V-fold cross-validation and V-fold penalization in least-squares density estimation

Sylvain Arlot, Matthieu Lerasle

V-FOLD CROSS-VALIDATION AND V-FOLD PENALIZATION IN LEAST-SQUARES DENSITY ESTIMATION

SYLVAIN ARLOT AND MATTHIEU LERASLE

Abstract. This paper studies V-fold cross-validation for model selection in least-squares density estimation. The goal is to provide theoretical grounds for choosing V in order to minimize the least-squares risk of the selected estimator. We first prove a non asymptotic oracle inequality for V-fold cross-validation and its bias-corrected version (V-fold penalization), with an upper bound decreasing as a function of V. In particular, this result implies V-fold penalization is asymptotically optimal. Then, we compute the variance of V-fold cross-validation and related criteria, as well as the variance of key quantities for model selection performances. We show these variances depend on V like $1 + 1/(V-1)$ (at least in some particular cases), suggesting the performances increase much from $V = 2$ to $V = 5$ or 10, and then is almost constant. Overall, this explains the common advice to take $V = 10$—at least in our setting and when the computational power is limited—, as confirmed by some simulation experiments.

1. Introduction

Cross-validation methods are widely used in statistics, for estimating the risk of a given statistical estimator [Sto74, All74, Gei75] and for selecting among a family of estimators. For instance, cross-validation can be used for model selection, where a collection of linear spaces is given (the models) and the problem is to choose the best least-squares estimator over one of these models. We refer to [AC10] for more references about cross-validation for model selection.

Then, a natural question arises: which cross-validation method should be used for minimizing the risk of the selected estimator? For instance, a popular family of cross-validation methods is V-fold cross-validation [Gei75, often called k-fold cross-validation], which depends on an integer parameter V, and enjoys a smaller computational cost than other classical cross-validation methods. The question becomes (1) which V is optimal, and (2) can we do almost as well as the optimal V with a small computational cost, that is, a small V ? Answering the second question is particularly useful for practical applications where the computational power is limited.

Surprisingly, few theoretical results exist for answering these two questions, especially with a non asymptotic point of view [AC10]. In short, previous results in least-squares regression show that at first order, V-fold cross-validation is suboptimal for model selection if V stays bounded, because V-fold cross-validation is biased [Arl08]. When correcting for the bias [Bur89, Arl08], we recover asymptotic optimality whatever V, but without any theoretical result distinguishing among values of V in the non asymptotic second order terms in the risk bounds [Arl08, oracle inequality].

Intuitively, if there is no bias, increasing V should reduce the variance of the V-fold cross-validation estimator of the risk, hence a smaller risk for the final estimator, as confirmed by some simulation experiments [Arl08, for instance]. But variance computations for unbiased V-fold methods have only been made in a very specific regression setting and they are asymptotic [Bur89].

This paper aims at providing theoretical grounds for the choice of V by two means: a non-asymptotic oracle inequality with a second order term depending on V (Section 3) and exact

Key words and phrases. V-fold cross-validation, density estimation, model selection, penalization.

variance computations shedding light on the influence of V on the variance (Section 4). In particular, we would like to understand why the common advice in the literature is to take $V = 5$ or 10, based on simulation experiments [HTF09, for instance].

The results of the paper are proved in the least-squares density estimation framework, because we can then benefit from explicit closed-form formulas and simplifications for the V-fold criteria. In particular, we show V-fold cross-validation and all leave-p-out methods are particular cases of V-fold penalties in least-squares density estimation (Lemma 1).

The first main result of the paper (Theorem 1) is an oracle inequality with leading constant $1 + \varepsilon_n(V)$ with $\varepsilon_n(V) \to 0$ when the sample size n goes to infinity (for unbiased V-fold methods) and $\varepsilon_n(V)$ decreasing as a function of V. To the best of our knowledge, Theorem 1 is the first non asymptotic oracle inequality for V-fold methods enjoying such properties: the leading constant $1+o(1)$ is new in density estimation, and the fact that ε_n decreases with V has never been obtained whatever the framework. Theorem 1 relies on a new concentration inequality for the V-fold penalty (Proposition 4) with deviation terms that decrease when V increases and are sharp, in some cases at least.

The second main result of the paper (Theorem 2) are the first non asymptotic variance computations for V-fold criteria that allow to understand precisely how the model selection performance of V-fold cross-validation or penalization depend on V. Previous results only focused on the variance of the V-fold criteron [Bur89, Cel08, Cel12, CR08], which is not sufficient for our purpose, as explained in Section 4. In our setting, we can then explain theoretically why taking $V > 10$ is not necessary for getting a performance close to the optimum, as confirmed by experiments on synthetic data in Section 5.

2. Least-squares density estimation and definition of V-fold procedures

This first section introduces the framework of the paper, the main procedures studied, and some useful notation.

2.1. General statistical framework.

We observe a sample $\xi_{1:n} = (\xi_1, ..., \xi_n) \in \mathcal{X}^n$ of n independent random variables with common distribution P. We assume P has a density s with respect to some measure μ on $\mathcal{X}$ and $s \in L^2(\mu)$. The goal is to estimate s from $\xi_{1:n}$, that is, to build an estimator $\widehat{s} = \widehat{s}(\xi_{1:n}) \in L^2(\mu)$ such that its quadratic risk $\|s - \widehat{s}\|^2$ is as small as possible, where for any $t \in L^2(\mu)$, $\|t\|$ denotes its $L^2(\mu)$-norm: $\|t\|^2 := \int_{\mathcal{X}} t^2 d\mu$.

Projection estimators are among the most classical estimators in this framework, see for example [DL93, Mas07]. Given a linear subspace S_m of $L^2(\mu)$ (called a model), the projection estimator of s onto S_m is defined by

$$\widehat{s}_m := \operatorname{argmin}_{t \in S_m} \left\{ \|t\|^2 - 2P_n(t) \right\} \quad , \tag{1}$$

where $P_n = n^{-1} \sum_{i=1}^n \delta_{\xi_i}$ is the empirical measure and for any function $t \in L^2(\mu)$, $P_n(t) = \int t dP_n = n^{-1} \sum_{i=1}^n t(\xi_i)$. The quantity minimized in the definition of $\widehat{s}_m$ is often called the empirical risk, and can be denoted by

$$P_n\gamma(t) = \|t\|^2 - 2P_n(t) \qquad \text{where} \quad \forall x \in \mathcal{X}\,, \forall t \in L^2(\mu)\,, \quad \gamma(t;x) = \|t\|^2 - 2t(x) \ .$$

The function γ is called the least-squares contrast.

2.2. Model selection.

When a collection of models $(S_m)_{m \in \mathcal{M}_n}$ is given, the model selection problem [Mas07] consists in choosing from data one among the corresponding projection estimators $(\widehat{s}_m)_{m \in \mathcal{M}_n}$. The goal is to design a model selection procedure $\widehat{m} : \mathcal{X}^n \mapsto \mathcal{M}_n$ so that the final

estimator $\tilde{s} := \widehat{s}_{\widehat{m}}$ has a quadratic risk as small as possible, that is, comparable to the risk of the oracle $\inf_{m \in \mathcal{M}_n} \|\widehat{s}_m - s\|^2$. More precisely, we aim at proving an oracle inequality of the form

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

with large probability. As long as the remainder term R_n is negligible in front of the risk of the oracle, the main goal is to minimize the leading constant C_n, that should be close to 1, for the procedure $\widehat{m}$ to be optimal.

In this paper, we focus on model selection procedures of the form

$$\widehat{m} := \arg\min_{m \in \mathcal{M}_n} \{ \mathrm{crit}(m) \} \ ,$$

where $\mathrm{crit} : \mathcal{M}_n \mapsto \mathbb{R}$ is some data-driven criterion. Since our goal is to satisfy an oracle inequality,

$$\mathrm{crit}_{\mathrm{id}}(m) = \|\widehat{s}_m - s\|^2 - \|s\|^2 = -2P(\widehat{s}_m) + \|\widehat{s}_m\|^2 = P\gamma(\widehat{s}_m) \ .$$

is an ideal criterion.

A popular way of designing a model selection criterion is penalization [BBM99, BM97, BM01, Mas07]:

$$\mathrm{crit}(m) = P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \ ,$$

for some penalty function $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}$, possibly data-driven. From the ideal criterion $\mathrm{crit}_{\mathrm{id}}$, we get the ideal penalty

$$\begin{aligned} \mathrm{pen}_{\mathrm{id}}(m) &:= \mathrm{crit}_{\mathrm{id}}(m) - P_n\gamma(\widehat{s}_m) = (P - P_n)\gamma(\widehat{s}_m) \\ &= 2(P_n - P)(\widehat{s}_m) = 2(P_n - P)(\widehat{s}_m - s_m) + 2(P_n - P)(s_m) \ , \end{aligned}$$

where

$$s_m := \mathrm{argmin}_{t \in S_m} \{ P\gamma(t) \} = \mathrm{argmin}_{t \in S_m} \left\{ \|t - s\|^2 \right\}$$

is the orthogonal projection of s onto $S_m \subset L^1(P)$.

2.3. **V-fold cross validation.** A standard approach for model selection is cross-validation. We refer the reader to [AC10] for references and a complete survey on cross-validation for model selection. This section only provides the minimal definitions and notation necessary for the remainder of the paper.

For any subset $A \subset \{1, \ldots, n\}$, with cardinality a, let

$$P_n^{(A)} := \frac{1}{a} \sum_{i \in A} \delta_{\xi_i} \qquad \widehat{s}_m^{(A)} := \mathrm{argmin}_{t \in S_m} \left\{ \|t\|^2 - 2P_n^{(A)}(t) \right\} \ ,$$

$P_n^{(-A)} = P_n^{(A^c)}$ and $\widehat{s}_m^{(-A)} = \widehat{s}_m^{(A^c)}$, where $A^c = \{1, .., n\} \setminus A$ denotes the complementary of A.

The main idea of cross-validation is data splitting: in order to estimate $\mathrm{crit}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$, some $T \subset \{1, \ldots, n\}$ is chosen, one first trains $\widehat{s}_m(\cdot)$ with $(\xi_i)_{i \in T}$, then test the trained estimator on the remaining data $(\xi_i)_{i \in T^c}$. This provides the hold-out criterion

$$\mathrm{crit}_{\mathrm{HO}}(m, T) := P_n^{(-T)}\gamma\left(\widehat{s}_m^{(T)}\right) = -2P_n^{(-T)}\left(\widehat{s}_m^{(T)}\right) + \left\|\widehat{s}_m^{(T)}\right\|^2 \ , \tag{2}$$

and all cross-validation criteria are defined as averages of hold-out criteria with various subsets T.

This paper focuses on V-fold cross-validation: Let $V \leq n$ be a positive integer and let $\mathcal{B} = (\mathcal{B}_1, ..., \mathcal{B}_V)$ be some partition of $\{1, ..., n\}$. The V-fold cross validation criterion is defined by

$$\mathrm{crit}_{\mathrm{VFCV}}(m, \mathcal{B}) := \frac{1}{V} \sum_{K=1}^{V} \mathrm{crit}_{\mathrm{HO}}(m, \mathcal{B}_K^c) \ .$$

Compared to the hold-out, one expects cross-validation to be less variable thanks to the averaging over V splits of the sample into $(\xi_i)_{i\in\mathcal{B}_K}$ and $(\xi_i)_{i\in\mathcal{B}_K^c}$.

Since $\mathrm{crit}_{\mathrm{VFCV}}(m,\mathcal{B})$ is known to be a biased estimator of $\mathbb{E}\left[\mathrm{crit}_{\mathrm{id}}(m)\right]$, Burman [Bur89] proposed the bias-corrected V-fold cross-validation criterion

$$\mathrm{crit}_{\mathrm{corr,VFCV}}(m,\mathcal{B}) := \mathrm{crit}_{\mathrm{VFCV}}(m,\mathcal{B}) + P_n\gamma\left(\widehat{s}_m\right) - \frac{1}{V}\sum_{K=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-\mathcal{B}_K)}\right) \ .$$

2.4. **Resampling-based and V-fold penalties.** Another approach for building general data-driven model selection criterion is penalization with a resampling-based estimator of the ideal penalty, as proposed by Efron [Efr83] with the bootstrap and recently generalized to all resampling schemes [Arl09]. Let $W \sim \mathcal{W}$ be some random vector of $\mathbb{R}^n$ independent from $\xi_{1:n}$ with $n^{-1}\sum_{i=1}^n W_i = 1$, and denote by $P_n^W = n^{-1}\sum_{i=1}^n W_i\delta_{\xi_i}$ the weighted empirical distribution of the sample. Then, the resampling-based penalty associated with $\mathcal{W}$ is defined as

$$\mathrm{pen}_{\mathcal{W}}(m) := C_{\mathcal{W}}\mathbb{E}_W\left[(P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)\right] \ , \tag{3}$$

where $\widehat{s}_m^W \in \mathrm{argmin}_{t\in S_m}\left\{P_n^W\gamma(t)\right\}$, $\mathbb{E}_W[\cdot]$ denotes the expectation with respect to W only (that is, conditionally to the sample $\xi_{1:n}$), and $C_{\mathcal{W}}$ is some positive constant. Resampling-based penalties have been studied recently in the least-squares density estimation framework [Ler12b], assuming W is exchangeable (i.e., its distribution is invariant by any permutation of its coordinates).

Since computing exactly $\mathrm{pen}_{\mathcal{W}}(m)$ has a large computational cost in general for exchangeable W, some non-exchangeable resampling schemes were introduced in [Arl08], inspired by V-fold cross-validation: given some partition $\mathcal{B} = (\mathcal{B}_1,\dots,\mathcal{B}_V)$ of $\{1,\dots,n\}$, the weight vector W is defined by $W_i = (1-\mathrm{Card}(\mathcal{B}_J)/n)^{-1}\mathbf{1}_{i\notin\mathcal{B}_J}$ for some random variable J with uniform distribution over $\{1,\dots,V\}$. Then, $P_n^W = P_n^{(-\mathcal{B}_J)}$ so that the associated resampling penalty, called *V-fold penalty*, is defined by

$$\begin{aligned}\mathrm{pen}_{\mathrm{VF}}(m,\mathcal{B},C) &:= \frac{C}{V}\sum_{K=1}^{V}\left[\left(P_n - P_n^{(-\mathcal{B}_K)}\right)\gamma\left(\widehat{s}_m^{(-\mathcal{B}_K)}\right)\right] \\ &= \frac{2C}{V}\sum_{K=1}^{V}\left(P_n^{(-\mathcal{B}_K)} - P_n\right)\left(\widehat{s}_m^{(-\mathcal{B}_K)}\right)\end{aligned} \tag{4}$$

where $C > 0$ is left free for flexibility, which is quite useful according to Lemma 1.

2.5. **Links between V-fold penalties, resampling penalties and (corrected) V-fold cross-validation.** In this paper, we focus our study on V-fold penalties because formula (4) covers all V-fold and resampling-based procedures mentioned in Sections 2.3 and 2.4.

First, when $V = n$, the only possible partition is $\mathcal{B}_{\mathrm{LOO}} = (\{1\},\dots,\{n\})$, and the V-fold penalty is called the leave-one-out penalty $\mathrm{pen}_{\mathrm{LOO}}(m,C) := \mathrm{pen}_{\mathrm{VF}}(m,\mathcal{B}_{\mathrm{LOO}},C)$. The associated weight vector W is exchangeable, hence Eq. (4) leads to all exchangeable resampling penalties since they are all equal up to a deterministic multiplicative factor in the least-squares density estimation framework, as proved in [Ler12b].

For V-fold methods, let us assume $\mathcal{B}$ is regular, that is,

$$\textbf{(H5}^\star\textbf{)} \qquad \mathcal{B} \text{ is a partition of } \{1,\dots,n\} \text{ and } \forall K\in\{1,\dots,V\}\ , \quad \mathrm{Card}(\mathcal{B}_K) = \frac{n}{V} \ .$$

Then, we get the following connection between V-fold penalization and cross-validation methods.

Lemma 1. *In least-squares density estimation, under assumption* (**H5***),

$$\mathrm{crit}_{\mathrm{corr,VFCV}}(m,\mathcal{B}) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}\left(m,\mathcal{B},V-1\right) \tag{5}$$

$$\mathrm{crit}_{\mathrm{VFCV}}(m,\mathcal{B}) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}\left(m,\mathcal{B},V-\frac{1}{2}\right) \tag{6}$$

$$\mathrm{crit}_{\mathrm{LPO}}(m,p) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{LPO}}\left(m,p,\frac{n}{p}-\frac{1}{2}\right) \tag{7}$$

$$= P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{LOO}}\left(m,(n-1)\frac{n/p-1/2}{n/p-1}\right) \tag{8}$$

$$= P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}\left(m,\mathcal{B}_{\mathrm{LOO}},(n-1)\frac{n/p-1/2}{n/p-1}\right)$$

where for any $p \in \{1,\dots,n-1\}$*, the leave-*p*-out cross-validation criterion is defined by*

$$\mathrm{crit}_{\mathrm{LPO}}(m,p) := \frac{1}{\mathrm{Card}(\mathcal{E}_p)}\sum_{A\in\mathcal{E}_p} P_n^{(A)}\gamma\left(\widehat{s}_m^{(-A)}\right)$$

$$\text{with} \quad \mathcal{E}_p := \{A \subset \{1,\dots,n\} \ \ s.t. \ \ \mathrm{Card}(A) = p\}$$

and the leave-p*-out penalty is defined by*

$$\forall C>0\,, \quad \mathrm{pen}_{\mathrm{LPO}}(m,p,C) := \frac{C}{\mathrm{Card}(\mathcal{E}_p)}\sum_{A\in\mathcal{E}_p}(P_n - P_n^{(-A)})\gamma\left(\widehat{s}_m^{(-A)}\right) \ .$$

Lemma 1 is proved in Section F.

Remark 1. Eq. (5) was first proved in [Arl08] in a general framework that includes least-squares density estimation, assuming only (**H5***). Eq. (6) shows that V-fold cross-validation and V-fold penalization (with $C = V - 1/2$) yield the same criterion. Similarly, Eq. (7) shows leave-p-out cross-validation and leave-p-out penalization (with $C = n/p - 1/2$) yield the same criterion.
Eq. (8) follows from Lemma 6.11 in [Ler12b] since $\mathrm{pen}_{\mathrm{LPO}}$ belongs to the family of exchangeable resampling penalties, with weights $W_i := (1-p/n)^{-1}\mathbf{1}_{i\notin B}$ and B is randomly chosen uniformly over $\mathcal{E}_p$. It can also be deduced from Proposition 3.1 in [Cel12], see Section F.

As a conclusion of this section, in the least-squares density estimation framework and assuming only (**H5***), it is sufficient to study V-fold penalization with a free multiplicative factor $C \geq V-1$ in front of the penalty for studying also V-fold cross-validation, corrected V-fold cross-validation and exchangeable resampling penalties. Therefore, in Sections 3–6, we focus our study on V-fold methods. Additional results on hold-out (penalization) are given in Section 7.1 for completing the picture.

3. Oracle inequalities

In this section, we state our first main result, that is, a non-asymptotic oracle inequality satisfied by V-fold procedures. The main novelty of this result is that it holds for any $V \in \{1,\dots,n\}$, any constant $C > (V-1)/2$ in front of the penalty, and the leading constant of the oracle inequality is as small as $1+o(1)$ when C is well-chosen. In addition, as proved by Section 2.5, they imply oracle inequalities satisfied by leave-p-out procedures for all p.

Recall that, given a partition $\mathcal{B}$ of $\{1,\dots,n\}$ into V regular blocks, the V-fold estimator is defined by $\tilde{s}(\mathcal{B},C) := \widehat{s}_{\widehat{m}(\mathcal{B},C)}$, where

$$\widehat{m} \in \mathrm{argmin}_{m\in\mathcal{M}_n}\left\{P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}(m,\mathcal{B},C)\right\} \ . \tag{9}$$

3.1. **Main Assumptions.** In order to state the main results, we assume the existence of some constants $L_\star$, $c_{\mathcal{M}}$, $\alpha_{\mathcal{M}}$, c_R^-, c_R^+, $r > 0$ such that

- *For all* $m \in \mathcal{M}_n$,

$$\text{(H1)} \qquad \left\| \sup_{t \in S_m, t \neq 0} \left(\frac{t}{\|t\|} \right)^2 \right\|_\infty \leq L_\star P \left(\sup_{t \in S_m, t \neq 0} \left(\frac{t}{\|t\|} \right)^2 \right) \ .$$

- *The models are nested, i.e.,*

$$\text{(H2)} \qquad \forall m, m' \in \mathcal{M}_n \, , \, S_m \cup S_{m'} \in \{ S_m, S_{m'} \} \ .$$

- *The number of models is polynomial, i.e.,*

$$\text{(H3)} \qquad \forall n \in \mathbb{N}^\star, \ \mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}} \ .$$

- *The oracle risk* $R_n^\star := n \inf_{m \in \mathcal{M}_n} \mathbb{E} \left(\| \widehat{s}_m - s \|^2 \right)$ *satisfies*

$$\text{(H4)} \qquad \forall n \in \mathbb{N}^\star, \ c_R^- (\ln n)^{4+r} \leq R_n^\star \leq c_R^+ n (\ln n)^{-2} \ .$$

- *Pseudo-regularity of the partition* $\mathcal{B} = (\mathcal{B}_K)_{K=1,..,V}$, *i.e.,*

$$\text{(H5)} \qquad \mathcal{B} \text{ is a partition of } \{1, \ldots, n\} \text{ and } \sup_{1 \leq K \leq V} \left| \mathrm{Card}(\mathcal{B}_K) - \frac{n}{V} \right| \leq 1 \ .$$

The assumptions will be discussed in Section 3.3.

