{\alpha,r} |s|^2_{B^{\alpha}_{\infty,2}} (D_m)^{2\alpha},$$

for a positive constant $K_{\alpha,r}$.
Since $s \in \mathcal{B}^{\alpha}_{\infty,2}(R)$, it comes

$$\|s - s_m\|^2 \leq K_{\alpha,r} R^2 (D_m)^{2\alpha}.$$

As for the variance term,

$$\mathbb{E}\left[\|s_m - \widehat{s}_m\|^2 \right] \leq \frac{\|\phi_m\|_\infty}{n} \leq \frac{\Phi D_m}{n} \qquad (\text{by } (Reg3)).$$

Under (Reg), $(Reg3)$, (Ad) and (Ran) we apply Theorem 3.2 to derive

$$\mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq \Gamma\left(K'_{\alpha,r} \inf_{m \in \mathcal{M}_n} \left\{ R^2 D_m^{-2\alpha} + \frac{D_m}{n} \right\} \right) + \frac{\kappa}{n},$$

where $K'_{\alpha,r}$ is a positive constant.
The conclusion results from the same calculation as in the proof of Theorem 4.1 with

$$\frac{1}{2} R^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}} \leq D_{m_n} \leq 2R^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}.$$

□

Unlike the previous result, we now turn to Besov balls $\mathcal{B}^{\alpha}_{\infty,2}(R)$ for any value of $\alpha > 0$, which is enabled by the use of trigonometric polynomials in (**Tp**).

Theorem 4.3. *Let us consider the collection (**Tp**) made of trigonometric polynomials and assume that (Reg), $(Reg2)$, $(Reg3)$, (Ad) and (Ran) hold.*
Then for $R > 0$ and $\alpha > 0$,

$$\sup_{s \in \mathcal{B}^{\alpha}_{\infty,2}(R)} \mathbb{E}\left[\|s - \widehat{s}_{\widehat{m}}\|^2 \right] \leq C'_\alpha R^{\frac{2}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}} + O\left(\frac{1}{n}\right), \tag{10}$$

for a given constant C'_α independent from n and s.

Proof. The same scheme of proof is used, except we need for an approximation result applying to trigonometric polynomials, which is also provided in page 205 of the book by DeVore and Lorentz [18]. Indeed if we consider models in collection (**Tp**) for any $\alpha > 0$, we get

$$\inf_{u \in S_m} \|s - u\|^2 \leq K_\alpha |s|^2_{B^{\alpha}_{\infty,2}} (D_m)^{2\alpha},$$

for a constant $K_\alpha > 0$. Assumption $(Reg3)$ enables to conclude as in the previous theorem. □

5 Proofs

5.1 Closed-form Lpo estimator

5.1.1 Proof of Lemma 2.1

The first remark is that for each $e \in \mathcal{E}_p$, we have $\forall t \in [0,1]$,

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

Then, the Lemma follows from the following combinatorial results

Lemma 5.1. *For any $i \neq j \neq k \in \{1,\dots,n\}$,*

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

where we stress that the sum is made over the resamples i, j and k are kept fixed.

Proof. $\sum_{e\in\mathcal{E}_p}\mathbb{1}_{(j\in\overline{e})}$ may be interpreted as the number of subsets of $\{1,\dots,n\}$ of size p (denoted by e) which do not contain j, since $j \in \overline{e}$. Thus, it is the number of possible choices of p non ordered and different elements among $n-1$.
The other equalities follow from a similar argument. □

5.2 Moments calculations

5.2.1 Proof of Proposition

2.2 The expectation is a straightforward consequence of (3).
The variance calculation is not difficult, but very technical. We only give the main step of this proof. First, let us define $A_\lambda = \sum_{j=1}^{n}\varphi_\lambda^2(X_j)$ and $B_\lambda = \sum_{j\neq k}\varphi_\lambda(X_j)\varphi_\lambda(X_k)$. Set $\alpha = n-1$ and $\beta = n-p+1$, such that

$$n(n-1)(n-p)\widehat{R}_p(m) = \sum_{\lambda}\left(\alpha A_\lambda + \beta B_\lambda\right).$$

Then,

$$\left[\sum_{\lambda}\left(\alpha A_\lambda + \beta B_\lambda\right)\right]^2 = \sum_{\lambda}\left(\alpha^2 A_\lambda^2 + \beta^2 B_\lambda^2 + 2\alpha\beta A_\lambda B_\lambda\right) + \sum_{\lambda\neq\lambda'}\left(\alpha^2 A_\lambda A_{\lambda'} + \beta^2 B_\lambda B_{\lambda'} + 2\alpha\beta A_\lambda B_{\lambda'}\right).$$

After some calculation, the different terms are respectively equal to

$$\mathbb{E}\sum_{\lambda} A_\lambda^2 = \sum_{\lambda}\left[\, n\, P\varphi_\lambda^4 + t_1\left(P\varphi_\lambda^2\right)^2\right],$$

$$\mathbb{E}\sum_{\lambda} B_\lambda^2 = \sum_{\lambda}\left[\, 4t_2 P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2 + 2t_1\left(P\varphi_\lambda^2\right)^2 + t_3\left(P\varphi_\lambda\right)^4\right],$$

$$\mathbb{E}\sum_{\lambda} A_\lambda B_\lambda = \sum_{\lambda}\left[\, 2t_1 P\varphi_\lambda^3 P\varphi_\lambda + t_2 P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\right],$$

$$\mathbb{E}\sum_{\lambda\neq\lambda'} A_\lambda A_{\lambda'} = n\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\right)^2 - \sum_{\lambda} P\varphi_\lambda^4\right] + t_1\left[\left(\sum_{\lambda} P\varphi_\lambda^2\right)^2 - \sum_{\lambda}\left(P\varphi_\lambda^2\right)^2\right],$$