3.2. **Oracle inequality for V-fold procedures.** The first main result of the paper is the oracle inequality satisfied by V-fold estimators.

Theorem 1. *Let* $\xi_{1:n}$ *be an i.i.d sample with marginal density* $s \in L^2(\mu)$. *Let* $(S_m)_{m \in \mathcal{M}_n}$ *be a collection of linear spaces satisfying assumptions* (**H1**), (**H2**), (**H3**), (**H4**) *w.r.t. the density* s. *Let* $\mathcal{B}$ *be satisfying assumption* (**H5**), $C > (V-1)/2$ *and*

$$\kappa := \frac{C}{V-1}, \qquad \delta := 2 \left(\kappa - 1 \right), \qquad \varepsilon := \left(1 \vee \frac{\ln n}{V^{1/3}} \right) \frac{\sqrt{\ln n}}{(R_n^\star)^{1/4}} \ .$$

Let $\tilde{s}$ *be the estimator defined by* (9). *A constant* $L > 0$, *depending only on* $L_\star$, $c_{\mathcal{M}}$, $\alpha_{\mathcal{M}}$, c_R^-, c_R^+ *and* r, *exists such that, with probability at least* $1 - n^{-2}$,

$$\text{(10)} \qquad \frac{1 - L\kappa\varepsilon - \delta_-}{1 + L\kappa\varepsilon + \delta_+} \| s - \tilde{s} \|^2 \leq \inf_{m \in \mathcal{M}_n} \left\{ \| s - \widehat{s}_m \|^2 \right\} \ ,$$

where $\forall u \in \mathbb{R}$, $u_+ := \max \{ u, 0 \}$ *and* $u_- := \max \{ -u, 0 \}$.

Theorem 1 is proved in Section G. Let us make a few comments.

- The rate of convergence of the leading constant $\sqrt{\ln n} (R_n^\star)^{-1/4}$ was the one obtained—in an upper bound—for resampling penalties with exchangeable weights in [Ler12b]. Theorem 1 proves that V-fold penalties achieve at least the same rates as soon as $V \geq O\left((\ln n)^3 \right)$. This is interesting, because V-fold penalties are exchangeable resampling penalties only when $V = n$.
- Compared to previous results on V-fold penalization [Arl08], an important feature of Theorem 1 is that all values $V \in \{1, \ldots, n\}$ are allowed. Furthermore, the remainder term $\varepsilon(n, V)$ decreases with V, which gives some theoretical confirmation that increasing the number V of data splits may improve the performance of V-fold methods, as observed empirically. See also Sections 4 and 7.1 on this point.

- Theorem 1 and Lemma 1 imply oracle inequalities satisfied by several estimators based on resampling criterions, according to Section 2.5. In particular, V-fold cross validation corresponds to $C = V - 1/2$, corrected V-fold cross-validation to $C = V - 1$, and the leave-p-out to $V = n$ and $C = (n-1)(n-p/2)/(n-p)$.
- The proof of Theorem 1 mainly relies on a new concentration inequality for V-fold penalties (Proposition 25) that is of independent interest.
- We say that (10) is a non-asymptotic oracle inequality in the sense that all parameters can vary with n as long as (**H1**), (**H2**), (**H3**), (**H4**) hold. However, the constants involved may not be sufficiently small to have a result interesting for small samples.

3.3. **Discussion of the assumptions.** (**H1**), (**H2**) and (**H3**) hold in classical collections of models considered in density estimation, such as regular dyadic histogram spaces, Fourier spaces and regular wavelet spaces, see for example [Ler11]. We refer also to [DL93] for a more complete presentation of these spaces and their approximation properties.
All our results can also be applied to regular histograms provided that $\|s\|_\infty < \infty$. Moreover, (**H2**) can be replaced by one among (**H2′**) and (**H2$^\diamond$**) below:

(**H2′**) $$\|s\|_\infty < \infty, \quad \sup_{(m,m')\in\mathcal{M}_n^2} \sup_{t\in S_m+S_{m'},\|t\|\leq 1} \|t\|_\infty^2 \leq \Gamma n \ .$$

(**H2$^\diamond$**) $(\phi_\lambda)_{\lambda\in\Lambda}$ is an orthonormal basis of $L^2(\mu)$ and $\exists \Lambda_m \subset \Lambda,\ S_m =< (\phi_\lambda)_{\lambda\in\Lambda_m} > \ .$

These two assumptions are considered in [Mas07]. They ensure a general assumption (**H2g**), introduced in Lemma 12 in the proof of Theorem 1 holds. (**H2g**) is sufficient to prove the main theorems.
(**H4**) is a technical assumption. The lower bound means essentially that we are in a non parametric situation. The upper bound roughly means that at least one of the estimators is consistent. Note also that (**H4**) is weaker than the assumption on the bias made in [Arl08]. We refer to [Ler11] for more details on the latter point.
Assumption (**H5***) could be relaxed, at the price of enlarging the bound (10), depending on how far $\mathcal{B}$ is from being regular. Throughout the paper, we choose to focus on (**H5**), or even on (**H5***), to keep the results and their proofs simple. Note also that regular partitions are the most classical ones for V-fold methods.

3.4. **Comparison with previous works.** Few non-asymptotic oracle inequalities have been proved for V-fold penalization or cross-validation procedures.

Concerning cross-validation, previous oracle inequalities are listed in the survey [AC10]. In the least-squares density estimation framework, oracle inequalities were proved by [vdLDK04] in the V-fold case, but compared the risk of the selected estimator with the risk of an oracle trained with $n(V-1)/V$ data. Optimal oracle inequalities were proved by [Cel12] for leave-p-out estimators with $p \ll n$. In comparison, Theorem 1 gives an oracle inequality for any V and considers the strongest possible oracle, that is, trained with n data. Moreover, leave-p-out criterions are studied for $V = n$, $C = (n-1)(n/p - 1/2)/(n/p - 1)$. In that case, we have $C = V - 1 + o(V-1)$. Hence the leading constant in Theorem 1 is asymptotically equal to 1 and we recover the result of [Cel12].

Concerning V-fold penalization, previous results were either valid for $V = n$ only (in least-squares density estimation [Ler12b] and for regressogram estimators [Arl09]), or for V bounded when n tends to infinity (for regressogram estimators [Arl08]). In comparison, Theorem 1 provides a result valid for all $V \in \{1, \dots, n\}$. In particular, the leading constant of the oracle inequality of [Arl08] increases with V, whereas the leading constant in Eq. (10) decreases as V increases, as observed in simulation experiments (for instance, in Section 5 and in [Arl08]).

4. Dependence on V of V-fold penalization and cross-validation

An interesting feature of our oracle inequality is that the remainder term $\varepsilon(n, V)$ improves with V. However, our concentration inequality provide the same deviation rate as in the exchangeable case for $V = O((\ln n)^3)$. In this section we would like to investigate further in this direction and understand more precisely the dependence in V of the V-fold procedures.

In order to do so, let us consider the first step of the proof of Theorem 1: by definition (9), $\tilde{s}(\mathcal{B}, C)$ satisfies, for all $m \in \mathcal{M}_n$,

$$\|s - \tilde{s}\|^2 \leq \|\widehat{s}_m - s\|^2 + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) - \mathrm{pen}_{\mathrm{id}}(m) + \mathrm{pen}_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}_{\mathrm{VF}}(\widehat{m}, \mathcal{B}, C) \ . \tag{11}$$

Then, two quantities play a key role for deriving an oracle inequality from Eq. (11): the expectation and the deviations of the normalized increments

$$\Delta(m, m', \mathcal{B}, C) := \sqrt{n}\left(\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) - \mathrm{pen}_{\mathrm{id}}(m) - (\mathrm{pen}_{\mathrm{VF}}(m', \mathcal{B}, C) - \mathrm{pen}_{\mathrm{id}}(m'))\right)$$

for all $m, m' \in \mathcal{M}_n$, or at least for m, m' that are "close to the oracle $m^\star$ or likely to be selected", since only $m, m' \in \{m^\star, \widehat{m}\}$ truly matter at the end. The influence of the expectations $\mathbb{E}[\Delta(m, m', \mathcal{B}, C)]$ is clearly enlightened by Theorem 1, thanks to the term $\delta(C, V)$ which appears in the leading constant of the oracle inequality (10). In this section, we further investigate the amplitude of deviations of $\Delta(m, m', \mathcal{B}, C)$ by computing their variance as a function of V. The quantity $\Delta(m, m', \mathcal{B}, C)$ is related to relative bounds [Cat07, Section 1.4] which can be used as a tool for model selection [Aud04].

We focus here on $C = V - 1/2$, that is, the V-fold cross-validation criterion, and on $C = V - 1$ which corresponds to its corrected version (see Lemma 1). Results valid for any $C > 0$ are given in Proposition 15 in the appendix. For simplicity, we assume all blocks have the same size n/V, that is, assumption (**H5***) holds true; in particular, V divides n.

Theorem 2. *Let S_m, $S_{m'}$ be two linear subspaces of $L^4(\mu)$, and let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ (resp. $(\psi_\lambda)_{\lambda \in \Lambda_{m'}}$) be some orthonormal basis of S_m (resp. $S_{m'}$) in $L^2(\mu)$. For every $q, r \in \{1, 2\}$, $\lambda, \lambda' \in \Lambda_m \cup \Lambda_{m'}$ and $\Lambda, \Lambda' \in \{\Lambda_m, \Lambda_{m'}\}$, let us define*

$$C_{\lambda,\lambda'}^{(q,r)} := \mathbb{E}\left[(\psi_\lambda(\xi_1) - P(\psi_\lambda))^q (\psi_{\lambda'}(\xi_1) - P(\psi_{\lambda'}))^r\right] \ ,$$

$$v_\lambda := \mathrm{var}\left(\psi_\lambda(\xi_1)\right), \qquad \beta\left(\Lambda, \Lambda'\right) := \sum_{\lambda \in \Lambda, \lambda' \in \Lambda'} \left(C_{\lambda,\lambda'}^{(1,1)}\right)^2 \ ,$$

$$\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) = \beta\left(\Lambda_m, \Lambda_m\right) + \beta\left(\Lambda_{m'}, \Lambda_{m'}\right) - 2\beta\left(\Lambda_m, \Lambda_{m'}\right) \ .$$

*If $\mathcal{B}$ satisfies (**H5***), then,*

$$\mathrm{var}\left(\Delta(m, m', \mathcal{B}, V-1)\right) = 4\,\mathrm{var}_P\left(s_m - s_{m'}\right) + \frac{8V}{V-1}\frac{\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right)}{n} \ . \tag{12}$$

Moreover, for every $\Lambda, \Lambda' \in \{\Lambda_m, \Lambda_{m'}\}$, let us define

$$\gamma\left(\Lambda, \Lambda'\right) := \sum_{\lambda \in \Lambda, \lambda' \in \Lambda'} \left[P(\psi_\lambda) C_{\lambda,\lambda'}^{(1,2)} + P(\psi_{\lambda'}) C_{\lambda,\lambda'}^{(2,1)}\right]$$

$$\mathbf{C}\left(\Lambda_m, \Lambda_{m'}\right) := \gamma\left(\Lambda_m, \Lambda_m\right) + \gamma\left(\Lambda_{m'}, \Lambda_{m'}\right) - 2\gamma\left(\Lambda_m, \Lambda_{m'}\right)$$

$$\zeta\left(\Lambda, \Lambda'\right) := \sum_{\lambda \in \Lambda, \lambda' \in \Lambda'} C_{\lambda,\lambda'}^{(2,2)} - \left(\sum_{\lambda \in \Lambda} v_\lambda\right)\left(\sum_{\lambda \in \Lambda'} v_\lambda\right)$$

$$\textit{and} \quad \mathbf{D}\left(\Lambda_m, \Lambda_{m'}\right) := \zeta\left(\Lambda_m, \Lambda_m\right) + \zeta\left(\Lambda_{m'}, \Lambda_{m'}\right) - 2\zeta\left(\Lambda_m, \Lambda_{m'}\right) \ .$$

Let $\kappa = 1 + [2(V-1)]^{-1}$ *and* $\nu = 1 + \kappa^2 (V-1)^{-1} - [4n(V-1)^2]^{-1}$. *For every* $m, m' \in \mathcal{M}_n$, *if*

$$\begin{aligned}\Delta_{\mathrm{VFCV}}(m,m') &:= \Delta\left(m, m', \mathcal{B}, V - \frac{1}{2}\right) \\ &= \sqrt{n}\left[\mathrm{crit}_{\mathrm{VFCV}}(m) - \mathrm{crit}_{\mathrm{id}}(m) - (\mathrm{crit}_{\mathrm{VFCV}}(m') - \mathrm{crit}_{\mathrm{id}}(m'))\right] \ ,\end{aligned}$$

then,

$$\begin{aligned}\mathrm{var}\left(\Delta_{\mathrm{VFCV}}(m,m')\right) &= 4\,\mathrm{var}_P\left(s_m - s_{m'}\right) + \frac{8\nu}{n}\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) \\ &\quad - \frac{2\mathbf{C}\left(\Lambda_m, \Lambda_{m'}\right)}{(V-1)n} + \frac{\mathbf{D}\left(\Lambda_m, \Lambda_{m'}\right)}{(V-1)^2 n^2} \ .\end{aligned} \tag{13}$$

Finally, for the (corrected) V-fold criterion itself, we have

$$\begin{aligned}\mathrm{var}\left(\mathrm{crit}_{\mathrm{VFCV}}(m;\mathcal{B})\right) &= \frac{4}{n}\mathrm{var}_P\left(s_m\right) + \frac{10V^2}{(V-1)^2 n^2}\left[1 - \frac{6}{5V} + \frac{2}{5V^2} - \frac{1}{5n}\right]\beta\left(\Lambda_m\right) \\ &\quad - \frac{2V}{(V-1)n^2}\gamma\left(\Lambda_m\right) + \frac{V^2}{(V-1)^2 n^3}\zeta\left(\Lambda_m\right)\end{aligned} \tag{14}$$

$$\begin{aligned}\mathrm{var}\left(\mathrm{crit}_{\mathrm{corr,VFCV}}(m;\mathcal{B})\right) &= \frac{4}{n}\mathrm{var}_P\left(s_m\right) + \frac{2}{n^2}\left[\frac{V+3}{V-1} - \frac{1}{n}\right]\beta\left(\Lambda_m\right) \\ &\quad - \frac{2}{n^2}\gamma\left(\Lambda_m\right) + \frac{1}{n^3}\zeta\left(\Lambda_m\right) \ .\end{aligned} \tag{15}$$

Theorem 2 is proved in Section H, where the variance of $\Delta(m, m', \mathcal{B}, C)$ is computed for any $C > 0$, as well as the variances of $\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C)$, $\mathrm{pen}_{\mathrm{id}}(m)$, $\mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C) - \mathrm{pen}_{\mathrm{id}}(m)$ and $P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, C)$ (Proposition 15). By Lemma 1, the variance of the increments of several resampling criterions (V-fold cross-validation and leave-p-out) can then be deduced from Proposition 15.

The key quantities $\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right)$, $\mathbf{C}\left(\Lambda_m, \Lambda_{m'}\right)$ and $\mathbf{D}\left(\Lambda_m, \Lambda_{m'}\right)$ appearing in Theorem 2 do not depend on the choice of particular bases $(\psi_\lambda)_{\lambda\in\Lambda_m}, (\psi_\lambda)_{\lambda\in\Lambda_{m'}}$, see Section I.2 in the supplementary material.

Interpretation of Theorem 2 with regular histogram models. Assuming a particular structure for the models $S_m, S_{m'}$, Eq. (12) and (13) can be simplified, allowing to compare them, and to make their dependence on V clearer.

Let S_m and $S_{m'}$ be the two regular histograms models of respective sizes d_m^{-1} and $d_{m'}^{-1}$. Formally, S_m is defined as the vector space of functions constant on each interval $I_{k,m} := [k/d_m, (k+1)/d_m)$, $k \in \mathbb{Z}$, and $S_{m'}$ is defined similarly. Then, if for any $x \in \mathbb{R}$, $\psi_{k,m}(x) = \sqrt{d_m}\mathbf{1}_{\left\{x \in I_{k,m}\right\}}$, the family $(\psi_{k,m})_{k\in\mathbb{Z}}$, is an orthonormal basis of S_m.

By Proposition 21 in Section I.2, Eq. (13) becomes

$$\begin{aligned}\mathrm{var}\left(\Delta_{\mathrm{VFCV}}(m,m')\right) &= \frac{8}{n} f_{\mathbf{B}}\left(V\right)\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) + 4\left(1 + \delta\left(V,n\right)\right)\mathrm{var}_P\left(s_m - s_{m'}\right) \ ,\end{aligned} \tag{16}$$

$$\text{where} \quad f_{\mathbf{B}}\left(V\right) := 1 + \frac{1}{V-1} + \frac{1}{(V-1)^2} + \frac{1}{4(V-1)^3} - \frac{1}{4n(V-1)^2}$$

$$\text{and} \quad \delta\left(V,n\right) = \frac{2}{(V-1)n} + \frac{1}{(V-1)^2 n^2} = o(1) \ .$$

So, the variance is slightly larger for V-fold cross-validation than for V-fold penalization, but not much larger. If V stays bounded as n tends to infinity, only the first term is multiplied by a bounded factor. If $V \to_{n\to\infty} \infty$, both procedures yield the same variance asymptotically (uniformly over m, m').

Eq. (12) and (16) show that the variance term of V-fold penalization and cross-validation depend on V like

$$\frac{8}{n} f(V) \mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right) + 4 \operatorname{var}_P\left(s_m - s_{m'}\right)$$

for some decreasing function f that depends on the procedure considered. So, in both cases, increasing V decreases the variance of the procedure. In order to understand by which factor the variance decreases when V increases, we have to compare the terms $\frac{\mathbf{B}(\Lambda_m, \Lambda_{m'})}{n}$ and $\operatorname{var}_P\left(s_m - s_{m'}\right)$.

Let us now assume in addition that $S_{m'} \subset S_m$, that is, $d_{m'}$ divides d_m since we consider regular histogram models. This holds for instance with dyadic regular partitions. Then, Remark 4 in Section I.2 shows that $\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right)$ is of the order of $d_{m'}$ (at least when $d_{m'}$ is large enough and $\|s_{m'}\| \geq L \|s\| > 0$ for some constant L). In addition,

$$0 \leq \operatorname{var}_P\left(s_m - s_{m'}\right) \leq \|s\|_\infty \left\|s_m - s_{m'}\right\|^2 \leq \|s\|_\infty \left\|s - s_{m'}\right\|^2$$

and if we assume S_m and $S_{m'}$ are both "close to the oracle", the bias terms $\|s - s_m\|^2 \approx \|s - s_{m'}\|^2$ and the expected variances $n^{-1} d_m \approx n^{-1} d_{m'}$ approximately match.

Overall, these informal arguments suggest that when $S_{m'} \subset S_m$ are both "close to the oracle",

$$L_1 f(V) \frac{d_m}{n} \leq 8 f(V) \frac{\mathbf{B}\left(\Lambda_m, \Lambda_{m'}\right)}{n} + 4 \operatorname{var}_P\left(s_m - s_{m'}\right) \leq L_2 \left(f(V) + L_3\right) \frac{d_m}{n} \tag{17}$$

for some positive constants L_1, L_2, L_3. Since $1 \leq f(V) \leq 4$ whatever V for both cross-validation and penalization, the maximal and minimal values of the variance (obtained with $V = 2$ and $V = n$ respectively) allowed by Eq. (17) only differ by a constant factor. More precisely, for cross-validation $f(2) = 3.25 + o(1)$ and $f(10) \leq 1.124$, and for penalization $f(2) = 2$ and $f(10) = 10/9 \leq 1.12$. So, increasing V from 2 to 10 already puts the variance very close to its minimal value. Increasing V again (say, from 10 to 50) may not improve much the performance, at least in terms of variance.

The conclusion of these informal arguments is confirmed by the simulation study of Section 5, see in particular Section 5.4.

Another interesting feature of this informal argument is that the parameter V appears in the first order term in the variance of the increments $\Delta_{\mathrm{VFCV}}(m, m')$. Most of the existing results focused on the variance of $\operatorname{var}\left(\operatorname{crit}_{\mathrm{VFCV}}(m)\right)$. Burman [Bur89] obtained asymptotic estimates of $\operatorname{var}\left(\operatorname{crit}_{\mathrm{VFCV}}(m)\right)$ in a regression framework. Celisse [Cel08, Cel12] and Celisse and Robin [CR08] computed exactly the variances of $\operatorname{var}\left(\operatorname{crit}_{\mathrm{VFCV}}(m)\right)$ and $\operatorname{var}\left(\operatorname{crit}_{\mathrm{LPO}}(m)\right)$ in the least-squares regression framework with projection estimators. These variances do not show clearly the influence of the parameters V since it only appears in second-order terms, of order $O(n^{-2})$, in $\operatorname{var}\left(\operatorname{crit}_{\mathrm{VFCV}}(m)\right)$. Actually, Eq. (14) shows that, in the histogram case,

$$\begin{aligned}\operatorname{var}\left(\operatorname{crit}_{\mathrm{VFCV}}(m; \mathcal{B})\right) = {} & \left[\frac{1}{n} + \frac{4}{n^2}\left(1 + \frac{1}{V-1} + O\left(\frac{1}{n}\right)\right)\right] \operatorname{var}_P\left(s_m\right) \\ & + \frac{2}{n^2}\left(1 + \frac{1}{V-1}\right)^2 \left[1 + \frac{2}{V} + \frac{1}{V(V-1)} - \frac{1}{n}\right] \beta\left(\Lambda_m\right) \quad .\end{aligned} \tag{18}$$

and Eq. (15) becomes

$$\operatorname{var}\left(\operatorname{crit}_{\mathrm{corr,VFCV}}(m; \mathcal{B})\right) = \frac{1 + O(1/n)}{n} \operatorname{var}_P\left(s_m\right) + \frac{2}{n^2}\left[1 + \frac{4}{V-1} - \frac{1}{n}\right] \beta\left(\Lambda_m\right) \quad . \tag{19}$$

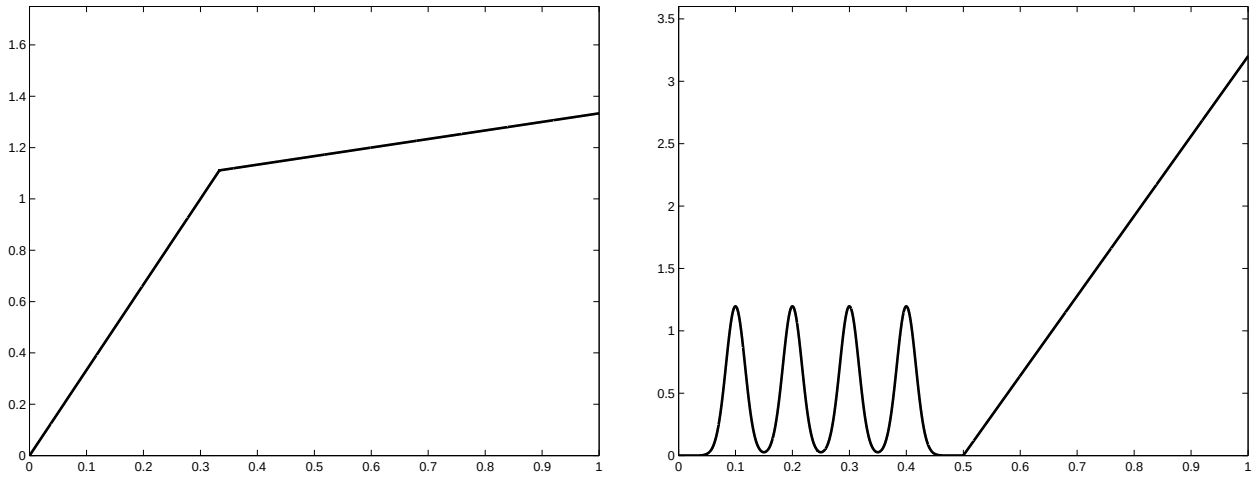

FIGURE 1. The two densities considered. Left: setting L. Right: setting S.

5. Simulation study

This section illustrates the main theoretical results of the paper with some experiments on synthetic data.

5.1. **Setting.** In this section, we consider $\mathcal{X} = [0,1]$ and μ is the Lebesgue measure on $\mathcal{X}$. Two examples are considered for the target density s and for the collection of models $(S_m)_{m\in\mathcal{M}}$.
Two density functions. s are considered, see Figure 1:

- Setting L: $s(x) = \frac{10x}{3}\mathbf{1}_{0\le x<1/3} + (1+\frac{x}{3})\mathbf{1}_{1\ge x\ge 1/3}$.
- Setting S: s is the mixture of the piecewise linear density $s_0(x) = (8x-4)\mathbf{1}_{1\ge x\ge 1/2}$ (with weight 0.8) and four truncated gaussians with means $(k/10)_{k=1,\dots,4}$ and common standard deviation 1/60 (each with weight 0.05).

Two collections of models. are considered, both leading to histogram estimators: for every $m \in \mathcal{M}$, S_m is the set of piecewise constant functions on some partition Λ_m of $\mathcal{X}$.

- "Regu" for regular histograms: $\mathcal{M} = \{1,\dots,n\}$ where for every $m \in \mathcal{M}$, Λ_m is the regular partition of $[0,1]$ into m bins.
- "Dya2" for dyadic regular histograms with two bin sizes and a variable change-point: $\mathcal{M} = \bigcup_{k\in\{1,\dots,\widetilde{n}\}} \{k\} \times \{0,\dots,\lfloor \ln_2(k)\rfloor\} \times \{0,\dots,\lfloor \ln_2(\widetilde{n}-k)\rfloor\}$ where $\widetilde{n} = \lfloor n/\ln(n)\rfloor$ and for every $(k,i,j) \in \mathcal{M}$, $\Lambda_{(k,i,j)}$ is the union of the regular partition of $[0,k/\widetilde{n})$ into 2^i pieces and the regular partition of $[k/\widetilde{n},1]$ into 2^j pieces.

The difference between "Regu" and "Dya2" can be visualized on Figure 2, where the corresponding oracle models have been plotted in setting S. While "Regu" is one of the simplest and most classical collections for density estimation, the flexibility of "Dya2" allows to adapt to the variability of the smoothness of s. Intuitively, in settings L and S, the optimal bin size is smaller on $[0,1/2]$ (where s is varying fastly) than on $[1/2,1]$ (where $|s'|$ is much smaller).

Another point of comparison of Regu and Dya2 is given by Table 1, that reports values of the quadratic risks obtained depending on the collection of models considered. Table 1 shows that in settings L and S, the collection Dya2 helps reducing the quadratic risk by approximately 20% (when comparing the best data-driven procedures of our experiment), and even more when comparing oracle estimators (30% in setting S, 59% in setting L). Therefore, in settings L and S, it is worth considering more complex collections of models (such as Dya2) than regular histograms.

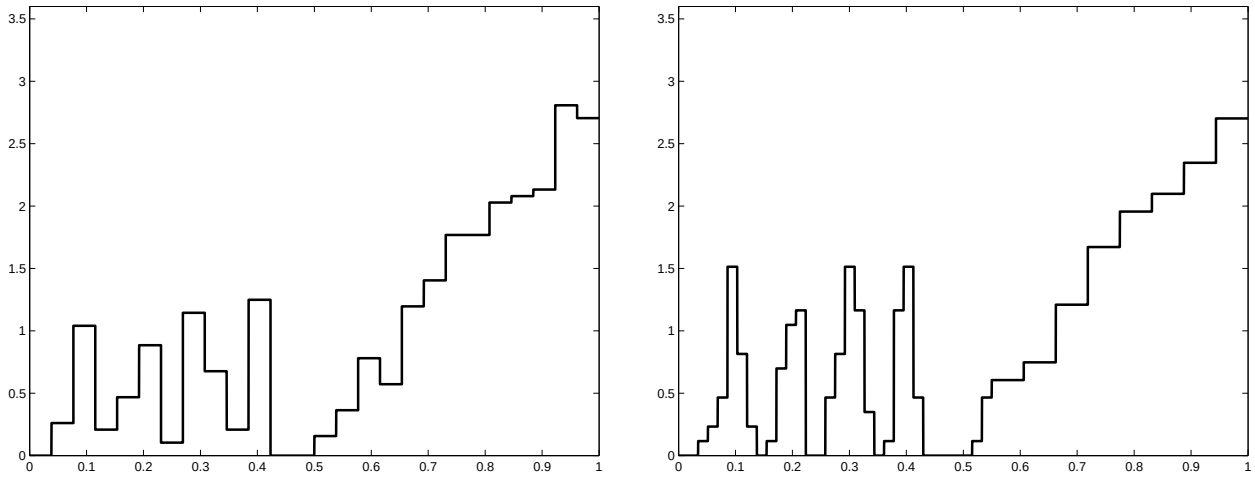

FIGURE 2. Oracle model for one sample, in setting S. Left: Regu. Right: Dya2.

TABLE 1. Comparison of Regu and Dya2: quadratic risks $\mathbb{E}[\|s - \widehat{s}_{\widehat{m}}\|^2]$ of "Oracle" and "Best" estimators (multiplied by 10^3) with the two collections of models. "Best" means that $\widehat{m}$ is the data-driven procedure minimizing $\mathbb{E}[\|s - \widehat{s}_{\widehat{m}}\|^2]$ among the data-driven procedures appearing in Table 3. "Oracle" means that $\widehat{m} \in \arg\min_{m \in \mathcal{M}} \|s - \widehat{s}_m\|^2$ is the oracle model for each sample.

Setting	Oracle(Regu)	Oracle(Dya2)	Best(Regu)	Best(Dya2)
L	13.3 ± 0.2	5.49 ± 0.06	25.5 ± 0.3	19.8 ± 0.3
S	62.7 ± 0.4	43.9 ± 0.3	101.0 ± 0.8	83.7 ± 0.7

Let us finally remark that Dya2 does not reduce the quadratic risk in all settings as significantly as in settings L and S. We performed similar experiments with a few other density functions, sometimes leading to less important differences between Regu and Dya2 in terms of risk (results not shown). The oracle model was always better with Dya2, but in two cases, the risk of the best data-driven procedure with Dya2 was larger than with Regu by 6 to 8%.

5.2. **Procedures compared.** In each setting, we considered the following model selection procedures:

- Mallows' C_p: penalization with $\mathrm{pen}(m) = 2d_m/n$, where $d_m = \mathrm{Card}(\Lambda_m)$ denotes the number of bins.
- V-fold cross-validation with $V \in \{2, 5, 10\}$ and the leave-one-out (that is, V-fold cross-validation with $V = n$), see Section 2.3.
- V-fold penalties (with $C = V - 1$), for $V \in \{2, 5, 10\}$ and the leave-one-out penalty (that is, V-fold penalty $V = n$), see Section 2.4.
- for comparison, the expectation of the ideal penalty $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$.

Since it is often suggested to multiply the usual penalties by some factor larger than one [Arl08], we considered all penalties above multiplied by a factor chosen among $\{1, 1.25, 1.5, 2\}$. Then, in every setting, the risks of the estimators selected with $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ gave us the optimal factor $C^\star$ by which all penalties should be multiplied. In Table 2, we only kept results corresponding to each penalty multiplied by $C^\star$, and we reported the value of $C^\star$. Complete results can be found in the appendix (Table 3).

TABLE 2. Estimated model selection performances, see text. Performances of the best data-driven procedures (that is, the best one and all procedures not significantly worse) are bolded. All penalties are multiplied by the factor $C^\star$, which is chosen according to results obtained with $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$, see Section 5.2.

Procedure	L–Dya2	S–Dya2
C_p	4.38 ± 0.09	3.01 ± 0.04
pen2F	5.12 ± 0.12	2.10 ± 0.02
pen5F	3.80 ± 0.07	1.95 ± 0.02
pen10F	$\mathbf{3.66 \pm 0.06}$	$\mathbf{1.91 \pm 0.02}$
penLOO	$\mathbf{3.61 \pm 0.06}$	$\mathbf{1.91 \pm 0.02}$
2FCV	6.41 ± 0.16	2.10 ± 0.02
5FCV	6.27 ± 0.16	2.09 ± 0.03
10FCV	6.25 ± 0.16	2.07 ± 0.03
LOO	6.41 ± 0.18	2.08 ± 0.03
$C^\star \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	3.66 ± 0.06	1.93 ± 0.02
$C^\star$	2	1.5

5.3. **Model selection performances.** In each setting, all procedures have been compared on $N = 1000$ independent synthetic data sets of size $n = 500$. For measuring their respective model selection performances, we estimated for each procedure $\widehat{m}(\cdot)$

$$C_{\mathrm{or}} := \mathbb{E}\left[\frac{\|s - \widehat{s}_{\widehat{m}}\|^2}{\inf_{m \in \mathcal{M}} \|s - \widehat{s}_m\|^2}\right]$$

which represents the constant that would appear in front of an oracle inequality. The uncertainty of estimation of C_{or} is measured by the empirical standard deviation of $\frac{\|s-\widehat{s}_{\widehat{m}}\|^2}{\inf_{m\in\mathcal{M}}\|s-\widehat{s}_m\|^2}$ divided by $\sqrt{N}$. The results are reported in Table 2 for settings L and S, with the collection Dya2.

Results for Regu are not reported here since C_p is already known to work well with Regu see [Ler12b], so V-fold methods would not improve significantly its performance, with a larger computational cost. Complete results (including Regu) are given in Table 3 in the appendix, showing the performances of C_p and V-fold methods indeed are very close.

Performance as a function of V. Let us first consider V-fold penalization. In both settings L and S, as suggested by our theoretical results, C_{or} decreases when V increases. The improvement is large when V goes from 2 to 5 (26% for L, 7% for S), small but significant when V goes from 5 to 10 (4% for L, 2% for S), and not significant when V goes from 10 to $n = 500$. Since the main influence of V is on the variance of the V-fold penalty, these experiments confirm our interpretation of Theorem 2 in Section 4: increasing V helps much more from 2 to 5 or 10 than from 10 to n.

The picture is less clear for V-fold cross-validation, for which no significant difference is observed among $V \in \{2, 5, 10, n\}$, and C_{or} is minimized for $V \in \{5, 10\}$. Indeed, as explained in a previous work in regression [Arl08], increasing V simultaneously decreases the bias and the variance of the V-fold cross-validation criterion, leading to various possible behaviours of C_{or} as a function of V depending on the setting.

Other comments. Table 2 confirms in the least-squares density estimation framework several facts previouly observed in least-squares regression [Arl08]:

- C_p performs much worse than V-fold penalization (except $V = 2$ in setting L) with the collection Dya2. On the contrary, C_p does well with Regu (see Table 3 in the appendix), but V-fold penalization then performs as well.
- V-fold cross-validation performs significantly worse than V-fold penalization (except in setting S with $V = 2$, where 2-fold cross-validation coincides with 2-fold penalization multiplied by 1.5, as shown in Section 2.4). Nevertheless, making a bad choice for $C^\star$ (which depends on the setting) can lead to worse performance with V-fold penalization, especially when $V = 2$ (see Table 3 in the appendix).

In other settings considered in a preliminary phase of our experiments, differences between $V = 2$ and $V = 5$ were sometimes smaller or not significant, but always with the same ordering (that is, the worse performance for $V = 2$). In a few settings, for which the "change-point" in the smoothness of s was close to the median of $sd\mu$, we found C_p among the best procedures with collection Dya2; then, V-fold penalization and cross-validation always had a performance very close to C_p. Both phenomena lead us to discard all settings for which there were no significant difference to comment.

5.4. **Variance as a function of V.** We now focus on illustrating theoretical results of Section 4 about the variance of V-fold penalization and its influence on model selection. Let us go back to the informal arguments at the beginning of Section 4, in order to understand precisely the role of deviations of $\Delta(m, m', \mathcal{B}, C)$ in the corresponding model selection procedure. For the sake of simplicity, we focus on the unbiased case ($C = V - 1$) in this subsection.

By definition (9) of $\tilde{s}(\mathcal{B}, V-1)$, a model $m \in \mathcal{M}_n$ can be selected if and only if for all $m' \in \mathcal{M}_n$,

$$P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) \leq P_n\gamma\left(\widehat{s}_{m'}\right) + \mathrm{pen}_{\mathrm{VF}}(m', \mathcal{B}, V-1) \ ,$$

which is equivalent to

$$\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m') \leq n^{-1/2}\Delta(m', m, \mathcal{B}, V-1) \ .$$

The left-hand side is of order $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')]$, whereas the right-hand side, which is centered, is at most of order $n^{-1/2}\sqrt{\mathrm{var}\left(\Delta(m', m, \mathcal{B}, V-1)\right)}$. Moreover, when selecting m instead of the oracle, the risk increase is of order $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \inf_{m' \in \mathcal{M}_n} \mathrm{crit}_{\mathrm{id}}(m')]$.

Therefore, the influence of V (through the variance of the criterion) on the model selection performance can be visualized by plotting the difference $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')] - n^{-1/2}\sqrt{\mathrm{var}\left(\Delta(m', m, \mathcal{B}, V-1)\right)}$ for all $m \in \mathcal{M}_n$ and, say, $m' = m^\star \in \mathrm{argmin}_{m \in \mathcal{M}_n} \mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m)]$ the oracle model. When this quantity is negative for some m, it means the corresponding procedure can loose up to $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')]$ in terms of risk. So, the smaller the set of such $m \in \mathcal{M}_n$ is, the better the procedure should be. Such information is provided on Figure 3.

An alternative visualization of the same phenomenon is to determine the set

$$\mathcal{M}^{\mathrm{sel}}(V) := \Big\{ m \in \mathcal{M} \ \text{ s.t. } \ \forall m' \in \mathcal{M}\,, \\ \mathbb{E}\left[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')\right] \leq \sqrt{\mathrm{var}\left(\Delta(m', m, \mathcal{B}, V-1)\right)/n} \Big\} \ , \tag{20}$$

which can be interpretated as "the set of models that could be selected by penalization with $\mathrm{pen}_{\mathrm{VF}}(\cdot, \mathcal{B}, V-1)$", according to the above informal argument. The smaller is this set, the better should be the procedure. Such information is provided on the right of Figure 3.

More precisely, Figure 3 considers the setting S with a sample size $n = 500$ and the collection Regu (for which models are naturally indexed by their dimension). On the left part, $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')] - n^{-1/2}\sqrt{\mathrm{var}\left(\Delta(m', m, \mathcal{B}, V-1)\right)}$ is plotted as a function of the dimension d_m for $V \in \{2, 5, 10, n\}$, as well as $\mathbb{E}[\mathrm{crit}_{\mathrm{id}}(m) - \mathrm{crit}_{\mathrm{id}}(m')]$ (black line). Figure 3 confirms the result of Section 4: the term $\sqrt{\mathrm{var}\left(\Delta(m', m, \mathcal{B}, V-1)\right)}$ decreases when V increases, which can explain the improvement of the model selection procedure performance observed in Table 2. More

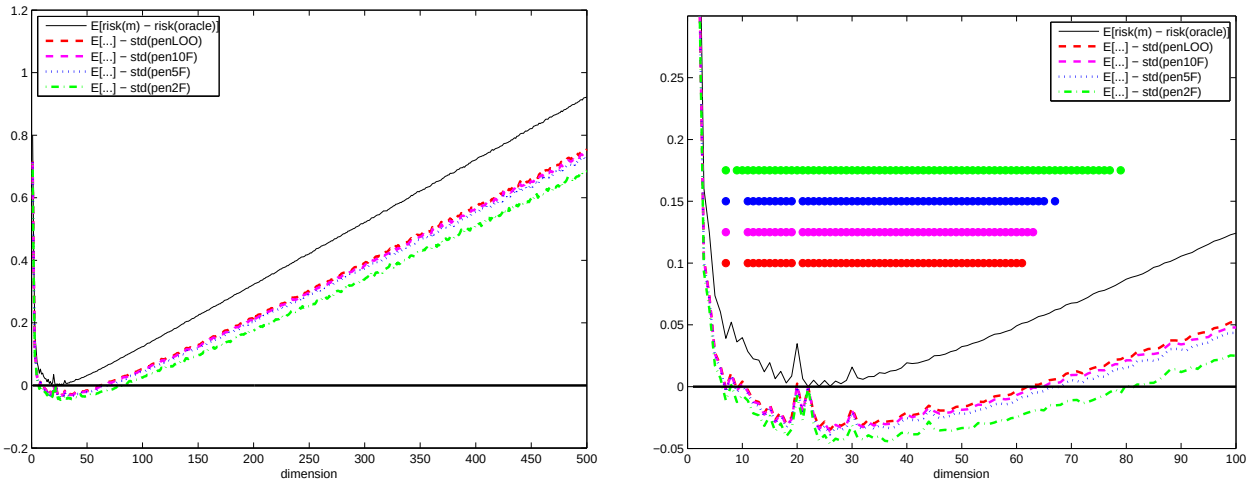

FIGURE 3. Visualization of standard-deviations as a function of V in experiment S–Regu, see text. Left: global picture ($1 \leq m \leq 500$). Right: zoom on the left part; the colored dots represent the set of "selectable models" $\mathcal{M}^{\text{sel}}$ for $V \in \{2, 5, 10, n\}$, with the same color code as the colored lines.

precisely, the standard-deviation term decreases much more from $V = 2$ to $V = 5$ than from $V = 5$ to $V = 10$ or n. Even when zooming on the graph (right of Figure 3), $V = 10$ and $V = n$ are hard to distinguish. On the contrary, going from $V = 2$ to $V \geq 5$ seems to reduce the standard-deviation of $\Delta(m^\star, m, \mathcal{B}, V-1)$ by a factor $\kappa > 1$ for all m, which confirms the informal arguments provided at the end of Section 4: in setting S–Regu with $n = 500$, when $m' = m^\star$, it seems

$$\operatorname{var}\big(\Delta(m', m, \mathcal{B}, V-1)\big) \approx \left(K + \frac{K'}{V-1}\right) \frac{d_m \vee d_{m'}}{n}$$

for some constants $K, K' > 0$.