$$\mathbb{E}\sum_{\lambda\neq\lambda'} B_\lambda B_{\lambda'} = 2t_1\sum_{\lambda\neq\lambda'}\left(P\varphi_\lambda\varphi_{\lambda'}\right)^2 + 4t_2\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda(X)P\varphi_\lambda\right)^2 - \sum_{\lambda} P\varphi_\lambda^2\left(P\varphi_\lambda\right)^2\right] + t_3\left[\left(\sum_{\lambda}\left(P\varphi_\lambda\right)^2\right)^2 - \sum_{\lambda}\left(P\varphi_\lambda\right)^4\right],$$

$$\mathbb{E}\sum_{\lambda\neq\lambda'} A_\lambda B_{\lambda'} = 2t_1\sum_{\lambda\neq\lambda'} P\varphi_\lambda^2\varphi_{\lambda'}P\varphi_{\lambda'} + t_2\left[\mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\right)\sum_{\lambda'}\left(P\varphi_{\lambda'}\right)^2 - \mathbb{E}\left(\sum_{\lambda}\varphi_\lambda^2(X)\left(P\varphi_\lambda\right)^2\right)\right].$$

On the other hand,

$$\left(n(n-1)(n-p)\mathbb{E}\left[\,\widehat{R}_p(m)\,\right]\right)^2 = n^2\alpha^2\left(\sum_{\lambda} P\varphi_\lambda^2\right)^2 + t_1^2\beta^2\left(\sum_{\lambda}\left[\,P\varphi_\lambda\,\right]^2\right)^2 + 2n\alpha\beta t_1\left(\sum_{\lambda} P\varphi_\lambda^2\right)\sum_{\lambda'}\left(P\varphi_{\lambda'}\right)^2.$$

Combining these two expressions yields the variance after some simplifications.

5.2.2 Proof of Corollary 2.4

We have to compute $R_n(m)$ for any model m.

$$\begin{aligned} R_n(m) &:= \mathbb{E}\left[\,\|\widehat{s}_m\|^2\,\right] - 2\mathbb{E}\left[\int_{[0,1]} s\widehat{s}_m\right], \\ &= \sum_{\lambda}\mathbb{E}\left(P_n\varphi_\lambda\right)^2 - 2\sum_{\lambda}\left(P\varphi_\lambda\right)^2, \\ &= \frac{1}{n}\sum_{\lambda}\mathrm{Var}\left(\varphi_\lambda(X)\right) - \sum_{\lambda}\left(P\varphi_\lambda\right)^2. \end{aligned}$$

5.3 Theorem 3.1

5.3.1 Outline of the strategy

Let us first describe the outlines of our strategy. We start with the definition of $\widehat{s}_{\widehat{m}}$ as the minimizer of the Lpo risk estimator, which leads to an inequality (11) written so as we stress the discrepancy between the Lpo estimator and its expectation, for each model in the collection. Then, we show that this discrepancy can be studied on a set of high probability (Lemma 5.4) rather than on the whole space . The gap between the Lpo risk and its expectation is evaluated through the use of two concentration inequalities: Bernstein's and a version of Talagrand's inequality (Proposition 5.1 and Proposition 5.2). By recombination of these different results, we derive the main inequality which holds except on a set of small probability (16). The conclusion results from the following lemma:

Lemma 5.2. *Let X and Y be two random variables such that $\forall z>0,\quad \mathbb{P}\left(X \geq Y + K_1 z + K_2\right) \leq \Sigma e^{-z}$, where $K_1,\ K_2,\ \Sigma > 0$. Then, we have*

$$\mathbb{E}X \leq \mathbb{E}Y + K_1\,\Sigma + K_2.$$

Proof. Set $Z = X - Y - K_2$. We have

$$\mathbb{P}\left(Z \geq K_1 z\right) \leq \Sigma e^{-z}.$$

Then,

$$\begin{aligned}
\mathbb{E}Z &\leq \mathbb{E}\left[\int_0^{+\infty} \mathbb{1}_{(t\leq Z)}\, dt\right], \\
&= \int_0^{+\infty} \mathbb{E}\left[\,\mathbb{1}_{(t\leq Z)}\,\right] dt, \\
&= \int_0^{+\infty} \mathbb{P}\left[\,t \leq Z\,\right] dt, \\
&\leq K_1 \int_0^{+\infty} \Sigma e^{-z}\, dz = K_1\Sigma.
\end{aligned}$$

□

5.3.2 Preliminaries

Notations First of all, let us define a few notations that will be useful in the sequel.
For any $p \in \{1, \ldots, n-1\}$ the Lpo risk estimator associated with the estimator $\widehat{s}_m$ is denoted by $\widehat{R}_p(m)$. For the sake of clarity, we define

$$\forall m, \qquad L_p(m) = \mathbb{E}\widehat{R}_p(m),$$

such that $L_p(\widehat{m}) := \mathbb{E}\left[\,\widehat{R}_p(m)\,\right]_{|m=\widehat{m}}$. For each m, $\{\varphi_\lambda\}_{\lambda\in\Lambda(m)}$ denotes an orthonormal basis of S_m. Moreover, we set

$$\begin{aligned}
\phi_m = \sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 \qquad &\text{and} \qquad V_m = \mathbb{E}\left[\,\phi_m(X)\,\right], \\
s_m = \sum_{\lambda\in\Lambda(m)} \beta_\lambda\varphi_\lambda \qquad &\text{and} \qquad \beta_\lambda = P\varphi_\lambda, \\
\widehat{s}_m = \sum_{\lambda\in\Lambda(m)} \widehat{\beta}_\lambda\varphi_\lambda \qquad &\text{and} \qquad \widehat{\beta}_\lambda = P_n\varphi_\lambda, \\
\chi^2(m) = \left\|s_m - \widehat{s}_m\right\|^2 \quad &= \sum_\lambda \nu_n^2\left(\varphi_\lambda\right), \\
E_m = \mathbb{E}\left[\,\chi^2(m)\,\right] \qquad &\text{and} \qquad \theta_{n,p} = \frac{2n-p}{(n-1)(n-p)}\,.
\end{aligned}$$

Remark: $\chi^2(m)$ is not a true χ^2 statistic, but is only somewhat similar to it.

We also stress two elementary but useful properties. For any $a, b \geq 0$,

$$\begin{aligned}
&(Roo) \qquad \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \\
&(Squ) \qquad 2ab \leq \eta a^2 + \eta^{-1}b^2, \quad \forall \eta > 0.
\end{aligned}$$

Intermediate results The first intermediate result deals with the relationship between $\widehat{R}_p$ and its expectation for each model.

Lemma 5.3. *For any $m \in \mathcal{M}_n$,*

$$\begin{aligned}
L_p(m) - L_p(\widehat{m}) &= \frac{n}{n-p}\left[\,E_m - E_{\widehat{m}}\,\right] - \left(\left\|s - s_{\widehat{m}}\right\|^2 - \left\|s - s_m\right\|^2\right), \\
\widehat{R}_p(m) - L_p(m) &= \theta_{n,p}\,\nu_n(\phi_m) - (1+\theta_{n,p})\left[\,\chi^2(m) - E_m\,\right] - 2(1+\theta_{n,p})\nu_n(s_m).
\end{aligned}$$

In Lemma 5.3, we see that $\nu_n(\phi_m)$ appears in the expressions. The following Proposition enables to upper bound the deviation of this quantity. It is a consequence of Bernstein's inequality [28].

Proposition 5.1. *With the above notations, let $z > 0$ and $C > 0$ be any positive constants and for each m, let us define $y_m = z + C\,nE_m$. Then, we have*

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left[|\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \leq 2e^{-y_m}.$$

Moreover if (Ad) holds, we have

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid |\nu_n(\phi_m)| \geq \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} + \frac{\|\phi_m\|_\infty}{n} y_m \right] \leq \Sigma_1 e^{-z},$$

where Σ_1 is a positive constant independent from n.

We now recall that $\chi^2(m) = \sum_\lambda \nu_n^2(\varphi_\lambda)$. A handy way to study this χ^2-like statistic is to introduce an event of large probability on which we are able to get some control. That is the reason why we introduce the event $\Omega_n(\epsilon)$ for any $\epsilon > 0$.

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.
Another use of Bernstein's inequality provides the following Lemma.

Lemma 5.4. *Set $\epsilon > 0$ and assume that (Reg) , $(Reg2)$ and (Pol) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq 2n^{2+\delta}\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi} (\log n)^2},$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

This Lemma turns out to be useful in order to assess the concentration of $\chi^2(m)$ around its expectation. This result may be found in Massart [28] and is a consequence of Talagrand's inequality.

Proposition 5.2. *Set $\epsilon > 0$ and for any C', $z > 0$, $x_m = z + C'\,nE_m$. Let us assume that (Reg) , $(Reg2)$ and (Pol) are fulfilled. Then,*

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{P}\left[\sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty\, x_m} \right) \right] \leq e^{-x_m}.$$

Furthermore if (Ad) holds,

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\|s\|_\infty\, x_m} \right) \right] \leq \Sigma_2\, e^{-z},$$

where $\Sigma_2 > 0$ denotes a positive constant independent from n.

Finally, in Lemma 5.3, it remains $\nu_n(s_m)$ for which nothing has already been made. The control of this quantity comes from an upper bound, which results from the following lemma.

Lemma 5.5. *Set m, $m' \in \mathcal{M}_n$. Then for any $\rho > 0$,*

$$\sup_{t \in S_m + S_{m'}} \nu_n^2\left(\frac{t}{\|t\|} \right) \leq (1+\rho)\chi^2(m) + (1+\rho^{-1})\chi^2(m').$$

5.3.3 Proof of Theorem 3.1

Proof. (Theorem 3.1)
We are now in position to give the main inequality from which we derive Theorem 3.1.
From the definition of $\widehat{m}$ as $\widehat{m} = \text{Argmin}_{m\in\mathcal{M}_n} \widehat{R}_p(m)$, we deduce that

$$\forall m \in \mathcal{M}_n, \qquad \widehat{R}_p(\widehat{m}) \leq \widehat{R}_p(m),$$

which implies

$$\left[\, \widehat{R}_p(\widehat{m}) - L_p(\widehat{m}) \,\right] \leq \left[\, \widehat{R}_p(m) - L_p(m) \,\right] + \left[\, L_p(m) - L_p(\widehat{m}) \,\right]. \tag{11}$$

Then, we apply Lemma 5.3 to (11) and get

$$\begin{aligned} \|s - s_{\widehat{m}}\|^2 + n\theta_{n,p} E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \quad \leq \quad & \|s - s_m\|^2 + n\theta_{n,p} E_m - (1+\theta_{n,p})\chi^2(m) + \\ & \theta_{n,p}\nu_n\left(\phi_m - \phi_{\widehat{m}}\right) + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right). \end{aligned} \tag{12}$$

Remarks:

- An upper bound for $\nu_n\left(\phi_m - \phi_{\widehat{m}}\right)$ may be obtained through Bernstein's inequality, so that we may relate $\nu_n\left(\phi_{\widehat{m}}\right)$ to $E_{\widehat{m}}$. This is reached thanks to Proposition 5.1.
- Ideally in the oracle inequality we have in mind, the left-hand side of the final inequality is something like $\mathbb{E}\left[\, \|s - \widehat{s}_{\widehat{m}}\|^2 \,\right]$, which is equal to $\mathbb{E}\left[\, \|s - s_{\widehat{m}}\|^2 \,\right] + \mathbb{E}\left[\, \chi^2(\widehat{m}) \,\right]$ with the present notations. However in (12), we observe that the left-hand side is $\mathbb{E}\left[\, \|s - s_{\widehat{m}}\|^2 \,\right] + \mathbb{E}\left[\, E_{\widehat{m}} \,\right]$. In order to relate $\mathbb{E}\left[\, E_{\widehat{m}} \,\right]$ to $\mathbb{E}\left[\, \chi^2(\widehat{m}) \,\right]$, we will uniformly control the discrepancy $E_m - \chi^2(m)$ over $\mathcal{M}_n$ thanks to both Lemma 5.4 and Proposition 5.2.
- Finally, $\nu_n\left(s_{\widehat{m}} - s_m\right)$ may be upper bounded thanks to Lemma 5.5, independently from $E_{\widehat{m}}$ and will therefore be dealt with later.