On the right part of Figure 3, the colored dots represent the set $\mathcal{M}^{\text{sel}}(V)$ for $V = 2$, 5, 10 and n (from top to bottom). As expected from previous results, $\mathcal{M}^{\text{sel}}(V)$ is a decreasing function of V, and the difference between $V = 2$ and $V = 5$ is much larger than between $V = 5$ and $V \geq 10$. Reducing $\mathcal{M}^{\text{sel}}(V)$ then can explain how increasing V makes the excess risk of the selected model smaller, as observed in Table 2: with less "selectable" models, the selected model will be more likely to be close to the oracle (in terms of risk). This phenomenon explains in part why it can be sufficient to take $V = 5$ or $V = 10$ in practice.

6. Fast algorithm for computing V-fold penalties for least-squares density estimation

Since the use of V-fold algorithms is motivated by computational reasons, it is important to discuss the actual computational cost of V-fold penalization and cross-validation as a function of V. In the least-squares density estimation framework, two approaches are possible: a naive one (valid for all frameworks) and a faster one (specific to least-squares density estimation). For clarifying the exposition, we assume in this section (**H5***) holds true (so, V divides n). The general algorithm for computing the V-fold penalized criterion and/or the V-fold cross-validation criterion consists in training the estimator with data-sets $(\xi_i)_{i \notin \mathcal{B}_j}$ for $j = 1, \ldots, V$ and then testing each trained estimator on the data sets $(\xi_i)_{i \in \mathcal{B}_j}$ and $(\xi_i)_{i \notin \mathcal{B}_j}$. In the least-squares density estimation framework, for any model S_m given through an orthogonal family $(\psi_\lambda)_{\lambda \in \Lambda_m}$ of d_m elements of

$L^2(\mu)$, we get the "naive" algorithm described and analysed more precisely in Section I.3.1 in the appendix, whose complexity is of order nVd_m.

Several simplifications occur in the least-squares density estimation framework, that allow to avoid a significant part of the computations made in the naive algorithm. This leads to the following algorithm.

Algorithm 1.

Input: $\mathcal{B}$ some partition of $\{1,...,n\}$ satisfying (**H5***), $\xi_1,\dots,\xi_n \in \mathcal{X}$ and $(\psi_\lambda)_{\lambda\in\Lambda_m}$ a finite orthogonal family of $L^2(\mu)$, with $\mathrm{Card}(m) = d_m$.

(1) For $i \in \{1,\dots,V\}$ and $\lambda \in \Lambda_m$, compute $A_{i,\lambda} := \frac{V}{n}\sum_{j\in B_i}\psi_\lambda(\xi_j)$

(2) For $i,j \in \{1,\dots,V\}$, compute $C_{i,j} := \sum_{\lambda\in\Lambda_m} A_{i,\lambda}A_{j,\lambda}$

(3) Compute $\mathcal{S} = \sum_{1\le i,j\le V} C_{i,j}$ and $\mathcal{T} = \mathrm{tr}(C)$.

Output:

Empirical risk: $P_n\gamma(\widehat{s}_m) = -\mathcal{S}/V^2$

V-fold cross-validation criterion: $\mathrm{crit}_{\mathrm{VFCV}}(m) = \frac{\mathcal{T}}{V(V-1)} - \frac{\mathcal{S}-\mathcal{T}}{(V-1)^2}$,

V-fold penalty: $\mathrm{pen}_{\mathrm{VF}}(m) = (\mathrm{crit}_{\mathrm{VFCV}}(m) - P_n\gamma(\widehat{s}_m))\frac{V-1/2}{V-1}$.

Up to the best of our knowledge, Algorithm 1 is new, even for computing the V-fold cross-validation criterion. Its correctness and complexity are analyzed with the following proposition.

Proposition 2. *Algorithm 1 is correct and has a computational complexity of order* $(n+V^2)d_m$.

In the histogram case, that is, when Λ_m *is a partition of* $\mathcal{X}$ *and* $\forall\lambda\in\Lambda_m$, $\psi_\lambda = |\lambda|^{-1/2}\mathbf{1}_\lambda$, *the computational complexity of Algorithm 1 can be reduced to the order of* $n+V^2d_m$.

Proposition 2 is proved in Section I.3.2 in the supplementary material.

7. Discussion

7.1. Hold-out penalization.

Similarly to all previous results on V-fold methods, we can analyze the hold-out methods where data are only split once.

7.1.1. *Definition.* First, we recall the definition of the hold-out criterion given in (2). Given $T \subset \{1,\dots,n\}$, we train the estimators $\widehat{s}_m^{(T)}$ with the data set $(\xi_i)_{i\in T}$ and estimate its risks with the remaining data set $(\xi_i)_{i\in T^c}$, which gives the hold-out criterion

$$\mathrm{crit}_{\mathrm{HO}}(m,T) = -2P_n^{(-T)}\left(\widehat{s}_m^{(T)}\right) + \left\|\widehat{s}_m^{(T)}\right\|^2 = P_n^{(-T)}\gamma\left(\widehat{s}_m^{(T)}\right) \quad .$$

Similarly, the hold-out penalty is defined as the hold-out estimator of $\mathrm{pen}_{\mathrm{id}}(m)$, that is,

$$\mathrm{pen}_{\mathrm{HO}}(m,T,C) = 2C\left(P_n^{(T)} - P_n^{(-T)}\right)\left(\widehat{s}_m^{(T)} - \widehat{s}_m^{(-T)}\right) \tag{21}$$

and the hold-out penalization estimator is defined by $\tilde{s}_{HO} = \widehat{s}_{\widehat{m}_{\mathrm{HO}}}$, where

$$\widehat{m}_{\mathrm{HO}} = \widehat{m}_{\mathrm{HO}}(T,C) = \mathrm{argmin}_{m\in\mathcal{M}_n}\left\{P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{HO}}(m,T,C)\right\} \quad . \tag{22}$$

7.1.2. *Oracle inequality.* Similarly to Theorem 1, we prove in Section I.1.1 the following oracle inequality for hold-out penalization estimators.

Theorem 3. *Let* $\xi_{1:n}$ *be i.i.d real valued random variables with density* $s \in L^2(\mu)$. *Let* $(S_m)_{m\in\mathcal{M}_n}$ *be a collection of linear spaces satisfying Assumptions* (**H1**), (**H2**), (**H3**), (**H4**). *Let* $C > 0$, $T \subset \{1,\dots,n\}$ *be a training set with* $n_t = \mathrm{Card}(T)$, $n_v = n - n_t$, *let*

$$\kappa^{\mathrm{HO}} = C\frac{n^2}{n_t n_v}, \qquad \delta^{\mathrm{HO}} = 2\left(\kappa^{\mathrm{HO}} - 1\right), \quad \textit{and} \quad \varepsilon_n^{(T)} = \frac{\sqrt{\ln n}}{(n_v \wedge n_t \wedge R_n^\star)^{1/4}} \quad .$$

Let $\tilde{s}_{HO}$ be the hold-out penalization estimator defined by Eq. (22). An absolute constant $L > 0$ exists such that, with probability at least $1 - n^{-2}$,

$$\frac{1 - \delta_{-}^{\mathrm{HO}} - L(\kappa^{\mathrm{HO}} \vee 1)\varepsilon_n^{(T)}}{1 + \delta_{+}^{\mathrm{HO}} + L(\kappa^{\mathrm{HO}} \vee 1)\varepsilon_n^{(T)}} \left\| s - \tilde{s}_{HO} \right\|^2 \leq \inf_{m \in \mathcal{M}_n} \left\| s - \widehat{s}_m \right\|^2 \quad .$$

Let us make a few comments.

- Let V be a divisor of n and let $n_v = n/V$, $n_t = n - n_v$, $C = n_t n_v/(2n^2)$ so that $\kappa^{\mathrm{HO}} = 1$ and $\delta^{\mathrm{HO}} = 0$. Theorems 1 and 3 show the stabilization effect of V-fold procedures. For large V, $\varepsilon_n^{(T)}$ is of order $(n^{-1}V(\ln n)^2)^{1/4}$ whereas ε in Theorem 1 remains of the correct order $(R_n^\star)^{-1/4}\sqrt{\ln n}$.
- When $V = 2$, it is easy to check on formula (21) that the hold-out penalty $\mathrm{pen}_{\mathrm{HO}}^{(-T)}$ built with $\{1, ..., n\} \setminus T$ is exactly the same as $\mathrm{pen}_{\mathrm{HO}}^{(T)}$ built with T. Hence, the 2-fold cross validation penalty pen_{2F} is equal to $(\mathrm{pen}_{\mathrm{HO}}^{(-T)} + \mathrm{pen}_{\mathrm{HO}}^{(T)})/2 = \mathrm{pen}_{\mathrm{HO}}^{(T)}$. This proves the logarithmic loss in the rate $\varepsilon(n, V)$ in Theorem 1 is only due to technical reasons.
- Similarly to Theorem 2, the variance terms can be computed for the hold-out penalty in order to understand separately the roles of the training sample size and of averaging over the V splits, in the V-fold criteria. See Proposition 19 in Section I.1.2 for details.

7.2. **Other model selection procedures for density estimation.** Least-squares density estimation is a classical problem of non-parametric statistics and several model selection procedures have been studied in this framework. Oracle inequalities can be derived, for example, for ℓ_1 penalization methods [BTW07], aggregation procedures [RT07], blockwise Stein method [Rig06] or using T-estimators [Bir08]. Up to our knowledge, none of these methods yield oracle inequalities without remainder terms and with a leading constant asymptotically equal to one at the level of generality presented in this paper. For example, our results are valid for data taking value in any metric space and the models can be of infinite dimension. The results of [Bir08] hold for infinite dimensional models but the estimators are not computable in practice. Let us mention here [BR06] proposed a precise evaluation of the penalty term in the case of regular histogram. Their final penalty is a modification of C_p, performing very well on regular histograms. These performances are likely to become much worse on the collection Dya2 presented in Section 5. This can be seen, for example, in Table 3 where we presented the performances of C_p with different over-penalizing constants.

7.3. **Conclusion on the choice of V.** Overall, choosing V requires a trade-off between:

- Computational complexity, usually proportional to V, sometimes smaller, see Section 6.
- Statistical performance in terms of risk, which is better when the bias and the variance are small. The bias decreases as V increases for V-fold cross-validation, but it can be removed completely or fixed to any desired value by using V-fold penalization instead, see Section 2.4. The variance decreases as V increases, but it almost reaches its minimal value by taking, say, $V = 5$ or $V = 10$, as shown by theoretical and empirical arguments in Sections 4 and 5.

The most common advice for choosing V in the literature [HTF09, for instance, Section 7.10.1] are between $V = 5$ and $V = 10$. This article provides clear evidence why taking V larger does not reduce the variance significantly. Concerning the bias, Lemma 1 shows 5-fold (resp. 10-fold) cross-validation corresponds to overpenalization by a factor $1 + 1/8$ (resp. $1 + 1/18$), which is likely to be a good amount in many cases; in our simulation experiments, the best overpenalization factor was even larger, see also [Arl08].

Note however that our results are only valid for some least-squares algorithms, and it is reported in the literature [AC10] that V-fold cross-validation behaves differently as a function of V in other settings.

Finally, we would like to address the question of choosing between V-fold cross-validation and penalization. The answer is rather simple (at least in least-squares density estimation), since Lemma 1 shows V-fold cross-validation is a particular instance of V-fold penalization, with $C = V - 1/2$. So, if one wants to overpenalize by a factor $(V - 1/2)/(V - 1)$, V-fold cross-validation is definitely the good choice. Otherwise, the best choice would be V-fold penalization with another value for C, depending on how much one wants to overpenalize.

Acknowledgments. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01 (DETECT project).

References

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010.

[All74] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[Arl07] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. `http://tel.archives-ouvertes.fr/tel-00198803/`.

[Arl08] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[Arl09] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624 (electronic), 2009.

[Aud04] Jean-Yves Audibert. A better variance control for pac-bayesian classification. Technical Report 905b, Laboratoire de Probabilités et Modèles Aléatoires, 2004. Available online at `http://imagine.enpc.fr/publications/papers/04PMA-905Bis.pdf`.

[BBLM05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[Bir08] Lucien Birgé. Model selection for density estimation with l^2-loss. *Preprint*, 2008.

[BM97] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BR06] Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Statist.*, 10, 2006.

[BTW07] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 530–543. Springer, Berlin, 2007.

[Bur89] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[Cat07] Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Inst. Math. Statist., 2007.

[Cel08] Alain Celisse. *Model Selection Via Cross-Validation in Density Estimation, Regression and Change-Points Detection*. PhD thesis, University Paris-Sud 11, December 2008. `http://tel.archives-ouvertes.fr/tel-00346320/`.

[Cel12] Alain Celisse. Optimal cross-validation in density estimation. Technical report, arXiv, 2012.

[CR08] Alain Celisse and Stéphane Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Comput. Statist. Data Anal.*, 52(5):2350–2368, 2008.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[Gei75] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[HRB03] Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[Ler11] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011. arXiv:0911.1497.

[Ler12a] Matthieu Lerasle. Adaptive non-asymptotic confidence balls in density estimation. *ESAIM Probab. Statist.*, 16:61–85, 2012.

[Ler12b] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(3):884–908, 2012. arXiv:0910.1654.

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Rig06] Philippe Rigollet. Adaptive density estimation using the blockwise Stein method. *Bernoulli*, 12(2):351–370, 2006.

[RT07] Philippe Rigollet and Alexander B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

[Sto74] Mervyn 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.

[vdLDK04] Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, 3:Art. 4, 27 pp. (electronic), 2004.

Appendix A. Additional notation

Throughout the appendix, $\mathcal{B} = (B_1, \dots, B_V)$ denotes some fixed partition of $\{1, \dots, n\}$ satisfying (**H5***). We refer to Section I.5 in the supplementary material for the general case. For every

$m \in \mathcal{M}_n$, let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of $S_m \in L^2(\mu)$, and

$$U_m := \frac{1}{n^2}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda_m}(\psi_\lambda(\xi_i)-P\psi_\lambda)(\psi_\lambda(\xi_j)-P\psi_\lambda)$$

$$U_{\mathcal{B},m} := \frac{1}{n^2}\sum_{I=1}^{V}\sum_{\substack{i,j\in B_I\\ i\neq j}}\sum_{\lambda\in\Lambda_m}(\psi_\lambda(\xi_i)-P\psi_\lambda)(\psi_\lambda(\xi_j)-P\psi_\lambda) \ .$$

Finally, for all $m, m' \in \mathcal{M}_n$, we define

$$\begin{aligned}
\mathbb{B}_{m,m'} &:= \{\, t \in S_m + S_{m'},\ \|t\| \leq 1 \,\}, \quad \mathbb{B}_m = \mathbb{B}_{m,m} \ , \\
v^2_{m,m'} &:= \sup_{t\in\mathbb{B}_{m,m'}} P[(t-Pt)^2], \quad v_m = v_{m,m} \ , \\
b^2_{m,m'} &:= \sup_{t\in\mathbb{B}_{m,m'}} \|t\|^2_\infty \,, \quad b_m = b_{m,m} \ , \\
D_m &:= P\left(\sup_{t\in\mathbb{B}_m}(t-Pt)^2\right) = n\mathbb{E}\left[\|\widehat{s}_m - s_m\|^2\right] \ , \\
R_m &:= n\|s-s_m\|^2 + D_m = n\mathbb{E}\left[\|s-\widehat{s}_m\|^2\right] \ .
\end{aligned}$$

Appendix B. New results on the V-fold penalty

This section gathers two results on the V-fold penalty that can be of independent interest: an exact formula (Proposition 3) and a concentration inequality (Proposition 4). These two propositions play a key role in the proof of our main results.

B.1. **Exact formula.**

Proposition 3. *Let $V \geq 2$, $n \geq 4$ and $\mathcal{B}$ satisfying* (**H5***). *For every $m \in \mathcal{M}_n$, with the notation of Section A,*

$$\frac{V-1}{2C}\,\mathrm{pen}_{\mathrm{VF}}(m) = \|\widehat{s}_m - s_m\|^2 - \frac{V}{V-1}\,(U_m - U_{\mathcal{B},m}) \ . \tag{23}$$

Proposition 3 is proved in Section D. Note that $2\|\widehat{s}_m - s_m\|^2$ is the main part of the ideal penalty, as explained by Eq. (34). Therefore, Proposition 3 provides an exact formula for $\mathrm{pen}_{\mathrm{VF}}(m) - \mathrm{pen}_{\mathrm{id}}(m)$, which is crucial for proving an oracle inequality like Theorem 1.

B.2. **Concentration inequality.**

Proposition 4. *Let $V \geq 2$, $n \geq 4$, $\mathcal{B}$ satisfying* (**H5***) *and S_m be some model satisfying* (**H1**) *and* (**H4**) *with $R_n^\star$ replaced by $R_m = n\mathbb{E}[\|s-\widehat{s}_m\|^2]$. Let $\mathrm{pen}_{\mathrm{VF}}(m) = \mathrm{pen}_{\mathrm{VF}}(m,\mathcal{B},C)$ be the V-fold penalty on S_m defined by Eq.* (4). *Let $C_\star$ be the constant defined above Eq.* (35). *Let*

$$\varepsilon_1(m,V,x) = C_\star\frac{\sqrt{x}}{R_m^{1/4}}\left[1+\frac{x}{V^{1/3}}+\left(\frac{x^2R_m^{1/4}}{\sqrt{n}}\right)\right] \ .$$

Then, an absolute constant L' exists such that, for every $0 \leq x \leq \frac{\sqrt{R_m}}{C_\star^2} \wedge V^{1/6}R_m^{1/4}$,

$$\mathbb{P}\left(|\mathrm{pen}_{\mathrm{VF}}(m) - \mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m)\right]| > L'\frac{C}{V-1}\varepsilon_1(m,V,x)\frac{R_m}{n}\right) \leq (e^2+6)e^{-x} \ . \tag{24}$$

Furthermore, for some $R_\star$ such that $0 < R_\star \leq R_m$, let

$$\varepsilon_2(n, V, x) = C_\star \frac{\sqrt{x}}{R_\star^{1/4}} \left[1 + \frac{x}{V^{1/3}} + \left(\frac{x^2 R_\star^{1/4}}{\sqrt{n}} \right) \right] \ .$$

Then, an absolute constant $L > 0$ exists such that, for any $2 \leq x \leq \frac{\sqrt{R_\star}}{C_\star^2} \wedge V^{1/6} R_\star^{1/4}$

$$\mathbb{P}\left(\left| \frac{V-1}{C} \operatorname{pen}_{\mathrm{VF}}(m) - 2 \left\| \widehat{s}_m - s_m \right\|^2 \right| > L \varepsilon_2(n, V, x) \frac{R_m}{n} \right) \leq (e^2 + 4) e^{-x} \ . \tag{25}$$

Proposition 4 is proved in Section E.

Eq. (24) is a concentration inequality for the V-fold penalty around its expectation. Eq. (25) is a formulation directly useful for proving an oracle inequality like Theorem 1, since $2 \left\| \widehat{s}_m - s_m \right\|^2$ is the main part of the ideal penalty, as explained by Eq. (34).

Only one concentration inequality was proved for the V-fold penalty before Proposition 4, for regressograms with the least-squares loss [Arl08]. The main novelty of Proposition 4 is to apply to all values of V, and to provide smaller deviation terms when V increases. In [Arl08], it was assumed $V \leq \ln(n)$ and the deviation bounds get worse when V increases, which is highly non-intuitive.

When the variables $\xi_{1:n}$ are real valued, the concentration of the U-statistics $U_m - U_{\mathcal{B},m}$ can be alternatively derived from Theorem 3.1 in [HRB03] and the evaluation of the terms A, B, C, D, F in this result obtained in the proof of Lemma 6.2 in [Ler12a]. We provide in Proposition 4 a result valid in general measurable spaces. This extension is useful for densities on $\mathbb{R}^d$ or when the data $\xi_{1:n}$ are only assumed to be mixing, see for example [Ler11].

The sharpness of the deviation terms in Proposition 4 can be assessed by comparing them to the exact variance computations of Theorem 2 and Proposition 15. This comparison is made in the case of regular histogram models—for which all terms can be simplified-by Proposition 22 in Section I.2.

Appendix C. Elementary properties of least-squares density estimation

This section gathers some classical results on least-squares density estimation that will be used repeatedly in the proofs.