According to the preceding remarks, we first apply Proposition 5.1 to $\nu_n\left(\phi_m - \phi_{\widehat{m}}\right)$. The successive use of (Reg) , (Squ) with any $\eta > 0$, and (Roo) provides

$$\begin{aligned} \sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} &\leq \sqrt{2V_m \Phi y_m}, \\ &\leq \eta\Phi V_m + \eta^{-1} y_m. \end{aligned}$$

Moreover, note that

$$V_m = \sum_\lambda \mathbb{E}\left[\, \varphi_\lambda^2(X) \,\right] = nE_m + \|s_m\|^2 \leq nE_m + \|s\|^2.$$

Hence with $y_m = z + C\, nE_m$,

$$\sqrt{2V_m \frac{\|\phi_m\|_\infty}{n} y_m} \leq \left[\, \eta\Phi + \eta^{-1} C \,\right] nE_m + \eta\Phi \|s\|^2 + \eta^{-1} z.$$

Similarly, (Reg) entails that

$$\frac{\|\phi_m\|_\infty}{3n} y_m \leq \frac{\Phi C}{3} nE_m + \frac{\Phi}{3} z,$$

which leads us to

$$\begin{aligned} \left|\nu_n\left(\phi_m - \phi_{\widehat{m}}\right)\right| \quad \leq \quad & nE_m \left[\, \eta\Phi + C\eta^{-1} + \Phi\frac{C}{3} \,\right] + nE_{\widehat{m}} \left[\, \eta\Phi + C\eta^{-1} + \Phi\frac{C}{3} \,\right] + 2z \left[\, \frac{\Phi}{3} + \eta^{-1} \,\right] + \\ & 2\eta\Phi \|s\|^2, \end{aligned}$$

except on an event of probability less than $\Sigma_1 e^{-z}$.
Set $\epsilon'' > 0$ and let us choose $\eta = \epsilon''/(3\Phi)$ and $C = 2\epsilon''/\left(\eta^{-1} + \Phi/3\right)$. Then it comes that

$$|\nu_n\left(\phi_m - \phi_{\widehat{m}}\right)| \leq nE_m\epsilon'' + nE_{\widehat{m}}\epsilon'' + 2z\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right] + 2\frac{\epsilon''}{3}\left\|s\right\|^2,$$

Plugging this into (12) provides

$$\left\|s - s_{\widehat{m}}\right\|^2 + n\theta_{n,p}(1-\epsilon'')E_{\widehat{m}} - (1+\theta_{n,p})\chi^2(\widehat{m}) \leq \left\|s - s_m\right\|^2 + n\theta_{n,p}(1+\epsilon'')E_m - (1+\theta_{n,p})\chi^2(m) + 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \theta_{n,p}\left(2z\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right] + 2\frac{\epsilon''}{3}\left\|s\right\|^2\right), \quad (13)$$

except on an event of probability less than $\Sigma_1 e^{-z}$.

On the other hand, Proposition 5.2 implies that for a given $\epsilon > 0$, except on a set of probability less than $\Sigma_2 e^{-z}$, we have

$$\forall m \in \mathcal{M}_n, \quad \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2\left\|s\right\|_\infty x_m}\right).$$

Using $x_m = z + C'nE_m$ and (Roo) , we get

$$\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)\left(\sqrt{nE_m}\left[1 + \sqrt{2\left\|s\right\|_\infty C'}\right] + \sqrt{2\left\|s\right\|_\infty z}\right),$$

which in turn, combined with (Squ) , implies for any $x > 0$

$$\chi^2(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)^2\left((1+x)E_m\left[1 + \sqrt{2\left\|s\right\|_\infty C'}\right]^2 + (1+x^{-1})\frac{2\left\|s\right\|_\infty}{n}z\right). \quad (14)$$

It holds for the particular choices $x = \epsilon$ and $C' = \left(1 - \sqrt{1+\epsilon}\right)^2/\left(2\left\|s\right\|_\infty\right)$, which results in

$$\frac{1-\epsilon''}{(1+\epsilon)^4}\chi^2(\widehat{m})\mathbb{1}_{\Omega_n(\epsilon)} \leq (1-\epsilon'')E_{\widehat{m}} + \frac{1-\epsilon''}{\epsilon(1+\epsilon)}\frac{2\left\|s\right\|_\infty}{n}z.$$

with probability larger than $1 - \Sigma_2 e^{-z}$.
From the above result and (13), it comes that on $\Omega_n(\epsilon)$, with probability larger than $1 - (\Sigma_1 + \Sigma_2)\, e^{-z}$, we have

$$\begin{aligned}\left\|s - s_{\widehat{m}}\right\|^2 + \left(n\theta_{n,p}\frac{1-\epsilon''}{(1+\epsilon)^4} - (1+\theta_{n,p})\right)\chi^2(\widehat{m}) \quad \leq \quad & \left\|s - s_m\right\|^2 + n\theta_{n,p}(1+\epsilon'')E_m - (1+\theta_{n,p})\chi^2(m) + \\ & 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \\ & \theta_{n,p}z\left(\frac{1-\epsilon''}{\epsilon(1+\epsilon)}2\left\|s\right\|_\infty + 2\Phi\left[\frac{1}{3} + \frac{3}{\epsilon''}\right]\right) + \\ & 2\theta_{n,p}\frac{\epsilon''}{3}\left\|s\right\|^2.\end{aligned}$$

Now for any $\epsilon > 0$, we define $\epsilon' > 0$ such that $\sqrt{1-\epsilon'} = (1+\epsilon)^{-4}$ and let us take ϵ'' satisfying $1 - \epsilon'' = \sqrt{1-\epsilon'}$. Then, the above inequality becomes

$$\begin{aligned}\left\|s - s_{\widehat{m}}\right\|^2 + \left[n\theta_{n,p}\left(1-\epsilon'\right) - (1+\theta_{n,p})\right]\chi^2(\widehat{m}) \quad \leq \quad & \left\|s - s_m\right\|^2 + n\theta_{n,p}\left[2 - \sqrt{1-\epsilon'}\right]E_m - (1+\theta_{n,p})\chi^2(m) + \\ & 2(1+\theta_{n,p})\nu_n\left(s_{\widehat{m}} - s_m\right) + \\ & \theta_{n,p}z\left(\frac{\sqrt{1-\epsilon'}}{\epsilon(1+\epsilon)}2\left\|s\right\|_\infty + 2\Phi\left[\frac{1}{3} + \frac{3}{1-\sqrt{1-\epsilon'}}\right]\right) + \\ & 2\theta_{n,p}\frac{1-\sqrt{1-\epsilon'}}{3}\left\|s\right\|^2. \quad (15)\end{aligned}$$

The following point consists in deriving an upper bound for $\nu_n(s_{\widehat{m}} - s_m)$. It results from the following inequalities and Lemma 5.5. Indeed, we have

$$\begin{aligned} 2\nu_n(s_{\widehat{m}} - s_m) &\leq 2\nu_n\left(\frac{s_{\widehat{m}} - s_m}{\|s_{\widehat{m}} - s_m\|}\right)\|s_{\widehat{m}} - s_m\|, \\ &\leq 2\sup_{t \in S_{\widehat{m}} + S_m} \nu_n\left(\frac{t}{\|t\|}\right)\|s_{\widehat{m}} - s_m\|. \end{aligned}$$

Moreover, $\|s_{\widehat{m}} - s_m\| \leq \|s_{\widehat{m}} - s\| + \|s - s_m\|$ and a double use of (Squ) give for any $x > 0$:

$$2\nu_n(s_{\widehat{m}} - s_m) \leq (1+x)\sup_{t \in S_{\widehat{m}} + S_m} \nu_n^2\left(\frac{t}{\|t\|}\right) + \frac{2}{2+x}\|s_{\widehat{m}} - s\|^2 + \frac{2}{x}\|s_m - s\|^2.$$

Finally, Lemma 5.5 yields that for any $\rho > 0$, we have

$$2\nu_n(s_{\widehat{m}} - s_m) \leq (1+x)\left[(1+\rho)\chi^2(\widehat{m}) + (1+\rho^{-1})\chi^2(m)\right] + \frac{2}{2+x}\|s_{\widehat{m}} - s\|^2 + \frac{2}{x}\|s_m - s\|^2.$$

With $x = \epsilon'$ and $\rho = \epsilon'(1+\epsilon')^{-1}$, we get

$$2\nu_n(s_{\widehat{m}} - s_m) \leq (1+2\epsilon')\chi^2(\widehat{m}) + (1+\epsilon')\frac{1+2\epsilon'}{\epsilon'}\chi^2(m) + \frac{2}{2+\epsilon'}\|s_{\widehat{m}} - s\|^2 + \frac{2}{\epsilon'}\|s_m - s\|^2.$$

Plugging this in (15) yields:
On the event $\Omega_n(\epsilon)$, with probability larger than $1 - (\Sigma_1 + \Sigma_2)e^{-z}$, we have for any $m \in \mathcal{M}_n$

$$\begin{aligned} \left[\frac{\epsilon' - 2\theta_{n,p}}{2+\epsilon'}\right]\|s - s_{\widehat{m}}\|^2 + \left[n\theta_{n,p}(1-\epsilon') - 2(1+\theta_{n,p})(1+\epsilon')\right]\chi^2(\widehat{m}) \leq \left[1 + \frac{2}{\epsilon'}(1+\theta_{n,p})\right]\|s - s_m\|^2 + \\ n\theta_{n,p}\left[2 - \sqrt{1-\epsilon'}\right]E_m + \\ \left[\frac{1+2\epsilon'+2\epsilon'^2}{\epsilon'}\right](1+\theta_{n,p})\chi^2(m) + \\ \theta_{n,p}(Az + B), \end{aligned} \tag{16}$$

where $A = \left(\frac{\sqrt{1-\epsilon'}}{\epsilon(1+\epsilon)}2\|s\|_\infty + 2\Phi\left[\frac{1}{3} + \frac{3}{1-\sqrt{1-\epsilon'}}\right]\right)$ and $B = 2\frac{1-\sqrt{1-\epsilon'}}{3}\|s\|^2$.
Then, Lemma 5.2 allows us to take the expectation and get the following result.

$$(\psi_1 \wedge \psi_2)\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)}\|s - \widehat{s}_{\widehat{m}}\|^2\right] \leq (\psi_3 \vee \psi_4)\mathbb{E}\left[\|s - \widehat{s}_m\|^2\right] + \theta_{n,p}\left[A(\Sigma_1 + \Sigma_2) + B\right], \tag{17}$$

where

$$\begin{aligned} \psi_1 &= \frac{\epsilon' - 2\theta_{n,p}}{2+\epsilon'} \\ \psi_2 &= n\theta_{n,p}(1-\epsilon') - 2(1+\theta_{n,p})(1+\epsilon') \\ \psi_3 &= 1 + \frac{2}{\epsilon'}(1+\theta_{n,p}) \\ \psi_4 &= n\theta_{n,p}\left[2 - \sqrt{1-\epsilon'}\right] + (1+\theta_{n,p})\left[\frac{1+2\epsilon'+2\epsilon'^2}{\epsilon'}\right]. \end{aligned}$$

In order to obtain a meaningful inequality, a necessary requirement is $\psi_1, \psi_2, \psi_3, \psi_4 \geq 0$. This is already satisfied for ψ_3 and ψ_4. We have only to check it for both ψ_1 and ψ_2.
It turns out that if $\epsilon' > 2/(n-1)$, then p must satisfy

$$\frac{4\epsilon'}{1+3\epsilon'} + \frac{2}{n}\frac{1+\epsilon'}{1+3\epsilon'} \leq \frac{p}{n} \leq 1 - \frac{2}{\epsilon'(n-1) - 2}, \tag{18}$$

provided

$$\frac{4\epsilon'}{1+3\epsilon'} + \frac{2}{n}\frac{1+\epsilon'}{1+3\epsilon'} \leq 1 - \frac{2}{\epsilon'(n-1)-2},$$

which is established by Lemma 5.6 for $n \geq 29$.
REMARK: In (18) since $0 < \epsilon' \leq 1$ by definition, we have $\frac{4\epsilon'}{1+3\epsilon'} \leq 1$.