For all $m \in \mathcal{M}_n$ and any non-empty $A \subset \{1, \ldots, n\}$,

$$\widehat{s}_m^{(A)} = \operatorname{argmin}_{t \in S_m} \left\{ P_n^{(A)} \gamma(t) \right\} = \sum_{\lambda \in \Lambda_m} \left(P_n^{(A)} \psi_\lambda \right) \psi_\lambda \ . \tag{26}$$

Classical computations show that

$$s_m = \arg\min_{t \in S_m} \left\{ -2Pt + \|t\|^2 \right\} = \sum_{\lambda \in \Lambda_m} (P\psi_\lambda) \psi_\lambda \ , \tag{27}$$

so that

$$(P_n^{(A)} - P)(\widehat{s}_m^{(A)} - s_m) = \sum_{\lambda \in \Lambda_m} ((P_n^{(A)} - P)\psi_\lambda)^2 = \left\| \widehat{s}_m^{(A)} - s_m \right\|^2 \tag{28}$$

for every $A \subset \{1, \ldots, n\}$, by using Eq. (26).

Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measured space and let $S_\Lambda := \operatorname{span}(\psi_\lambda)_{\lambda \in \Lambda}$ be a linear subspace of measurable functions. Let Π_Λ be the orthogonal projection on $S_\Lambda \cap L_2(\mu)$ w.r.t. the scalar product $\langle t, u \rangle = \int tu d\mu$. Let $\mathbb{B}_\Lambda := \{ t \in S_\Lambda \text{ s.t. } \|t\| \leq 1 \}$.

Lemma 5. *Let f be a function in $L_2(\mu)$. Then*

$$\sup_{t \in \mathbb{B}_\Lambda} \left(\int t f d\mu \right)^2 = \|\Pi_\Lambda(f)\|^2 \ . \tag{29}$$

In particular, for any linear map $L : S_\Lambda \to \mathbb{R}$ such that $\sum_{\lambda\in\Lambda} (L(\psi_\lambda))^2 < \infty$, we have

$$\sum_{\lambda\in\Lambda} (L(\psi_\lambda))^2 = \sup_{t\in\mathbb{B}_\Lambda} (L(t))^2 \quad . \tag{30}$$

Proof of Lemma 5. For every $t \in S_\Lambda$,

$$\int t f d\mu = \langle t,\, f\rangle_{L_2(\mu)} = \langle t,\, \Pi_\Lambda f\rangle_{L_2(\mu)}$$

since Π_Λ is the orthogonal projection on S_Λ. By Cauchy-Schwarz inequality,

$$\forall t \in \mathbb{B}_\Lambda\,, \quad \left|\int t f d\mu\right| \le \|t\|\,\|\Pi_\Lambda f\| \le \|\Pi_\Lambda f\| \quad ,$$

with equality when $\|\Pi_\Lambda f\| = 0$ or when $t = (\|\Pi_\Lambda f\|)^{-1}\Pi_\Lambda f$. This proves (29).

Now, $\sum_{\lambda\in\Lambda} (L(\psi_\lambda))^2$ is the square of the ℓ^2-norm of the sequence $(L(\psi_\lambda))_{\lambda\in\Lambda}$, hence, from (29) and the linearity of L,

$$\sum_{\lambda\in\Lambda} (L(\psi_\lambda))^2 = \sup_{(\beta_\lambda)_{\lambda\in\Lambda}, \sum_{\lambda\in\Lambda}\beta_\lambda^2\le 1} \left(\sum_{\lambda\in\Lambda} \beta_\lambda L(\psi_\lambda)\right)^2 = \sup_{t\in\mathbb{B}_\Lambda} (L(t))^2 \quad .$$

□

Choosing respectively $L(t) = \psi_\lambda(t)$, $L(t) = t(x) - Pt$ and $L(t) = (P_N - P)t$, (29) readily implies the following corollary, which can be found essentially in [Mas07].

Corollary 6. *Assume that for every $x \in \mathcal{X}$, $\sum_{\lambda\in\Lambda} (\psi_\lambda(x))^2 < +\infty$, then $\sum_{\lambda\in\Lambda} (\psi_\lambda(x) - P(\psi_\lambda))^2 < +\infty$ and, $\forall x \in \mathcal{X}$,*

$$\sum_{\lambda\in\Lambda} (\psi_\lambda(x) - P(\psi_\lambda))^2 = \sup_{t\in\mathbb{B}_\Lambda} \left\{ (t(x) - P(t))^2 \right\} \quad , \tag{31}$$

$$\sum_{\lambda\in\Lambda} (\psi_\lambda(x))^2 = \sup_{t\in\mathbb{B}_\Lambda} \left\{ t(x)^2 \right\} \quad . \tag{32}$$

Moreover, $\forall \xi_1, \ldots, \xi_N \in \mathcal{X}$,

$$\sup_{t\in\mathbb{B}_\Lambda} \left\{ ((P_N - P)t)^2 \right\} = \sum_{\lambda\in\Lambda} ((P_N - P)\psi_\lambda)^2 \quad . \tag{33}$$

Proof of Corollary 6. We only have to check that $\sum_{\lambda\in\Lambda} (\psi_\lambda(x) - P(\psi_\lambda))^2 < \infty$, this follows from the fact that $\sum_{\lambda\in\Lambda} (\psi_\lambda(x))^2 < +\infty$ by assumption, and that $\sum_{\lambda\in\Lambda} (P\psi_\lambda)^2 < \infty$ since it is equal to the L^2-norm of $\Pi_\Lambda s$. □

Exact formula for the ideal penalty. Note that

$$\mathrm{pen}_{\mathrm{id}}(m) = 2(P_n - P)(\widehat{s}_m - s_m) + 2(P_n - P)(s_m) = \|\widehat{s}_m - s_m\|^2 + 2(P_n - P)(s_m) \ , \tag{34}$$

using Eq. (28), so the following lemma provides an exact formula for the ideal penalty.

Lemma 7 ([Ler12b])**.** *For all $m \in \mathcal{M}_n$ and $A \subset \{1, \ldots, n\}$,*

$$\left\|\widehat{s}_m^{(A)} - s_m\right\|^2 = \sup_{t\in\mathbb{B}_m} \left\{ \left((P_n^{(A)} - P)t \right)^2 \right\} \quad \textit{and} \quad \|\widehat{s}_m - s_m\|^2 = \frac{1}{n} P_n T_m + U_m \ .$$

Proof of Lemma 7. The first equality comes from Eq. (28) and (33) (Corollary 6). The second equality is proved in [Ler12b, Lemma 6.11]. □

Control of some remainder terms. We will use repeatedly in the proofs that under assumptions (**H1**) and (**H4**),

$$b_m^2 \le L_\star \left(D_m + \|s_m\|^2 \right) \le L_\star \left(1 + \frac{\|s_m\|^2}{c_R^-} \right) R_m \ ,$$

$$v_m^2 \le b_m \|s\| \le \sqrt{L_\star \left(1 + \frac{\|s_m\|^2}{c_R^-} \right)} \|s\| \sqrt{R_m} \ ,$$

which can be rewritten as follows by defining $C_\star := \left(L_\star \left(1 + \frac{\|s\|^2}{c_R^-} \right) \right)^{1/4} \left(1 \vee \sqrt{\|s\|} \right)$:

$$b_m^2 \le C_\star^4 R_m \qquad \text{and} \qquad v_m^2 \le C_\star^2 \sqrt{R_m} \le \frac{C_\star^2}{\sqrt{R_n^\star}} R_m \ . \tag{35}$$

Appendix D. Proof of Proposition 3

A key result in the proof of Proposition 3 is the following lemma about the "covariance" of the weights W_i.

Lemma 8. *Assume that $V \ge 2$ and $n \ge 4$ and that (**H5***) holds. For all $i \in \{1, ..., n\}$, let K_i be the index of the block $\mathcal{B}_K$ such that $i \in B_{K_i}$. For all $i, j \in \{1, \dots, n\}$, we have*

$$E_{i,j}^{(\mathrm{VF})} := \mathbb{E}\left((W_i - 1)(W_j - 1) \right) = \frac{1}{V-1} - \frac{V}{(V-1)^2} \mathbf{1}_{K_i \neq K_j} \ . \tag{36}$$

Lemma 8 is proved in Section I.4.1.

We can now prove Proposition 3. Let us compute $\mathrm{pen}_{\mathrm{VF}}(m)$ on an orthonormal basis $(\psi_\lambda)_{\lambda \in \Lambda_m}$ of S_m. We have, from Eq. (4)

$$\begin{aligned}
\mathrm{pen}_{\mathrm{VF}}(m) &= 2C\mathbb{E}_W \left((P_n^W - P_n) \left(\widehat{s}_m^W - \widehat{s}_m \right) \right) = 2C\mathbb{E}_W \left(\sum_{\lambda \in \Lambda_m} \left[(P_n^W - P_n)\psi_\lambda \right]^2 \right) \\
&= 2C\mathbb{E}_W \left(\sum_{\lambda \in \Lambda_m} \left[(P_n^W - P_n) \left(\psi_\lambda - P\psi_\lambda \right) \right]^2 \right) \\
&= \frac{2C}{n^2} \sum_{\lambda \in \Lambda_m} \sum_{1 \le i,j \le n} \mathbb{E}\left[(W_i - 1)(W_j - 1) \right] \left(\psi_\lambda(X_i) - P\psi_\lambda \right) \left(\psi_\lambda(X_j) - P\psi_\lambda \right) \ . \qquad (37)
\end{aligned}$$

Thanks to Lemma 8, we deduce that

$$\begin{aligned}
\frac{\mathrm{pen}_{\mathrm{VF}}(m)}{2C} &= \frac{1}{n^2(V-1)} \sum_{\lambda \in \Lambda_m} \sum_{1 \le i,j \le n} \left(\psi_\lambda(X_i) - P\psi_\lambda \right) \left(\psi_\lambda(X_j) - P\psi_\lambda \right) \\
&\quad - \frac{V}{(V-1)^2} \frac{1}{n^2} \sum_{\lambda \in \Lambda_m} \sum_{K \neq K'=1}^{V} \sum_{i \in K,\, j \in K'} \left(\psi_\lambda(X_i) - P\psi_\lambda \right) \left(\psi_\lambda(X_j) - P\psi_\lambda \right) \\
&= \frac{1}{V-1} \left\| \widehat{s}_m - s_m \right\|^2 - \frac{V}{(V-1)^2} \left(U_m - U_{\mathcal{B},m} \right) \quad \text{by Eq. (28).}
\end{aligned}$$

□

Appendix E. Proof of Proposition 4

From Proposition 3, it is sufficient to prove concentration inequalities for $U_{\mathcal{B},m}$, U_m and $\|\widehat{s}_m - s\|^2$, which is respectively done by Lemmas 10 and 11 below.

In the proof, we use in particular that $R^\star \leq R_m \leq n$ by (**H4**), at least for n large enough. So, when applying Lemma 11 with $a = n$, we have $\min\{R^\star, n\} = R^\star$. Furthermore, we get $R^\star/(xn) \leq 1/x \leq 1/2$ since we can assume $x \geq 2$ (otherwise, the probability bounds are greater than 1), so this term that appears in ε_5 when applying Lemma 10 can be merged with the constant term in $\varepsilon_2(m, V, x)$. □

First, we prove a general concentration result for U-statistics of the form of $U_{\mathcal{B},m}$ that is valid with no assumption on the partition $\mathcal{B}$.

Lemma 9. *Let S_m be a linear space of function and let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of S_m. Let $(\mathcal{B}_K)_{K=1}^V$ be a partition of $\{1, ..., n\}$ and for all $K = 1, ..., V$, let $n_K = \mathrm{Card}(\mathcal{B}_K)$. Let $Z = n^2 U_{\mathcal{B},m}$. Then, an absolute constant $C > 0$ exists such that, for all $x > 0$, for all $\eta \in (0, 1/\sqrt{x}]$, we have*

$$P\left(|Z| \leq C\left(\eta D_m\sqrt{\sum_{K=1}^V n_K^2} + \frac{v_m^2 x^2}{\eta}\sqrt{\sum_{K=1}^V n_K^2} + \frac{\sqrt{V} b_m^2 x^4}{\eta^3}\right)\right) \geq 1 - e^{2-x}.$$

In particular, if $\mathcal{R}_m > 0$ and $\nu > 0$ satisfy

$$D_m \leq \mathcal{R}_m,\ v_m^2 \leq \nu^2 \mathcal{R}_m,\ b_m^2 \leq \mathcal{R}_m, \tag{38}$$

taking $\eta = \epsilon x$ in the previous inequality yields, for all x such that $\epsilon\sqrt{x} \leq 1$,

$$P\left(|U_{\mathcal{B},m}| \leq C x \frac{\mathcal{R}_m}{n}\left(\left(\epsilon + \frac{\nu^2}{\epsilon}\right)\frac{2}{n}\sqrt{\sum_{K=1}^V n_K^2} + \frac{\sqrt{V}}{n\epsilon^3}\right)\right) \geq 1 - e^{2-x}.$$

Proof of Lemma 9. For all $K = 1, ..., V$, let

$$Z_K = \sum_{\lambda\in\Lambda_m}\sum_{i\neq j\in\mathcal{B}_K} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

As the random variables Z_K are independent, we can apply [BBLM05, Theorem 2] (see Lemma 33) to get that

$$\|Z\|_q \leq 2\sqrt{c}\sqrt{q\sum_{K=1}^V \|Z_K\|_{q/2}^2}.$$

From [Ler11, Corollary 4.3 in the supplementary material] (recalled in Corollary 29), we have, with probability larger than $1 - 4e^{-x}$,

$$|Z_K| \leq C'\left(\epsilon n_K D_m + \frac{n_K}{\epsilon} v_m^2 x + \frac{b_m^2 x^2}{\epsilon^3}\right) .$$

Integrating over x (see Lemma 31 for detailed computations), we deduce an absolute constant $C > 0$ exists such that

$$\|Z_K\|_{q/2} \leq C\left(\epsilon n_K D_m + \frac{n_K}{\epsilon} v_m^2 q + \frac{b_m^2 q^2}{\epsilon^3}\right)$$

Hence, there exists an absolute constant C such that

$$\|Z\|_q \leq C\left(\epsilon D_m\sqrt{q\sum_{K=1}^V n_K^2} + \frac{v_m^2 q^{3/2}}{\epsilon}\sqrt{\sum_{K=1}^V n_K^2} + \frac{\sqrt{V} b_m^2 q^{5/2}}{\epsilon^3}\right).$$

Using [Arl07, Lemma 8.10] (recalled in Lemma 30), we obtain that there exists an absolute constant C such that, with probability $1 - e^{2-x}$,

$$|Z| \leq C \left(\epsilon D_m \sqrt{x \sum_{K=1}^{V} n_K^2} + \frac{v_m^2 x^{3/2}}{\epsilon} \sqrt{\sum_{K=1}^{V} n_K^2} + \frac{\sqrt{V} b_m^2 x^{5/2}}{\epsilon^3} \right).$$

We conclude the proof, taking $\epsilon = \eta/\sqrt{x}$. □

In particular, for regular partitions $\mathcal{B}$, Lemma 9 implies the following.

Lemma 10. *Let $\xi_{1:n}$ be i.i.d random variables and let S_m be a linear space satisfying Assumptions* (**H1**), (**H4**). *For all $m \in \mathcal{M}_n$, let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ be an orthonormal basis of S_m. Let $(\mathcal{B}_K)_{K=1,\dots,V}$ be a partition of $\{1, ..., n\}$ satisfying* (**H5***). *Let $C_\star$ be the constant defined in Eq.* (35). *Let $R_\star$ be such that $0 < R_\star \leq R_m$ and*

$$\varepsilon_3(n, V, x) = \frac{x}{V^{1/3}} \vee \left\{ \frac{x^2 R_\star^{1/4}}{\sqrt{n}} \right\} \vee \frac{R_\star}{xn}.$$

Then, an absolute constant $L > 0$ exists such that, for all $2 \leq x \leq C_\star^{-2} R_\star^{1/2} \wedge V^{1/6} R_\star^{1/4}$,

$$\mathbb{P}\left(|U_{\mathcal{B},m}| > L \varepsilon_3(n, V, x) \frac{C_\star x^{1/2}}{R_\star^{1/4}} \frac{R_m}{n} \right) \leq e^{2-x} \ .$$

Proof of Lemma 10. From (**H5***), we have $n_K = n/V$, hence

$$\sum_{K=1}^{V} n_K^2 = \frac{n^2}{V} \qquad \text{thus} \qquad \frac{1}{n} \sqrt{\sum_{K=1}^{V} n_K^2} = \sqrt{\frac{1}{V}} \ .$$

Condition (38) of Lemma 9 holds with $\nu = \sqrt{x} R_\star^{-1/4}$, $\mathcal{R}_m = C_\star R_m$ from (35). We deduce from this lemma that there exists an absolute constant L such that, for all $\epsilon > 0$ satisfying $\epsilon\sqrt{x} \leq 1$,

$$P\left(|U_{\mathcal{B},m}| \leq C_\star L x \frac{R_m}{n} \left(\left(\epsilon + \frac{\nu^2}{\epsilon} \right) \frac{1}{\sqrt{V}} + \frac{\sqrt{V}}{n\epsilon^3} \right) \right) \geq 1 - e^{2-x}.$$

Let $\epsilon = V^{1/6}\nu \wedge x^{-1/2}$, we have

$$\frac{\nu^2}{\epsilon\sqrt{V}} \vee \frac{\epsilon}{\sqrt{V}} \leq \frac{\nu}{V^{1/3}}, \qquad \frac{\sqrt{V}}{n\epsilon^3} \leq \frac{1}{n\nu^3} \vee \left\{ \frac{x R_\star^{1/4}}{\sqrt{n}} \nu \right\} \leq \left(\frac{1}{x^2} \frac{R_\star}{n} \vee \left\{ \frac{x R_\star^{1/4}}{\sqrt{n}} \right\} \right) \nu \ .$$

□

Lemma 11. *Let $\xi_{1:n}$ be i.i.d random variables and let S_m be a linear space satisfying Assumptions* (**H1**), (**H4**). *Let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ be an orthonormal system in S_m and let $R_\star$ be such that $0 < R_\star \leq R_m$. Let $C_\star$ be the constant defined in* (35). *There exists an absolute constant L such that,*

$$\forall 2 \leq x \leq \frac{\sqrt{R_\star}}{C_\star^2}, \qquad \mathbb{P}\left(|U_m| > L \frac{C_\star \sqrt{x}}{R_\star^{1/4}} \frac{R_m}{n} \right) \leq 4e^{-x} \ ,$$

and for every $A \subset \{1, \dots, n\}$ with cardinality $a > 0$,

$$\forall 2 \leq x \leq \frac{\sqrt{R_\star \wedge a}}{C_\star^2}, \qquad \mathbb{P}\left(\left| \left\| \widehat{s}_m^{(A)} - s_m \right\|^2 - \frac{D_m}{a} \right| > L \frac{C_\star \sqrt{x}}{(R_\star \wedge a)^{1/4}} \frac{R_m}{a} \right) \leq 2e^{-x} \ .$$

Proof of Lemma 11. We combine Lemma 7 and Eq. (35) with two results from [Ler11, supplementary material]: Theorem 4.1 (recalled in Proposition 28) and Corollary 4.3 (recalled in Corollary 29). □

Appendix F. Proof of Lemma 1

Let us first recall here the proof of Eq. (5) (coming from [Arl08]) for the sake of completeness. By (**H5***), $P_n - P_n^{(-\mathcal{B}_K)} = V^{-1}(P_n^{(\mathcal{B}_K)} - P_n^{(-\mathcal{B}_K)})$ and $P_n^{(\mathcal{B}_K)} - P_n = (V-1)V^{-1}(P_n^{(\mathcal{B}_K)} - P_n^{(-\mathcal{B}_K)})$, so that

$$\begin{aligned}
\mathrm{crit}_{\mathrm{pen}_{\mathrm{VF}}}(m, \mathcal{B}, V-1) &:= P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}(m, \mathcal{B}, V-1) \\
&= P_n\gamma\left(\widehat{s}_m\right) + \frac{V-1}{V^2}\sum_{K=1}^{V}\left[\left(P_n^{(\mathcal{B}_K)} - P_n^{(-\mathcal{B}_K)}\right)\gamma\left(\widehat{s}_m^{(-\mathcal{B}_K)}\right)\right] \\
&= P_n\gamma\left(\widehat{s}_m\right) + \frac{1}{V}\sum_{K=1}^{V}\left[\left(P_n^{(\mathcal{B}_K)} - P_n\right)\gamma\left(\widehat{s}_m^{(-\mathcal{B}_K)}\right)\right] \\
&= \mathrm{crit}_{\mathrm{corr,VFCV}}(m, \mathcal{B}) \ .
\end{aligned}$$

We can prove Eq. (6) and (7) simultaneously, by proving a slightly more general result, namely Eq. (42) below. Let $\mathcal{E}$ be a set of subsets of $\{1, \ldots, n\}$ such that

$$\forall A \in \mathcal{E}\,, \quad \mathrm{Card}(A) = p \quad \text{and} \quad \frac{1}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}} P_n^{(-A)} = P_n \ . \tag{39}$$

Let us consider the associated penalty

$$\mathrm{pen}_{\mathcal{E}}(m, C) = \frac{C}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}}(P_n - P_n^{(-A)})\gamma\left(\widehat{s}_m^{(-A)}\right) = \frac{2C}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}}(P_n^{(-A)} - P_n)\left(\widehat{s}_m^{(-A)}\right)$$

and the associated cross-validation criterion

$$\mathrm{crit}_{\mathcal{E}}(m) = \frac{1}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}} P_n^{(A)}\gamma\left(\widehat{s}_m^{(-A)}\right) \ .$$

When $\mathcal{E} = \{\mathcal{B}_1, \ldots, \mathcal{B}_V\}$, we get the V-fold penalty $\mathrm{pen}_{\mathrm{VF}} = \mathrm{pen}_{\mathcal{E}}$ and the V-fold cross-validation criterion $\mathrm{crit}_{\mathrm{VFCV}} = \mathrm{crit}_{\mathcal{E}}$, and Eq. (39) holds true with $p = n/V$ under assumption (**H5***). When $\mathcal{E} = \mathcal{E}_p := \{A \subset \{1, \ldots, n\} \text{ s.t. } \mathrm{Card}(A) = p\}$, Eq. (39) always holds true and we get the leave-p-out penalty $\mathrm{pen}_{\mathrm{LPO}} = \mathrm{pen}_{\mathcal{E}}$ and the leave-p-out cross-validation criterion $\mathrm{crit}_{\mathrm{LPO}} = \mathrm{crit}_{\mathcal{E}}$.