Finally to assert the existence if the constant Γ in Theorem 3.1, we need to make sure that the ratio $(\psi_3 \vee \psi_4) / (\psi_1 \wedge \psi_2)$ is bounded.
It may be easily checked that all ψ_ks may be reshaped as

$$\psi_k = \frac{F(p,n)}{1-p/n},$$

where F is a bounded quantity. Moreover by construction, the bounds in (18) lead to $\psi_1 = 0$ and $\psi_2 = 0$, which should be prohibited since we would like to consider the ratio $(\psi_3 \vee \psi_4) / (\psi_1 \vee \psi_2)$. That is the reason why p/n must be slightly larger (resp. lower) than each one of the above bounds, hence (Ran) . Furthermore since no bound depend on s, (Ran) gives the required constant Γ. A similar reasoning shows that it exists a constant $\kappa > 0$ depending on s and the constants of the problem but independent from n, such that

$$\frac{\theta_{n,p}}{\psi_1 \wedge \psi_2} \leq \frac{\kappa}{n},$$

which yields

$$\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] \leq \Gamma(\epsilon,\alpha,\beta) \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\left\|s - \widehat{s}_m\right\|^2\right] + \frac{\kappa(\epsilon,s,\Phi,\alpha,\beta,\delta)}{n}.$$

We now simply add the missing term $\mathbb{E}\left[\mathbb{1}_{\Omega_n(\epsilon)^c} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right]$ to both sides of the above inequality. It only remains to show that this term is of the right order:

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s - \widehat{s}_{\widehat{m}}\right\|^2\right] &\leq \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s - s_{\widehat{m}}\right\|^2\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \left\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\right\|^2\right], \\ &\leq \|s\|^2 \mathbb{P}\left[\Omega_n^c(\epsilon)\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_n^c(\epsilon)} \sum_{\lambda\in\widehat{m}} \left[\nu_n(\varphi_\lambda)^2\right]\right].\end{aligned}$$

Lemma 5.4 then enables to deduce that the first term in the right-hand side inequality satisfies

$$\forall n, \qquad \|s\|^2 \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq \|s\|^2 \frac{n_0}{n},$$

for an appropriate choice of $n_0 > 0$, depending on ϵ, δ and Φ.
For the second one, Jensen's inequality yields

$$\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2(\varphi_\lambda)\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \mathbb{E}\left[\sum_{\lambda\in\widehat{m}} (\varphi_\lambda(X) - P\varphi_\lambda)^2\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right].$$

Moreover, (Squ) with any $\eta > 0$ provides

$$(\varphi_\lambda(X) - P\varphi_\lambda)^2 \leq (1+\eta)\varphi_\lambda^2(X) + (1+\eta^{-1})P\varphi_\lambda^2.$$

Finally, $\sum_{\lambda\in\Lambda(m)} \varphi_\lambda^2 = \phi_m$ and $P\phi_{\widehat{m}} \leq \|\phi_{\widehat{m}}\|_\infty$ lead to

$$\mathbb{E}\left[\sum_{\lambda\in\widehat{m}} \nu_n^2(\varphi_\lambda)\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \left(2+\eta+\eta^{-1}\right) \mathbb{E}\left[\|\phi_{\widehat{m}}\|_\infty\, \mathbb{1}_{\Omega_n^c(\epsilon)}\right] \leq \left(2+\eta+\eta^{-1}\right) \frac{\Phi n}{(\log n)^2} \mathbb{P}\left[\Omega_n^c(\epsilon)\right]$$

thanks to (Reg), and Lemma 5.4 enables to conclude.

□

5.3.4 Proof of Proposition 5.1

Proof. Bernstein's inequality [28] states

$$\forall x > 0, \qquad \mathbb{P}\left[\, |\nu_n(\phi_m)| \geq \frac{1}{n}\sqrt{2v\,x} + \frac{b}{3n}x \,\right] \leq e^{-x},$$

with $b \geq |\phi_m(X_i) - \mathbb{E}\phi_m(X_i)|$ and $v = \sum_{i=1}^n \mathrm{Var}\,[\,\phi_m(X_i)\,]$.
Since X_i are *i.i.d.* and $\phi_m \geq 0$, we have

$$b = \|\phi_m\|_\infty \qquad \text{and} \qquad v \leq n\,V_m\,\|\phi_m\|_\infty\,,$$

hence the first part of the proposition.
For the second part of the result, a union bound provides

$$\begin{aligned}
&\mathbb{P}\left[\, \exists m \in \mathcal{M}_n \mid |\nu_n(\phi_m)| \geq \sqrt{2V_m\frac{\|\phi_m\|_\infty}{n}y_m} + \frac{\|\phi_m\|_\infty}{n}y_m \,\right] \\
&\leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left[|\nu_n(\phi_m)| \geq \sqrt{2V_m\frac{\|\phi_m\|_\infty}{n}y_m} + \frac{\|\phi_m\|_\infty}{n}y_m \right], \\
&\leq \sum_{m\in\mathcal{M}_n} e^{-y_m}, \\
&\leq e^{-z} \sum_{m\in\mathcal{M}_n} e^{-C\,nE_m} \qquad (y_m = z + C\,nE_m), \\
&\leq e^{-z} \sum_{m\in\mathcal{M}_n} e^{-C\xi D_m}, \qquad (Ad) \\
&\leq e^{-z} \sum_{D\geq 1} e^{-C\xi D + \delta\log(D)}, \qquad (Pol), \\
&\leq \Sigma_1\, e^{-z}.
\end{aligned}$$