Let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be some orthogonal basis of S_m in $L^2(\mu)$. On the one hand, using Eq. (39) and (26), we get

$$\begin{aligned}
\mathrm{pen}_{\mathcal{E}}(m, C) &= \frac{2C}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}}(P_n^{(-A)} - P_n)\left(\widehat{s}_m^{(-A)}\right) \\
&= \frac{2C}{\mathrm{Card}(\mathcal{E})}\sum_{A\in\mathcal{E}}\sum_{\lambda\in\Lambda_m}\left[\left(P_n^{(-A)}(\psi_\lambda) - P_n(\psi_\lambda)\right)P_n^{(-A)}(\psi_\lambda)\right] \\
&= \frac{2C}{\mathrm{Card}(\mathcal{E})}\sum_{\lambda\in\Lambda_m}\left[\sum_{A\in\mathcal{E}}\left(P_n^{(-A)}(\psi_\lambda)\right)^2 - P_n(\psi_\lambda)\sum_{A\in\mathcal{E}}P_n^{(-A)}(\psi_\lambda)\right] \\
&= \frac{2C}{\mathrm{Card}(\mathcal{E})}\sum_{\lambda\in\Lambda_m}\sum_{A\in\mathcal{E}}\left[\left(P_n^{(-A)}(\psi_\lambda)\right)^2 - \left(P_n(\psi_\lambda)\right)^2\right] \ . \qquad (40)
\end{aligned}$$

On the other hand, using that $P_n^{(A)} = \frac{n}{p} P_n - \frac{n-p}{p} P_n^{(-A)}$ by (39),

$$
\begin{aligned}
&\mathrm{crit}_{\mathcal{E}}(m) - P_n\gamma\left(\widehat{s}_m\right) \\
&= \frac{1}{\mathrm{Card}(\mathcal{E})} \sum_{A\in\mathcal{E}} \left[P_n^{(A)}\gamma\left(\widehat{s}_m^{(-A)}\right) - P_n\gamma\left(\widehat{s}_m\right) \right] \\
&= \frac{1}{\mathrm{Card}(\mathcal{E})} \sum_{A\in\mathcal{E}} \left[\left\|\widehat{s}_m^{(-A)}\right\|^2 - 2P_n^{(A)}\left(\widehat{s}_m^{(-A)}\right) - \|\widehat{s}_m\|^2 + 2P_n\left(\widehat{s}_m\right) \right] \\
&= \frac{1}{\mathrm{Card}(\mathcal{E})} \sum_{A\in\mathcal{E}} \sum_{\lambda\in\Lambda_m} \left[\left(P_n^{(-A)}(\psi_\lambda)\right)^2 - 2P_n^{(A)}(\psi_\lambda)P_n^{(-A)}(\psi_\lambda) + \left(P_n(\psi_\lambda)\right)^2 \right] \\
&= \frac{1}{\mathrm{Card}(\mathcal{E})} \sum_{\lambda\in\Lambda_m} \sum_{A\in\mathcal{E}} \left[\left(\frac{2n}{p} - 1\right) \left(P_n^{(-A)}(\psi_\lambda)\right)^2 - \frac{2n}{p} P_n(\psi_\lambda)P_n^{(-A)}(\psi_\lambda) + \left(P_n(\psi_\lambda)\right)^2 \right] \\
&= \left(\frac{2n}{p} - 1\right) \frac{1}{\mathrm{Card}(\mathcal{E})} \sum_{\lambda\in\Lambda_m} \sum_{A\in\mathcal{E}} \left[\left(P_n^{(-A)}(\psi_\lambda)\right)^2 - \left(P_n(\psi_\lambda)\right)^2 \right] \quad , \qquad (41)
\end{aligned}
$$

where we used again Eq. (26) and (39). Comparing Eq. (40) and (41) gives

$$
\mathrm{crit}_{\mathcal{E}}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathcal{E}}\left(m, \frac{n}{p} - \frac{1}{2}\right) \qquad (42)
$$

which implies Eq. (6) and (7). Eq. (8) follows by [Ler12b]. □

Note than Eq. (8) can also be deduced from Proposition 2.1 in [Cel12], which proves that

$$
\mathrm{crit}_{\mathrm{LPO}}(m,p) = \frac{1}{n(n-p)} \sum_{\lambda\in\Lambda_m} \left(\sum_{i=1}^n \psi_\lambda(X_i)^2 - \frac{1}{n-1} \sum_{i\neq j=1}^n \psi_\lambda(X_i)\psi_\lambda(X_j) \right) \quad .
$$

Elementary algebraic computations show then that

$$
\mathrm{crit}_{\mathrm{LPO}}(m,p) - P_n\gamma(\widehat{s}_m) = \frac{2n-p}{n^2(n-p)} \sum_{\lambda\in\Lambda_m} \left(\sum_{i=1}^n \psi_\lambda(X_i)^2 - \frac{1}{n-1} \sum_{i\neq j=1}^n \psi_\lambda(X_i)\psi_\lambda(X_j) \right) \quad .
$$

From Lemma 1 and the latter equation, we obtain that, for any p, p'

$$
\frac{n/p-1}{n/p-1/2} \left(\mathrm{crit}_{\mathrm{LPO}}\left(m,p\right) - P_n\gamma(\widehat{s}_m)\right) = \frac{n/p'-1}{n/p'-1/2} \left(\mathrm{crit}_{\mathrm{LPO}}\left(m,p'\right) - P_n\gamma(\widehat{s}_m)\right) \quad .
$$

In particular, when $p' = 1$, from (7), since $\mathrm{pen}_{\mathrm{LPO}}(m,1,C) = \mathrm{pen}_{\mathrm{LOO}}(m,C)$,

$$
\begin{aligned}
\mathrm{pen}_{\mathrm{LPO}}\left(m,p,\frac{n}{p}-\frac{1}{2}\right) &= \frac{n/p-1/2}{n/p-1} \frac{n-1}{n-1/2} \mathrm{pen}_{\mathrm{LPO}}\left(m,1,n-\frac{1}{2}\right) \\
&= \mathrm{pen}_{\mathrm{LOO}}\left(m,(n-1)\frac{n/p-1/2}{n/p-1}\right) \quad .
\end{aligned}
$$

Appendix G. Proof of Theorem 1

The proof of Theorem 1 can be sketched as follows. First, we prove a general oracle inequality valid for any penalty approximately equal to $C\|\widehat{s}_m - s_m\|^2$ for some constant $C > 1$ (Lemma 12). Then, we show that the V-fold penalty satisfies this condition, which is mostly a consequence of Proposition 4. Finally, we check the assumptions of Theorem 1 imply those of Lemma 12.

G.1. **A general model selection theorem.**

Lemma 12. *Assume* (**H1**) *and* (**H4**) *hold true, and that*

$$\text{(H2g)}\qquad \exists y \geq 2, \Phi > 0\,,\ \exists \pi \in \mathcal{M}_1(\mathcal{M}_n) \quad \text{such that} \quad \forall m, m' \in \mathcal{M}_n\,, \quad \text{if} \quad x_m := -\ln\left(\pi(m)\right)\ ,$$
$$v_{m,m'}^2 \leq \Phi \frac{R_m \vee R_{m'}}{\sqrt{R_n^\star}} \quad \text{and} \quad (x_m + x_{m'} + y) b_{m,m'}^2 \leq 9\Phi n \frac{R_m \vee R_{m'}}{\sqrt{R_n^\star}}\ .$$

Let $\mathrm{pen} : \mathcal{M}_n \to [0, +\infty)$ *be some penalty function and*

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

Assume moreover, for y, $(x_m)_{m\in\mathcal{M}_n}$ *and* π *such that* (**H2g**) *holds true, that absolute constants* $(L_i)_{i=1,2}$, *a sequence* $(z_n)_{n\in\mathbb{N}}$ *and a constant* $c > 1/2$ *exist such that: for all* $m \in \mathcal{M}_n$, *a constant* c_m *and a function* $\varepsilon_{1,m}$ *exist such that*

$$\text{(C1)}\qquad y + x_m \leq z_n\,, \qquad \frac{z_n}{\sqrt{R_n^\star}} \xrightarrow[n\to+\infty]{} 0 \quad \text{and} \quad z_n \xrightarrow[n\to+\infty]{} +\infty\ ,$$

$$\text{(C2)}\qquad \forall x \leq c_m\,, \qquad \mathbb{P}\left\{ \left| c^{-1}\,\mathrm{pen}(m) - 2\left\|\widehat{s}_m - s_m\right\|^2 \right| \leq L_1 \varepsilon_{1,m}(x) \frac{R_m}{n} \right\} \geq 1 - L_2 e^{-x}\ .$$

Let $\nu_n := \sqrt{2\Phi}\left(R_n^\star\right)^{-1/4}\sqrt{z_n}$ *and* $\varepsilon_{2,m}(x) := c\varepsilon_{1,m}(x + x_m) + \nu_n$. *Then, some* $n_0 > 0$ *and absolute constants* $(L_i)_{i=3,4}$ *exist such that for all* $n \geq n_0$ *and for all* $x \leq y \wedge \inf_{m\in\mathcal{M}}(c_m - x_m)$, *with probability larger than* $1 - L_4 e^{-x}$, *for all* $m \in \mathcal{M}_n$,

$$\frac{1 - 2(c-1)_- - L_3\varepsilon_{2,\widehat{m}}(x)}{1 + 2(c-1)_+ + L_3\varepsilon_{2,m}(x)} \left\|\widehat{s}_{\widehat{m}} - s\right\|^2 \leq \left\|\widehat{s}_m - s\right\|^2\ . \tag{43}$$

The constant n_0 *depends on all parameters of assumptions* (**H2g**), (**C1**), (**C2**), *but it does not depend on* m.

Remark 2. Lemma 12 is an extension of Theorem 3.2 in [Ler12b], which corresponds to the particular case $c = 1$. We prove Lemma 12 here to study more general V-fold criteria, since some interesting ones are biased, as explained in Lemma 1. Lemma 12 will be used to prove both the results on V-fold procedures and those on hold-out penalties (see Sections 7.1 and I.1), this is why it is stated as a separate result.

Before proving Lemma 12, let us give a concentration inequality that we will need in the proof.

Lemma 13 (Corollary of Lemma 6.8 in [Ler12b])**.** *Assume* (**H2g**) *holds true for some* π, y, Φ, *and recall that* $x_m = -\ln(\pi(m))$. *Then, an absolute constant* $L > 0$ *exists such that, for all* $x \leq y$, *with probability larger than* $1 - e^{-x}$,

$$\exists m, m' \in \mathcal{M}_n,\ (P_n - P)(s_m - s_{m'}) \leq L\sqrt{\Phi}\frac{\sqrt{x + x_m + x_{m'}}}{(R_n^\star)^{1/4}}\left(\frac{R_m}{n} + \frac{R_{m'}}{n}\right). \tag{44}$$

Lemma 13 is proved in Section I.4.2.

Proof of Lemma 12. By definition of $\widehat{m}$ and Eq. (34), for all $m \in \mathcal{M}_n$,

$$\left\|s - \widehat{s}_{\widehat{m}}\right\|^2 \leq \left\|\widehat{s}_m - s\right\|^2 + \mathrm{pen}(m) - 2\left\|\widehat{s}_m - s_m\right\|^2 + 2\left\|\widehat{s}_{\widehat{m}} - s_{\widehat{m}}\right\|^2 - \mathrm{pen}(\widehat{m}) - 2(P_n - P)(s_m - s_{\widehat{m}})\ . \tag{45}$$

The idea of the proof is to use that $\mathrm{pen}(m) \approx \mathrm{pen}_{\mathrm{id}}(m)$ up to multiplying factors (condition (**C2**)) and a centered term (that is concentrated by Lemma 13). Then, we will show that the remainder terms are negligble in front of the risk, in particular by using the fact that $\|\widehat{s}_m - s_m\|$ is concentrated around its expectation.

Construction of a favorable event $\Omega(\mathcal{M}_n)$. Let $2 \le x \le y \wedge \inf_{m\in\mathcal{M}}(c_m - x_m)$. Recall that $C_\star$ is defined above Eq. (35). Let us define, for all $m \in \mathcal{M}_n$, $\zeta_m(x) = C_\star \left(R_n^\star\right)^{-1/4} \sqrt{x + x_m}$ and for some constant $L_5 > 0$ to be chosen later

$$\Omega_1(\mathcal{M}_n) := \left\{ \forall m \in \mathcal{M}_n\,,\ \left| \|\widehat{s}_m - s_m\|^2 - \frac{D_m}{n} \right| \le L_5 \zeta_m(x) \frac{R_m}{n} \right\}\ ,$$

$$\Omega_2(\mathcal{M}_n) := \left\{ \forall (m,m') \in \mathcal{M}_n^2\,,\ 2(P_n - P)(s_m - s_{m'}) \le L_5 \nu_n \left(\frac{R_m}{n} + \frac{R_{m'}}{n} \right) \right\}\ ,$$

$$\Omega_3(\mathcal{M}_n) := \left\{ \forall m \in \mathcal{M}_n\,,\ \left| c^{-1}\,\mathrm{pen}(m) - 2\,\|\widehat{s}_m - s_m\|^2 \right| \le L_5 \varepsilon_{1,m}(x_m + x) \frac{R_m}{n} \right\}\ .$$

From (**C1**), some constant $n_1 > 0$ exists such that, for all $n \ge n_1$, $z_n \le C_\star^{-2}\sqrt{R_n^\star}$. As $x_m \ge 0$ and $R_n^\star \le n$ from (**H4**), for any $2 \le x \le y$, we obtain $2 \le x + x_m \le \sqrt{R_n^\star \wedge n}/C_\star^2$, hence, Lemma 11 applies with $R_\star = R_n^\star$, and, using a union bound, we obtain an absolute constant L_6 such that, for all $n \ge n_1$, if $L_5 \ge L_6$,

$$\mathbb{P}\left\{\,\Omega_1(\mathcal{M}_n)^c\,\right\} \le 2 \sum_{m\in\mathcal{M}_n} e^{-x-x_m} = 2e^{-x}\ .$$

From (**H2g**) and the fact that $x \le y$, Lemma 13 applies. As $x_m + x_{m'} + x \le 2z_n$ by (**C1**), there exists n_2 and an absolute constant L_7 such that, for all $n \ge n_2$, if $L_5 \ge L_7$,

$$\mathbb{P}\left\{\,\Omega_2(\mathcal{M}_n)^c\,\right\} \le e^{-x}\ .$$

From condition (**C2**) and since $x + x_m \le c_m$, for all $m \in \mathcal{M}_n$ and all $L_5 \ge L_1$

$$\mathbb{P}\left\{\,\Omega_3(\mathcal{M}_n)^c\,\right\} \le L_2 e^{-x} \sum_{m\in\mathcal{M}_n} \pi(m) = L_2 e^{-x}\ .$$

Hence, choosing $L_5 = \max\{\,L_1, L_6, L_7\,\}$, $L_4 = L_2 + 3$, the event $\Omega(\mathcal{M}_n) := \Omega_1(\mathcal{M}_n) \cap \Omega_2(\mathcal{M}_n) \cap \Omega_3(\mathcal{M}_n)$ satisfies $\mathbb{P}\left\{\,\Omega(\mathcal{M}_n)^c\,\right\} \le L_4 e^{-x}$ if $n \ge \max\{\,n_1, n_2\,\}$.
Eq. (43) holds on $\Omega(\mathcal{M}_n)$. From (**C1**), there exists n_3 such that, for all $n \ge n_3$, $L_5 C_\star \left(R_n^\star\right)^{-1/4} \sqrt{z_n} \le 1/2$, so on $\Omega_1(\mathcal{M}_n)$, for all $n \ge n_3$, we have

$$\forall m \in \mathcal{M}_n\,, \qquad \|\widehat{s}_m - s\|^2 \ge \frac{1}{2}\frac{R_m}{n}\ .$$

Therefore, for $L_3 = 2L_5$, on $\Omega(\mathcal{M}_n)$, for all $m \in \mathcal{M}_n$ and $n \ge n_0 := \max\{\,n_1, n_2, n_3\,\}$,

$$\mathrm{pen}(m) - 2\,\|\widehat{s}_m - s_m\|^2 \le 2(c-1)_+\,\|s_m - \widehat{s}_m\|^2 + L_3 c \varepsilon_{1,m}(x_m + x)\,\|\widehat{s}_m - s\|^2\ ,$$

$$-\left(\mathrm{pen}(\widehat{m}) - 2\,\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\|^2\right) \le 2\,(c-1)_-\,\|s_{\widehat{m}} - \widehat{s}_{\widehat{m}}\|^2 + L_3 c \varepsilon_{1,\widehat{m}}(x_{\widehat{m}} + x)\,\|\widehat{s}_{\widehat{m}} - s\|^2\ ,$$

$$2(P_n - P)(s_m - s_{\widehat{m}}) \le L_3 \nu_n\,\|s - \widehat{s}_{\widehat{m}}\|^2 + L_3 \nu_n\,\|s - \widehat{s}_m\|^2\ .$$

Plugging these inequalities into Eq. (45) yields Eq. (43). □

G.2. **Proof of Theorem 1.** We will use in the proof the following lemma, showing the assumptions of Theorem 1 imply assumptions (**H2g**) and (**C1**) of Lemma 12.

Lemma 14. (**H2**), (**H2′**) *or* (**H2⋄**) *together with* (**H1**), (**H3**), (**H4**) *imply* (**H2g**) *and* (**C1**), *with* π *the uniform probability measure on* $\mathcal{M}_n$, $y = 2\ln n + \ln(L_4^{-1})$ *and* Φ *depending on the parameters appearing in the assumptions that hold true.*

Remark 3. Lemma 14 is the only part of the proof of Theorem 1 where we use one assumption among (**H2**), (**H2′**) or (**H2⋄**).

Proof of Lemma 14. (**H3**) ensures that $x_m = O(\ln n)$. Under (**H2′**), we have

$$v^2_{m,m'} \le \sup_{t\in\mathbb{B}_{m,m'}} \int t^2 s d\mu \le \|s\|_\infty \qquad \text{and} \qquad b^2_{m,m'} \le \Gamma n \ .$$

As $\frac{R_m \vee R_{m'}}{\sqrt{R^\star_n}} \ge \sqrt{R^\star_n} \ge \sqrt{c^-_R}(\ln n)^{2+r/2}$, (**C1**) and (**H2g**) hold for some constant $\Phi(\|s\|_\infty, \Gamma, c^-_R)$. Under (**H2**$^\diamond$), using successively the inequality $P((t-Pt)^2) \le P(t^2)$, Cauchy-Schwarz inequality, the triangular inequality and (**H1**), we have

$$\begin{aligned}
v^2_{m,m'} &\le \sup_{\sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}} a^2_\lambda \le 1} P\left[\left(\sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}} a_\lambda\psi_\lambda\right)^2\right] \\
&\le \sup_{\sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}} a^2_\lambda \le 1} \left\{ \left\| \sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}} a_\lambda\psi_\lambda \right\|_\infty \|s\| \left\| \sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}} a_\lambda\psi_\lambda \right\| \right\} \\
&\le (b_m + b_{m'})\,\|s\| \le 2\,\|s\| \sqrt{\frac{R_m \vee R_{m'}}{L_\star}} \ . \\
b^2_{m,m'} &\le 2(b^2_m + b^2_{m'}) \le \frac{4}{L_\star}\,(R_m \vee R_{m'}) \ .
\end{aligned}$$

These inequalities yield also (**H2g**) and (**C1**). Under Assumption (**H2**), we can choose a basis of $L^2(\mu)$ such that (**H2**$^\diamond$) holds, hence (**H2**) ensures also (**H2g**). □

Proof of Theorem 1. We apply Lemma 12. Proposition 4 ensures that Condition (**C2**) in Lemma 12 is fulfilled with L_1 for some absolute constant, $L_2 = e^2 + 4$,

$$c = \kappa, \quad \forall m \in \mathcal{M}\,, \quad c_m = \frac{\sqrt{R^\star_n}}{C^2_\star} \wedge V^{1/6}\,(R^\star_n)^{1/4}\,, \quad \varepsilon_{1,m}(x) = \varepsilon_2(n, V, x) \ .$$

From (**H4**), $\inf_{m\in\mathcal{M}_n} c_m \ge L\ln(n)^{1+r/4}$. Let $z_n = (3+\alpha_{\mathcal{M}})\ln n$, hence, for all n large enough, $x_m = \ln(\mathrm{Card}(\mathcal{M}_n))$, $x = 2\ln n + \ln(L_4^{-1})$, we have $x + x_m \le z_n$, $x + x_m \le c_m$ for all $m \in \mathcal{M}_n$, and $z_n/\sqrt{R^\star_n} \to 0$. Since (**C1**) holds for these values by Lemma 14, we deduce from Lemma 12 that Theorem 1 holds for n sufficiently large. Taking L such that $L\kappa\varepsilon(n, V) \ge 1$ for too small values of n yields the result. □

Appendix H. Proof of Theorem 2

Theorem 2 actually is a corollary of the following proposition, since $C_D = C_C = 0$ when $C = V - 1$, and V-fold cross-validation corresponds to $C = V - 1/2$ (see Lemma 1).

Proposition 15. *Let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ and $(\psi_\lambda)_{\lambda\in\Lambda_{m'}}$ be two finite orthonormal families of vectors of $L^4(\mu)$, $C > 0$ some constant, and define*

$$\begin{aligned}
C_B &= \frac{8}{n^3}\left[\frac{C^2(n-V+1)}{(V-1)^3} + \frac{2C}{V-1} + n - 1\right], \\
C_C &= \frac{4(C-V+1)}{n^2(V-1)}\,,\ C_D = \frac{4(C-V+1)^2}{(V-1)^2 n^3} \ .
\end{aligned}$$

Assume that $\mathcal{B}$ satisfies (**H5***) *and* $\mathrm{pen}_{\mathrm{VF}}(m) = \mathrm{pen}_{\mathrm{VF}}(m;\mathcal{B};C)$. *Then, with the notation of Theorem 2, and noting* $\beta(\Lambda) = \beta(\Lambda,\Lambda)$ *and similarly with* $\gamma(\cdot), \zeta(\cdot)$, *for every* $m \in \mathcal{M}_n$,

$$\mathrm{var}(\mathrm{pen}_{\mathrm{id}}(m)) = \frac{4}{n}\mathrm{var}_P(s_m) + \frac{8(1-n^{-1})}{n^2}\beta(\Lambda_m) + \frac{4}{n^2}\gamma(\Lambda_m) + \frac{4}{n^3}\zeta(\Lambda_m) \quad , \tag{46}$$

$$\mathrm{var}(2\|\widehat{s}_m - s_m\|^2) = \frac{8(1-n^{-1})}{n^2}\beta(\Lambda_m) + \frac{4}{n^3}\zeta(\Lambda_m) \quad , \tag{47}$$

$$\mathrm{var}\left(\mathrm{pen}_{\mathrm{VF}}(m)\right) = \frac{8C^2(n-V+1)}{n^3(V-1)^3}\beta(\Lambda_m) + \frac{4C^2}{n^3(V-1)^2}\zeta(\Lambda_m) \quad , \tag{48}$$

$$\mathrm{var}(\mathrm{pen}_{\mathrm{VF}}(m) - \mathrm{pen}_{\mathrm{id}}(m)) = \frac{4}{n}\mathrm{var}_P(s_m) + C_B\beta(\Lambda_m) - C_C\gamma(\Lambda_m) + C_D\zeta(\Lambda_m) \quad , \tag{49}$$

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

$$\begin{aligned}
&\mathrm{var}\left((\mathrm{pen}_{\mathrm{VF}}(m) - \mathrm{pen}_{\mathrm{id}}(m)) - (\mathrm{pen}_{\mathrm{VF}}(m') - \mathrm{pen}_{\mathrm{id}}(m'))\right) \\
&\qquad = \frac{4}{n}\mathrm{var}_P(s_m - s_{m'}) + C_B\mathbf{B}(\Lambda_m,\Lambda_{m'}) - C_C\mathbf{C}(\Lambda_m,\Lambda_{m'}) + C_D\mathbf{D}(\Lambda_m,\Lambda_{m'}) \quad ,
\end{aligned}$$

and for every $C > 0$ *and* $m \in \mathcal{M}_n$,

$$\begin{aligned}
&\mathrm{var}\left(P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{VF}}(m;\mathcal{B};C)\right) \\
&= \frac{4}{n}\mathrm{var}_P(s_m) + \frac{2}{n^2}\left[1 + \frac{4C^2}{(V-1)^3} - \frac{1}{n}\left(\frac{2C}{V-1} - 1\right)^2\right]\beta(\Lambda_m) \\
&\quad - \frac{2}{n^2}\left(\frac{2C}{V-1} - 1\right)\gamma(\Lambda_m) + \frac{1}{n^3}\left(\frac{2C}{V-1} - 1\right)^2\zeta(\Lambda_m) \quad .
\end{aligned} \tag{50}$$

The proof of Proposition 15 relies on the following lemma, by taking $\xi_{\lambda,i} = \psi_\lambda(X_i) - P(\psi_\lambda)$.

Lemma 16. *Let* Λ *be a discrete set,* $\Lambda_1, \Lambda_2 \subset \Lambda$ *non-empty,* $n \geq 2$, $\alpha \in \mathcal{M}_n(\mathbb{R})$ *symmetric,* $(\beta_\lambda)_{\lambda\in\Lambda} \in \mathbb{R}^\Lambda$ *and* $(\xi_{\lambda,1})_{\lambda\in\Lambda}, \ldots, (\xi_{\lambda,n})_{\lambda\in\Lambda}$ *a sequence of independent and identically distributed random variables with* $\forall i, \lambda$, $\mathbb{E}[\xi_{\lambda,i}] = 0$. *For every* $q, r \in \{1,2\}$ *and* $\lambda, \lambda' \in \Lambda$, *define*

$$\overline{v}_\lambda := \mathbb{E}\left[\xi_{\lambda,1}^2\right] \qquad \text{and} \qquad \overline{C}_{\lambda,\lambda'}^{(q,r)} := \mathbb{E}\left[\xi_{\lambda,1}^q \xi_{\lambda',1}^r\right] \quad .$$

Let us assume that $\sum_{\lambda\in\Lambda}\beta_\lambda^2 < \infty$, $\left\|\sum_{\lambda\in\Lambda}\xi_{\lambda,1}^2\right\|_\infty < \infty$. *Then, if for every* $\Lambda_a \subset \Lambda$,

$$Z(\Lambda_a) = \sum_{1\leq i,j\leq n}\sum_{\lambda\in\Lambda_a}(\alpha_{i,j}\xi_{\lambda,i}\xi_{\lambda,j}) + \sum_{1\leq i\leq n}\sum_{\lambda\in\Lambda_a}(\beta_\lambda\xi_{\lambda,i}) \quad , \tag{51}$$

$$\begin{aligned}
\mathrm{cov}(Z(\Lambda_1), Z(\Lambda_2)) &= n\sum_{\lambda\in\Lambda_1,\,\lambda'\in\Lambda_2}\left(\beta_\lambda\beta_{\lambda'}\overline{C}_{\lambda,\lambda'}^{(1,1)}\right) \\
&\quad + 2\left(\sum_{1\leq i\neq j\leq n}\alpha_{i,j}^2\right)\left(\sum_{\lambda\in\Lambda_1,\,\lambda'\in\Lambda_2}\left(\overline{C}_{\lambda,\lambda'}^{(1,1)}\right)^2\right) \\
&\quad + \left(\sum_{i=1}^n\alpha_{i,i}^2\right)\left[\sum_{\lambda\in\Lambda_1,\,\lambda'\in\Lambda_2}\overline{C}_{\lambda,\lambda'}^{(2,2)} - \left(\sum_{\lambda\in\Lambda_1}\overline{v}_\lambda\right)\left(\sum_{\lambda\in\Lambda_2}\overline{v}_\lambda\right)\right] \\
&\quad + \left(\sum_{i=1}^n\alpha_{i,i}\right)\sum_{\lambda\in\Lambda_1,\,\lambda'\in\Lambda_2}\left[\beta_{\lambda'}\overline{C}_{\lambda,\lambda'}^{(2,1)} + \beta_\lambda\overline{C}_{\lambda,\lambda'}^{(1,2)}\right] \quad .
\end{aligned} \tag{52}$$

Lemme 16 (p. 31) is proved in Section I.4.3. We can now prove Proposition 15.

Proof of Proposition 15. For all $i,j \in \{1\dots,n\}$ and $\lambda \in \Lambda := \Lambda_m \cup \Lambda_{m'}$, let

$$\xi_{\lambda,i} = \psi_\lambda(X_i) - P(\psi_\lambda) \qquad \text{and} \qquad E^{(W)}_{i,j} = \mathbb{E}\left[(W_i - 1)(W_j - 1)\right] \ ,$$

where, for all $i = 1,\dots,n$, $W_i = (1 - n^{-1}n_J)\mathbf{1}_{i\notin B_J}$ is the V-fold weight vector. For every $q,r \geq 0$ and $\lambda,\lambda' \in \Lambda_m \cup \Lambda_{m'}$, let $\overline{v}_\lambda = v_\lambda$ and $\overline{C}^{(q,r)}_{\lambda,\lambda'} = C^{(q,r)}_{\lambda,\lambda'}$. We have

$$\begin{aligned}
&\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}} \left[P(\psi_\lambda)P(\psi_{\lambda'})\overline{C}^{(1,1)}_{\lambda,\lambda'}\right] \\
&\qquad = \mathbb{E}\left[\sum_{\lambda\in\Lambda_m\,,\,\lambda'\in\Lambda_{m'}} \left(P(\psi_\lambda)\left(\psi_\lambda(\xi_1) - P(\psi_\lambda)\right)P(\psi_{\lambda'})\left(\psi_{\lambda'}(\xi_1) - P(\psi_{\lambda'})\right)\right)\right] \\
&\qquad = \mathbb{E}\left[\left(s_m(\xi_1) - \mathbb{E}\left[s_m(\xi_1)\right]\right)\left(s_{m'}(\xi_1) - \mathbb{E}\left[s_{m'}(\xi_1)\right]\right)\right] = \operatorname{cov}_P\left(s_m\,,\,s_{m'}\right) \qquad (53)
\end{aligned}$$

and remark that

$$\operatorname{var}_P(s_m) + \operatorname{var}_P(s_{m'}) - 2\operatorname{cov}_P\left(s_m\,,\,s_{m'}\right) = \operatorname{var}_P(s_m - s_{m'}) \ . \qquad (54)$$

Ideal penalty. For the ideal penalty, we simply notice that

$$\begin{aligned}
\operatorname{pen}_{\mathrm{id}}(m) &= 2(P_n - P)(\widehat{s}_m) = 2\sum_{\lambda\in\Lambda_m}\left[(P_n\psi_\lambda - P\psi_\lambda)(P_n\psi_\lambda)\right] \\
&= \frac{2}{n^2}\sum_{1\leq i,j\leq n}\sum_{\lambda\in\Lambda_m}\left(\xi_{\lambda,i}\xi_{\lambda,j}\right) + \frac{2}{n}\sum_{1\leq i\leq n}\sum_{\lambda\in\Lambda_m}\left(P(\psi_\lambda)\xi_{\lambda,i}\right) \ . \qquad (55)
\end{aligned}$$

Therefore, $\operatorname{pen}_{\mathrm{id}}(m)$ is of the form (51) with $\Lambda_a = \Lambda_m$ and

$$\forall i,j \in \{1,\dots,n\}\,, \quad \alpha_{i,j} = \frac{2}{n^2} \qquad \text{and} \qquad \forall \lambda \in \Lambda_m\,, \quad \beta_\lambda = \frac{2P(\psi_\lambda)}{n} \ ,$$

so that, by Lemma 16 and Eq. (53),

$$\operatorname{var}(\operatorname{pen}_{\mathrm{id}}(m)) = \frac{4}{n}\operatorname{var}_P(s_m) + \frac{8(n-1)}{n^3}\beta\left(\Lambda_m\right) + \frac{4}{n^2}\gamma\left(\Lambda_m\right) + \frac{4}{n^3}\zeta\left(\Lambda_m\right) \ .$$

Proof of Eq. (47). Since

$$\|s_m - \widehat{s}_m\|^2 = \sum_{\lambda\in\Lambda_m}\left[(P_n\psi_\lambda - P\psi_\lambda)^2\right] = \frac{1}{n^2}\sum_{1\leq i,j\leq n}\sum_{\lambda\in\Lambda_m}\left(\xi_{\lambda,i}\xi_{\lambda,j}\right) \ ,$$

$\|s_m - \widehat{s}_m\|^2$ is of the form (51) with $\Lambda_a = \Lambda_m$ and

$$\forall i,j \in \{1,\dots,n\}\,, \quad \alpha_{i,j} = \frac{1}{n^2} \qquad \text{and} \qquad \forall \lambda \in \Lambda_m, \quad \beta_\lambda = 0 \ ,$$

so that, by Lemma 16,

$$\operatorname{var}\left(\|s_m - \widehat{s}_m\|^2\right) = \frac{2(n-1)}{n^3}\beta\left(\Lambda_m\right) + \frac{1}{n^3}\zeta\left(\Lambda_m\right) \ .$$

V-fold penalty. It follows from Eq. (37) that

$$\operatorname{pen}_{\mathrm{VF}}(m) = \frac{2C}{n^2}\sum_{1\leq i,j\leq n}\sum_{\lambda\in\Lambda_m}\left(E^{(\mathrm{VF})}_{i,j}\xi_{\lambda,i}\xi_{\lambda,j}\right) \ , \qquad (56)$$

where $E^{(\mathrm{VF})}_{i,j}$ is computed in Lemma 8:

$$\forall I,J \in \{1,\dots,V\}\,, \forall i \in B_I, \forall j \in B_J, \quad E^{(\mathrm{VF})}_{i,j} = \frac{1}{V-1} - \frac{V\mathbf{1}_{I\neq J}}{(V-1)^2} \ ,$$

since all blocks B_I have the same size n/V. So,

$$\sum_{i=1}^{n} E_{i,i}^{(\mathrm{VF})} = \frac{n}{V-1} \qquad \sum_{i=1}^{n} \left(E_{i,i}^{(\mathrm{VF})} \right)^2 = \frac{n}{(V-1)^2} \ . \tag{57}$$

Using in addition that,

$$\sum_{1 \leq i \leq n} E_{i,i}^{(\mathrm{VF})} + \sum_{1 \leq i \neq j \leq n} E_{i,j}^{(\mathrm{VF})} = \sum_{1 \leq i,j \leq n} E_{i,j}^{(\mathrm{VF})} = \mathbb{E}\left[\left(\sum_{i=1}^{n} (W_i - \overline{W}_n) \right)^2 \right] = 0 \ , \tag{58}$$

Eq. (57) implies that

$$\sum_{1 \leq i \neq j \leq n} \left(E_{i,j}^{(\mathrm{VF})} \right) = \frac{-n}{V-1} \ . \tag{59}$$

In addition,

$$\sum_{1 \leq i \neq j \leq n} \left(E_{i,j}^{(\mathrm{VF})} \right)^2 = n \left(\frac{n}{V} - 1 \right) \frac{1}{(V-1)^2} + n \left(n - \frac{n}{V} \right) \frac{1}{(V-1)^4} = \frac{n(n-V+1)}{(V-1)^3} \ . \tag{60}$$

According to Eq. (56), $\mathrm{pen}_{\mathrm{VF}}(m) = \mathrm{pen}_{\mathrm{VF}}(m; \mathcal{B}; C)$ (without assuming $C = V-1$) is of the form (51) with $\Lambda_a = \Lambda_m$ and

$$\forall i,j \in \{1, \ldots, n\} \ , \quad \alpha_{i,j} = \frac{2C E_{i,j}^{(\mathrm{VF})}}{n^2} \qquad \text{and} \qquad \forall \lambda \in \Lambda_m \ , \quad \beta_\lambda = 0 \ .$$

From Eq. (57) and (60), we obtain that

$$\sum_{i=1}^{n} \alpha_{i,i}^2 = \frac{4C^2}{n^3(V-1)^2} \qquad \text{and} \qquad \sum_{1 \leq i \neq j \leq n} \alpha_{i,j}^2 = \frac{4C^2(n-V+1)}{n^3(V-1)^3} \ .$$

Therefore, by Lemma 16,

$$\mathrm{var}(\mathrm{pen}_{\mathrm{VF}}(m)) = \frac{8C^2(n-V+1)}{n^3(V-1)^3} \beta(\Lambda_m) + \frac{4C^2}{n^3(V-1)^2} \zeta(\Lambda_m) \ . \tag{61}$$

Difference of V-fold and ideal penalty. According to Eq. (55) and (56),

$$(\mathrm{pen}_{\mathrm{VF}}(m) - \mathrm{pen}_{\mathrm{id}}(m)) = Z_\Delta(\Lambda_m) \ , \qquad (\mathrm{pen}_{\mathrm{VF}}(m') - \mathrm{pen}_{\mathrm{id}}(m')) = Z_\Delta(\Lambda_{m'})$$

where $Z_\Delta(\cdot)$ is defined by Eq. (51) with

$$\forall i,j \in \{1, \ldots, n\}, \quad \alpha_{i,j} = \frac{2}{n^2} \left(C E_{i,j}^{(\mathrm{VF})} - 1 \right) \quad \text{and} \quad \forall \lambda \in \Lambda \ , \quad \beta_\lambda = \frac{-2P(\psi_\lambda)}{n} \ .$$

So, using Eq. (57), (59) and (60), we have

$$\sum_{i=1}^{n} \alpha_{i,i}^2 = \frac{4}{n^4} \left[C^2 \sum_{i=1}^{n} \left(E_{i,i}^{(\mathrm{VF})} \right)^2 - 2C \sum_{i=1}^{n} E_{i,i}^{(\mathrm{VF})} + n \right] = \frac{4(C-V+1)^2}{(V-1)^2 n^3} = C_D \ ,$$

$$\sum_{1\le i\ne j\le n}\alpha_{i,j}^2 = \frac{4}{n^4}\left[C^2\sum_{1\le i\ne j\le n}\left(E_{i,j}^{(\mathrm{VF})}\right)^2 - 2C\sum_{1\le i\ne j\le n}E_{i,j}^{(\mathrm{VF})} + n(n-1)\right]$$
$$= \frac{4}{n^4}\left[\frac{C^2n(n-V+1)}{(V-1)^3} + \frac{2Cn}{V-1} + n(n-1)\right]$$
$$= \frac{4}{n^3}\left[\frac{C^2(n-V+1)}{(V-1)^3} + \frac{2C}{V-1} + (n-1)\right] = \frac{C_B}{2}\ ,$$
$$\sum_{i=1}^n \alpha_{i,i} = \frac{2(C-V+1)}{n(V-1)} = \frac{C_C}{2}\ .$$

Therefore, by Lemma 16 with $\Lambda_a = \Lambda_m$, $\Lambda_b = \Lambda_{m'}$ and Eq. (53),

$$\mathrm{cov}\left(Z_\Delta(\Lambda_m),\, Z_\Delta(\Lambda_{m'})\right) = \frac{4}{n}\,\mathrm{cov}_P\left(s_m, s_{m'}\right) + C_B\beta\left(\Lambda_m,\Lambda_{m'}\right) - C_C\gamma\left(\Lambda_m,\Lambda_{m'}\right) + C_D\zeta\left(\Lambda_m,\Lambda_{m'}\right)$$

which gives Eq. (49) when $m = m'$. Eq. (**??**) follows then from Eq. (54) and the relation

$$\mathrm{var}\left(Z_\Delta(\Lambda_m) - Z_\Delta(\Lambda_{m'})\right) = \mathrm{var}\left(Z_\Delta(\Lambda_m)\right) + \mathrm{var}\left(Z_\Delta(\Lambda_{m'})\right) - 2\,\mathrm{cov}\left(Z_\Delta(\Lambda_m),\, Z_\Delta(\Lambda_{m'})\right)\ .$$