□

5.3.5 Proof of Lemma 5.4

Proof. We recall that

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n,\ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\}.$$

Then, we deduce that

$$\begin{aligned}
\mathbb{P}\,[\,\Omega_n^c(\epsilon)\,] &= \mathbb{P}\left[\left\{ \exists m \in \mathcal{M}_n,\ \exists \lambda \in \Lambda(m) \mid |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\} \right], \\
&\leq \sum_{m\in\mathcal{M}_n} \sum_{\lambda\in\Lambda(m)} \mathbb{P}\left[\left\{ |\nu_n(\varphi_\lambda)| \geq \frac{2\epsilon\,\|s\|_\infty \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\} \right], \\
&\leq \sum_{m\in\mathcal{M}_n} D_m\, e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad \text{(Bernstein)} \\
&\leq \sum_{D\geq 1} D^{\delta+1} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad (Pol) \\
&\leq n^{\delta+2} e^{-\frac{\|s\|_\infty \eta(\epsilon)}{\Phi}(\log n)^2}, \quad (D \leq n)
\end{aligned}$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

□

5.3.6 Proof of Proposition 5.2

Proof. First, we notice that $\chi(m) = \sqrt{\chi^2(m)}$ may be also expressed as

$$\chi(m) = \sup_{a/\sum_{\lambda\in\Lambda(m)}\alpha_\lambda^2=1} \left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right| \geq \sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right|,$$

where A is dense subset of

$$\left\{a=(a_1,\ldots,a_{D_m})\in\mathbb{R}^{D_m} \mid \sum_{\lambda\in\Lambda(m)}\alpha_\lambda^2 = 1 \text{ and } \sum_{\lambda\in\Lambda(m)}|\alpha_\lambda| \leq \frac{t}{z}\right\}.$$

Moreover, if we define the event

$$\Omega = \left\{\sup_{\lambda\in\Lambda(m)} \nu_n\left(\varphi_\lambda\right) \leq t\right\}$$

for $t>0$, then we deduce that

$$\chi(m) \leq \sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right| \tag{19}$$

on $\Omega\cap\{\chi(m)\geq z\}$.
Then, Talagrand's inequality to $\sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right|$ gives for $\epsilon>0$

$$\forall x>0, \qquad \mathbb{P}\left[\mathbb{1}_\Omega \sup_{a\in A}\left|\nu_n\left(\sum_{\lambda\in\Lambda(m)} a_\lambda\varphi_\lambda\right)\right| \geq (1+\epsilon)\left(\sqrt{\chi^2(m)}+\sqrt{\frac{2\left\|s\right\|_\infty}{n}x}\right)\right] \leq e^{-x},$$

with $z=\sqrt{2\left\|s\right\|_\infty/n}$ and $t = 2\epsilon\left\|s\right\|_\infty\left[\kappa(\epsilon)\Phi n/(\log n)^2\right]^{-1}$.
Finally, the first result comes from both (19) and $\Omega_n(\epsilon)=\Omega$.

As for the second inequality,

$$\begin{aligned}
&\mathbb{P}\left[\exists m\in\mathcal{M}_n \mid \sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m}+\sqrt{2\left\|s\right\|_\infty x_m}\right)\right]\\
&\leq \sum_{m\in\mathcal{M}_n} e^{-x_m},\\
&\leq e^{-z}\sum_{m\in\mathcal{M}_n} e^{-C'nE_m}, \quad (x_m = C'\xi D_m + z)\\
&\leq e^{-z}\sum_{D\geq 1} e^{-C'\xi D+\delta\log D}, \quad (Ad) \text{ and } (Pol)\\
&\leq \Sigma_2\, e^{-z}.
\end{aligned}$$

□

5.3.7 Proof of Lemma 5.6

Lemma 5.6. *For $n\geq 29$, there exists $0<\epsilon<1$ such that*

$$\zeta(\epsilon) > \frac{2}{n-1} \qquad \text{and} \qquad \frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)}+\frac{2}{n} < 1-\frac{2}{\zeta(\epsilon)(n-1)-2},$$

where $\zeta(\epsilon)=\left[1-(1+\epsilon)^{-8}\right]$.

Proof. The first part is obvious since for a given n, we can choose $0 < \epsilon < 1$ such that $\zeta(\epsilon) > 2/(n-1)$. Then with $\delta = \zeta(\epsilon) - 2/(n-1)$, we have

$$\delta(n-1) = \zeta(\epsilon)(n-1) - 2.$$

After some calculations, it is easy to see that

$$\begin{aligned} & \frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2}, \\ \Leftrightarrow \quad & \delta^2 \, \frac{n+6}{n} - \delta \, \frac{n-10}{n} + \frac{2n+10}{(n-1)^2} < 0, \end{aligned}$$

which is a polynomial of degree 2 in δ.
For $n \geq 29$, the discriminant is positive and any δ between the two distinct zeros yields a value for $\zeta(\epsilon)$ such that

$$\frac{4\zeta(\epsilon)}{1+3\zeta(\epsilon)} + \frac{2}{n} < 1 - \frac{2}{\zeta(\epsilon)(n-1)-2},$$

which enables to conclude. □

5.4 Theorem 3.2

5.4.1 Intermediate results

The proof of Theorem 3.2 follows the same structure as that of Theorem 3.1. We subsequently focus on the main differences, which essentially occur in the control of the χ^2-type statistic. Since they are nearly the same, the following results are given (without or) with only short proofs.

We start introducing another event of large probability on which we are able to get the desired control. For any $\epsilon > 0$,

$$\Omega_n(\epsilon) = \left\{ \forall m \in \mathcal{M}_n, \ \forall \lambda \in \Lambda(m), \quad |\nu_n(\varphi_\lambda)| \leq \frac{2\epsilon \|s\|_2 \sqrt{\Phi/\xi \, nE_m} \log n}{\kappa(\epsilon)\sqrt{\Phi n}} \right\},$$

where $\kappa(t) = 2(t^{-1} + 1/3)$.

The following lemma is the counterpart of Lemma 5.4 and is devoted to control the remainder terms. It heavily relies on Bernstein's inequality.

Lemma 5.7. *Set $\epsilon > 0$ and assume that (Reg), (Reg2), (Reg3) (Ad) and (Pol) hold. Then,*

$$\forall \alpha > 0, \qquad \mathbb{P}\left[\Omega_n^c(\epsilon)\right] \leq 2n^{2+\delta} \, e^{-\frac{\eta(\epsilon)}{\sqrt{\Phi}}(\|s\| \vee 1)(\log n)^2},$$

where $\eta(t) = \frac{2\epsilon^2}{\kappa(t)(\kappa(t)+2t/3)}$.

Now, we are in position to give the main result providing the desired control on the χ^2-type statistic.

Proposition 5.3. *Set $\epsilon > 0$ and for any C', $z > 0$, $x_m = z + C'\sqrt{nE_m}$. Assume that (Reg), (Reg2), (Reg3), (Ad) and (Pol) are fulfilled. Then,*

$$\forall m \in \mathcal{M}_n, \qquad \mathbb{P}\left[\sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\| \vee 1)\sqrt{\Phi/\xi \, nE_m} \, x_m} \right) \right] \leq e^{-x_m},$$

and furthermore,

$$\mathbb{P}\left[\exists m \in \mathcal{M}_n \mid \ \sqrt{n}\chi(m) \mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\| \vee 1)\sqrt{\Phi/\xi \, nE_m} \, x_m} \right) \right] \leq \Sigma_2 \, e^{-z},$$

where $\Sigma_2 > 0$ denotes a positive constant independent from n.

Proof. (sketch of proof) It relies on Talagrand's inequality as well as of the following straightforward upper bound.

$$\forall m, \qquad \sup_{t\in S_m,\ \|t\|_2=1} \mathrm{Var}\left[\, t(X)\,\right] \leq \|s\|\,\|t\|_2\sqrt{\|\phi_m\|_\infty} = \|s\|\,\sqrt{\|\phi_m\|_\infty} \leq (\|s\|\vee 1)\sqrt{\|\phi_m\|_\infty}.$$

□

5.4.2 Outline of the proof of Theorem 3.2

The first main difference in the proof comes from the use of Proposition 14, which yields

$$\sqrt{n}\chi(m)\mathbb{1}_{\Omega_n(\epsilon)} \geq (1+\epsilon)\left(\sqrt{nE_m} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m}\right)$$

on an event of high probability.
From several applications of (Squ) and (Roo) , we obtain

$$\begin{aligned}\forall \rho, C'>0, \qquad \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi\, nE_m}\, x_m} &\leq \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi}\,\sqrt{nE_m}\, z} + \sqrt{2(\|s\|\vee 1)\sqrt{\Phi/\xi}C' nE_m}\,,\\ &\leq \rho\,\sqrt{nE_m} + \rho^{-1}(\|s\|\vee 1)\sqrt{\Phi/\xi}\, z + C\,\sqrt{nE_m},\\ &\leq (\rho + C)\,\sqrt{nE_m} + \rho^{-1}(\|s\|\vee 1)\sqrt{\Phi/\xi}\, z,\end{aligned}$$

with $C' = C\left[\,2(\|s\|\vee 1)\sqrt{\Phi/\xi}\,\right]^{-1}$.
Thus in the same way as (14), we derive

$$\forall x>0, \qquad \chi^2(m)\mathbb{1}_{\Omega_n(\epsilon)} \leq (1+\epsilon)^2\left((1+x)E_m\left[\,1+(\rho+C)\,\right]^2 + (1+x^{-1})\frac{\left(\rho^{-1}(\|s\|\vee 1)\sqrt{\Phi/\xi}\right)^2}{n} z^2\right).$$

The following remains essentially the same.

The last point of the proof concerns the addition of the missing term $\mathbb{E}\left[\,\mathbb{1}_{\Omega_n(\epsilon)^c}\,\|s-\widehat{s}_{\widehat{m}}\|^2\,\right]$ to both sides of the inequality. We still have

$$\mathbb{E}\left[\,\mathbb{1}_{\Omega_n^c(\epsilon)}\,\|s-\widehat{s}_{\widehat{m}}\|^2\,\right] \leq \|s\|^2\,\mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] + \mathbb{E}\left[\,\mathbb{1}_{\Omega_n^c(\epsilon)}\,\|s_{\widehat{m}}-\widehat{s}_{\widehat{m}}\|^2\,\right],$$

and Lemma 5.4 gives that

$$\forall n, \qquad \|s\|^2\,\mathbb{P}\left[\,\Omega_n^c(\epsilon)\,\right] + \mathbb{E}\left[\,\textstyle\sum_{\lambda\in\widehat{m}}\nu_n^2\left(\varphi_\lambda\right)\mathbb{1}_{\Omega_n^c(\epsilon)}\,\right] \leq \left(\|s\|^2+1\right)\frac{n_0}{n},$$

for an appropriate choice of $n_0>0$, which depends on δ, Φ and ϵ.
This summarizes the main steps and concludes the proof.

References

[1] Hirotugu Akaike. 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, 1973.

[2] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with unknown variance. *The Annals of Statistics*, 00:00, 2008.

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

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

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

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

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

[8] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 2006.

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

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

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

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

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

[14] P. Burman, E. Chow, and D. Nolan. A Cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

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

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

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

[18] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer, 1993.

[19] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

[20] S. Dudoit and M. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

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

[22] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[23] I. Ibragimov and R. Khas'minskij. *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, Berlin, 1981.

[24] A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 2004.

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

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

[27] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[28] P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.

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

[30] J. Rissanen. Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

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

[32] Marie Sauvé. Histogram selection in non gaussian regression. Technical Report 5911, INRIA, may 2006.

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

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

[35] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

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

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

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

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

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

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