V-fold penalized criterion. Let $Z_C(\Lambda_m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}_{\mathrm{VF}}(m;\mathcal{B};C)$. Then,

$$P_n\gamma\left(\widehat{s}_m\right) = -\left\|\widehat{s}_m\right\|^2 = -\sum_{\lambda\in\Lambda_m}\left(P_n(\psi_\lambda)\right)^2$$
$$= -\sum_{\lambda\in\Lambda_m}\left((P_n-P)(\psi_\lambda)\right)^2 - 2\sum_{\lambda\in\Lambda_m}\left[P(\psi_\lambda)(P_n-P)(\psi_\lambda)\right] - \sum_{\lambda\in\Lambda_m}\left(P(\psi_\lambda)\right)^2$$
$$= \frac{-1}{n^2}\sum_{\lambda\in\Lambda_m}\sum_{1\le i,j\le n}\xi_{\lambda,i}\xi_{\lambda,j} - \frac{2}{n}\sum_{\lambda\in\Lambda_m}\left[P(\psi_\lambda)\sum_{i=1}^n\xi_{\lambda,i}\right] - \sum_{\lambda\in\Lambda_m}\left(P(\psi_\lambda)\right)^2$$

So, by Eq. (56), $Z_C(\Lambda_m) + \sum_{\lambda\in\Lambda_m}\left(P(\psi_\lambda)\right)^2$ is of the form of Eq. (51) with $\Lambda_a = \Lambda_m$,

$$\forall i,j\in\{1,\dots,n\}\,,\quad \alpha_{i,j} = \frac{2CE_{i,j}^{(\mathrm{VF})} - 1}{n^2} \qquad \text{and} \qquad \forall\lambda\in\Lambda_m\,,\quad \beta_\lambda = \frac{-2P(\psi_\lambda)}{n}$$

where $E_{i,j}^{(\mathrm{VF})}$ is defined in Lemma 8. Similarly to computations for Z_Δ,

$$\sum_{i=1}^n\alpha_{i,i} = \frac{1}{n}\left(\frac{2C}{V-1} - 1\right), \qquad \sum_{i=1}^n\alpha_{i,i}^2 = \frac{1}{n^3}\left(\frac{2C}{V-1} - 1\right)^2$$
$$\text{and} \quad \sum_{1\le i\ne j\le n}\alpha_{i,j}^2 = \frac{1}{n^3}\left[\frac{4C^2(n-V+1)}{(V-1)^3} + \frac{4C}{V-1} + n - 1\right]$$

so that Lemma 16 and Eq. (53) imply

$$\mathrm{var}\left(Z_C(\Lambda_m)\right) = \frac{4}{n}\,\mathrm{var}_P\left(s_m\right) + \frac{2}{n^3}\left[\frac{4C^2(n-V+1)}{(V-1)^3} + \frac{4C}{V-1} + n - 1\right]\beta\left(\Lambda_m\right)$$
$$- \frac{2}{n^2}\left(\frac{2C}{V-1} - 1\right)\gamma\left(\Lambda_m\right) + \frac{1}{n^3}\left(\frac{2C}{V-1} - 1\right)^2\zeta\left(\Lambda_m\right)\ ,$$

which proves Eq. (50). □

Appendix I. Supplementary material

The supplementary material is organized as follows. First, results concerning hold-out penalization are detailed in Section I.1, with the proof of the oracle inequality stated in Section 7.1 (Theorem 3) and an exact computation of the variance. Section I.2 gives expressions of the main terms appearing in Theorem 2 independent of the basis and evaluate these terms in the particular case of regular histograms. Section I.3 provides complements on the computational aspects stated in Section 6. In particular, we state and analyse the basic algorithm for computation V-fold criterions and we give the proof of Proposition 2. Then, several technical results of the main paper are proved in Section I.4. In Section I.5, we extend the results of the paper to pseudo-regular partitions, that is, when assumption (**H5***) is replaced by (**H5**). Some useful probabilistic tools are recalled in Section I.6. Finally, some simulation results are detailed in Section I.7, as a supplement to the ones of Section 5.

I.1. Results on hold-out penalization.

This section gathers the results we can prove on hold-penalization, similarly to the results already proved for V-fold penalization.

I.1.1. *Oracle inequality: proof of Theorem 3.* First note that (**H2**) implies (**H2g**) (Lemma 14) since we assume (**H1**), (**H3**) and (**H4**). So, we will prove Theorem 3 with (**H2**) replaced by (**H2g**).

Recall that for any $T \subset \{1,..,n\}$, $n_t = \mathrm{Card}(T)$, $n_v = n - n_t$ and $(\psi_\lambda)_{\lambda\in\Lambda_m}$ denotes an orthonormal basis of S_m. The hold-out penalty is equal to

$$\mathrm{pen}_{\mathrm{HO}}(m) = 2C(P_n^{(T)} - P_n^{(-T)})(\widehat{s}_m^{(T)} - \widehat{s}_m^{(-T)}) = 2C \sum_{\lambda\in\Lambda_m} \left[\left(P_n^{(T)} - P_n^{(-T)} \right) \psi_\lambda \right]^2 \quad . \tag{62}$$

As for Theorem 1, the sketch of the proof of Theorem 3 is to show the conditions of Lemma 12 are satisfied with the hold-out penalty. Since we need a concentration result for $\mathrm{pen}_{\mathrm{HO}}(m)$, we start by an exact formula for the hold-out penalty (Lemma 17, analogous to Proposition 3). Then, we get the concentration of $\mathrm{pen}_{\mathrm{HO}}(m)$ with Lemma 18 (analogous to Proposition 4) and Lemma 11.

Let us first state and prove the two lemmas that will be necessary in the proof.

Lemma 17. *For all $m \in \mathcal{M}_n$, we have*

$$\mathrm{pen}_{\mathrm{HO}}(m) = 2C \left[\left\| \widehat{s}_m^{(T)} - s_m \right\|^2 + \left\| \widehat{s}_m^{(-T)} - s_m \right\|^2 - 2(P_n^{(T)} - P) \left(\widehat{s}_m^{(-T)} - s_m \right) \right].$$

Proof of Lemma 17. By definition, we have

$$\begin{aligned}
\mathrm{pen}_{\mathrm{HO}}(m) &= 2C \sum_{\lambda\in\Lambda_m} \left\{ \left((P_n^{(-T)} - P)\psi_\lambda \right)^2 + \left((P_n^{(T)} - P)\psi_\lambda \right)^2 \right\} \\
&\quad -2C \sum_{\lambda\in\Lambda_m} \left\{ 2 \left((P_n^{(-T)} - P)\psi_\lambda \right) \left((P_n^{(T)} - P)\psi_\lambda \right) \right\} \\
&= 2C \left[\left\| \widehat{s}_m^{(-T)} - s_m \right\|^2 + \left\| \widehat{s}_m^{(T)} - s_m \right\|^2 \right. \\
&\quad \left. -2(P_n^{(T)} - P) \left(\sum_{\lambda\in\Lambda_m} \left((P_n^{(-T)} - P)\psi_\lambda \right) \psi_\lambda \right) \right].
\end{aligned}$$

□

Lemma 18. *An absolute constant $L > 0$ exists such that, for all $m \in \mathcal{M}_n$ and $x > 0$, with probability larger than $1 - 2e^{-x}$,*

$$\left|(P_n^{(T)} - P)\left(\widehat{s}_m^{(-T)} - s_m\right)\right| \leq LC_\star \left(\frac{\sqrt{x}}{(R_n^\star)^{1/4}} \vee \frac{x}{\sqrt{n_t}}\right)\left(\left\|\widehat{s}_m^{(-T)} - s_m\right\|^2 + \frac{R_m}{n_t}\right). \tag{63}$$

Proof of Lemma 18. Let us apply Bernstein's inequality conditionally to $(\xi_i)_{i\notin T}$ to the function $t = \left(\widehat{s}_m^{(-T)} - s_m\right)$. From Eq. (35),

$$\left\|\widehat{s}_m^{(-T)} - s_m\right\|_\infty \leq \left\|\widehat{s}_m^{(-T)} - s_m\right\| b_m \leq C_\star^2 \sqrt{R_m}\left\|\widehat{s}_m^{(-T)} - s_m\right\| \ ,$$
$$\operatorname{var}\left(\widehat{s}_m^{(-T)}(\xi) - s_m(\xi)\,\middle|\,(\xi_i)_{i\notin T}\right) \leq \left\|\widehat{s}_m^{(-T)} - s_m\right\|^2 v_m^2 \leq \frac{C_\star^2 R_m}{\sqrt{R_n^\star}}\left\|\widehat{s}_m^{(-T)} - s_m\right\|^2 \ .$$

Hence, for all $x > 0$, with probability larger than $1 - 2e^{-x}$, conditionally to $(\xi_i)_{i\notin T}$,

$$\begin{aligned}\left|(P_n^{(T)} - P)\left(\widehat{s}_m^{(-T)} - s_m\right)\right| &\leq \left\|\widehat{s}_m^{(-T)} - s_m\right\|\sqrt{\frac{C_\star^2 R_m}{n_t}}\left(\sqrt{\frac{2x}{\sqrt{R_n^\star}}} + \frac{C_\star x}{3\sqrt{n_t}}\right) \\ &\leq \frac{\eta}{2}\left\|\widehat{s}_m^{(-T)} - s_m\right\|^2 + \frac{C_\star^2 R_m}{\eta n_t}\left(\frac{2x}{\sqrt{R_n^\star}} + \frac{C_\star^2 x^2}{9n_t}\right) \ .\end{aligned}$$

As the bound on the probability does not depend on $(\xi_i)_{i\notin T}$, the same inequality holds unconditionally. We choose $\eta = (\sqrt{x}\,(R_n^\star)^{-1/4}) \wedge (xn_t^{-1/2})$ to conclude the proof. □

Proof of Theorem 3. For all $x > 0$, let $\varepsilon_n^{(T)}(x) := (R_n^\star \wedge n_v \wedge n_t)^{-1/2}\sqrt{x}$. We deduce from Lemmas 11, 17 and 18 that there exist absolute constants L, L' such that, for all $x \leq C_\star^{-2}\sqrt{R_n^\star \wedge n_v \wedge n_t}$, with probability larger than $1 - L'e^{-x}$

$$\forall m \in \mathcal{M}_n,\ \left|\frac{\mathrm{pen}_{\mathrm{HO}}(m)}{Cn^2(n_v n_t)^{-1}} - 2\left\|\widehat{s}_m - s_m\right\|^2\right| \leq L\varepsilon_n^{(T)}(x)\frac{R_m}{n} \ .$$

Therefore, conditions (**C2**) of Lemma 12 hold with

$$c = C\frac{n^2}{n_v n_t}, \qquad \varepsilon_{1,m}(x) = \varepsilon_n^{(T)}(x), \qquad c_m = \frac{\sqrt{R_n^\star \wedge n_t \wedge n_v}}{C_\star^2} \ .$$

Moreover, from Lemma 14, Condition (**C1**) holds with the uniform probability measure π on $\mathcal{M}_n$, $y = 2\ln n + \ln(L_4^{-1})$, $z_n = \ln(\pi(m)^{-1}) + y$. From Lemma 12, we deduce that there exist a constant L such that, with probability larger than $1 - n^{-2}$, $\forall m \in \mathcal{M}_n$,

$$\frac{1 + \left(2C\frac{n^2}{n_t n_v} - 1\right)_- - L\left(C\frac{n^2}{n_t n_v} \vee 1\right)\varepsilon_n^{(T)}(\ln n)}{1 + \left(2C\frac{n^2}{n_t n_v} - 1\right)_+ - L\left(C\frac{n^2}{n_t n_v} \vee 1\right)\varepsilon_n^{(T)}(\ln n)}\left\|\widehat{s}_{\widehat{m}} - s\right\|^2 \leq \left\|\widehat{s}_m - s\right\|^2 \ .$$

□

I.1.2. *Variance.*

Proposition 19. *Assume that* $card(T) = n_t \in \{1, \dots, n-1\}$. *Then, with the notations introduced in Proposition 15, for every* $m, m' \in \mathcal{M}_n$,

$$\text{(64)} \qquad \operatorname{var}(\operatorname{pen}_{\mathrm{HO}}(m)) = \frac{4C^2}{n_v^3}\left[\left(\frac{n}{n_t}-1\right)^3+1\right]\zeta(\Lambda_m) + \frac{8C^2 n}{n_v^3 n_t^3}\left(-3n_t^2+n(3n_t-n_t^2)+n^2(n_t-1)\right)\beta(\Lambda_m) \ ,$$

$$\text{(65)} \qquad \operatorname{var}(\operatorname{pen}_{\mathrm{HO}}(m)-\operatorname{pen}_{\mathrm{id}}(m)) = \frac{4}{n}\operatorname{var}_P(s_m) + 8\left[\frac{C^2}{n_v^3 n_t^3}\left(-3n_t^2+n(3n_t-n_t^2)+n^2(n_t-1)\right)-\frac{2C}{nn_v n_t}+\frac{n-1}{n^3}\right]\beta(\Lambda_m) - \frac{8}{n}\left(\frac{Cn}{n_v n_t}-\frac{1}{n}\right)\gamma(\Lambda_m)+4\left[\frac{C^2}{n_v^3}\left[\left(\frac{n_v}{n_t}\right)^3+1\right]-\frac{2C}{nn_v n_t}+\frac{1}{n^3}\right]\zeta(\Lambda_m) \ ,$$

and for every $m, m' \in \mathcal{M}_n$,

$$\text{(66)} \qquad \operatorname{var}\left((\operatorname{pen}_{\mathrm{HO}}(m)-\operatorname{pen}_{\mathrm{id}}(m))-(\operatorname{pen}_{\mathrm{HO}}(m')-\operatorname{pen}_{\mathrm{id}}(m'))\right) = \frac{4}{n}\operatorname{var}_P(s_m-s_{m'}) + 8\left[\frac{C^2}{n_v^3 n_t^3}\left(-3n_t^2+n(3n_t-n_t^2)+n^2(n_t-1)\right)-\frac{2C}{nn_v n_t}+\frac{n-1}{n^3}\right]\mathbf{B}(\Lambda_m,\Lambda_{m'}) - \frac{8}{n}\left(\frac{Cn}{n_v n_t}-\frac{1}{n}\right)\mathbf{C}(\Lambda_m,\Lambda_{m'})+4\left[\frac{C^2}{n_v^3}\left[\left(\frac{n_v}{n_t}\right)^3+1\right]-\frac{2C}{nn_v n_t}+\frac{1}{n^3}\right]\mathbf{D}(\Lambda_m,\Lambda_{m'}) \ .$$

Proof of Proposition 19. By definition we have $n_v P_n^{(-T)} = nP_n - n_t P_n^{(T)}$, i.e.,

$$P_n^{(T)} - P_n^{(-T)} = P_n^{(T)} - \frac{nP_n - n_t P_n^{(T)}}{n_v} = \frac{n}{n_v}\left(P_n^{(T)} - P_n\right) \ .$$

Therefore, we have

$$\begin{aligned}
\operatorname{pen}_{\mathrm{HO}}(m) &= 2C(P_n^{(T)}-P_n^{(-T)})(\widehat{s}_m^{(T)}-\widehat{s}_m^{(-T)}) = 2C\sum_{\lambda\in\Lambda_m}\left[\left(P_n^{(T)}-P_n^{(-T)}\right)\psi_\lambda\right]^2 \\
&= 2C\left(\frac{n}{n_v}\right)^2\sum_{\lambda\in\Lambda_m}\left[\left(P_n^{(T)}-P_n\right)\psi_\lambda\right]^2 \\
&= 2C\left(\frac{n}{n_v}\right)^2\sum_{\lambda\in\Lambda_m}\frac{1}{n^2}\left(\sum_{i=1}^n\left(\frac{n}{n_t}\mathbf{1}_{i\in T}-1\right)\psi_\lambda(X_i)\right)^2 \\
&= 2C\frac{1}{n_v^2}\sum_{\lambda\in\Lambda_m}\left(\sum_{i=1}^n\left(\frac{n}{n_t}\mathbf{1}_{i\in T}-1\right)(\psi_\lambda(X_i)-P\psi_\lambda)\right)^2 \\
&= 2C\frac{1}{n_v^2}\sum_{\lambda\in\Lambda_m}\sum_{i,j=1}^n E_{i,j}^{(\mathrm{HO})}\xi_{\lambda,i}\xi_{\lambda,j} \ . \qquad \text{(67)}
\end{aligned}$$

In the previous inequality, we wrote, for all $i, j \in \{1 \dots, n\}$ and $\lambda \in \Lambda_m$,

$$\xi_{\lambda,i} = \psi_\lambda(X_i) - P(\psi_\lambda) \qquad \text{and} \qquad E_{i,j}^{(\mathrm{HO})} = \left(\frac{n}{n_t}\mathbf{1}_{i\in T}-1\right)\left(\frac{n}{n_t}\mathbf{1}_{j\in T}-1\right) \ ,$$

the hold-out weight vector. We have

$$E^{(\mathrm{HO})}_{i,j} = \left(\frac{n}{n_t} - 1\right)^2 \mathbf{1}_{i,j\in T} + \left(1 - \frac{n}{n_t}\right) \mathbf{1}_{i\in T,\, j\notin T} + \left(1 - \frac{n}{n_t}\right) \mathbf{1}_{i\notin T,\, j\in T} + \mathbf{1}_{i,j\notin T} \ ,$$

Therefore,

$$\sum_{i=1}^{n} E^{(\mathrm{HO})}_{i,i} = n_t \left(\frac{n}{n_t} - 1\right)^2 + (n - n_t) = \frac{n(n-n_t)}{n_t} \tag{68}$$

$$\sum_{i=1}^{n} \left(E^{(\mathrm{HO})}_{i,i}\right)^2 = n_t \left(\frac{n}{n_t} - 1\right)^4 + (n - n_t) = (n-n_t)\left[\left(\frac{n}{n_t} - 1\right)^3 + 1\right] \ . \tag{69}$$

Eq. (58) also holds for the hold-out weight-vector, i.e.,

$$\sum_{1\le i,j\le n} E^{(\mathrm{HO})}_{i,j} = \mathbb{E}\left[\left(\sum_{i=1}^{n}\left(\frac{n}{n_t}\mathbf{1}_{i\in T} - 1\right)\right)^2\right] = 0 \ , \tag{70}$$

so Eq. (68) implies that

$$\sum_{1\le i\ne j\le n}\left(E^{(\mathrm{HO})}_{i,j}\right) = -\sum_{i=1}^{n} E^{(\mathrm{HO})}_{i,i} = \frac{-n(n-n_t)}{n_t} \ . \tag{71}$$

In addition,

$$\begin{aligned}
\sum_{1\le i\ne j\le n}\left(E^{(\mathrm{HO})}_{i,j}\right)^2 &= 2n_t(n-n_t)\left(\frac{n}{n_t} - 1\right)^2 + n_t(n_t-1)\left(\frac{n}{n_t} - 1\right)^4 + (n-n_t)(n-n_t-1) \\
&= \frac{(n-n_t)^3}{n_t}\left(2 + \frac{(n_t-1)(n-n_t)}{n_t^2} + \frac{n_t(n-n_t-1)}{(n-n_t)^2}\right) \\
&= \frac{n(n-n_t)}{n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right) \ . \qquad (72)
\end{aligned}$$

According to (67), $\mathrm{pen}_{\mathrm{HO}}(m)$ is of the form (51) with

$$\forall i,j \in \{1,\dots,n\}, \quad \alpha_{i,j} = \frac{2CE^{(\mathrm{HO})}_{i,j}}{n_v^2} \qquad \text{and} \qquad \forall \lambda \in \Lambda_m, \quad \beta_\lambda = 0 \ .$$

From Eq. (69) and (72), we obtain that

$$\begin{aligned}
\sum_{i=1}^{n} \alpha_{i,i}^2 &= \frac{4C^2}{n_v^3}\left[\left(\frac{n}{n_t} - 1\right)^3 + 1\right] \qquad \text{and} \\
\sum_{1\le i\ne j\le n} \alpha_{i,j}^2 &= \frac{4C^2 n}{n_v^3 n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right) \ .
\end{aligned}$$

Therefore, by Lemma 16 applied with $\Lambda_1 = \Lambda_2 = \Lambda_m$,

$$\begin{aligned}
\mathrm{var}(\mathrm{pen}_{\mathrm{HO}}(m)) &= \frac{4C^2}{n_v^3}\left[\left(\frac{n}{n_t} - 1\right)^3 + 1\right]\zeta(\Lambda_m) \\
&\quad + \frac{8C^2 n}{n_v^3 n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right)\beta(\Lambda_m) \ . \qquad (73)
\end{aligned}$$

According to Eq. (67) and (55), $\mathrm{pen}_{\mathrm{HO}}(m) - \mathrm{pen}_{\mathrm{id}}(m)$ is of the form (51) with

$$\forall i,j \in \{1,\dots,n\}, \quad \alpha_{i,j} = 2\left(C\frac{E_{i,j}^{(\mathrm{HO})}}{n_v^2} - \frac{1}{n^2}\right) \qquad \text{and} \qquad \forall \lambda \in \Lambda_m, \quad \beta_\lambda = \frac{-2P(\psi_\lambda)}{n} \ .$$

So, using Eq. (68), (69), (71) and (72), we have

$$\begin{aligned}
\sum_{i=1}^{n} \alpha_{i,i}^2 &= 4\left[\frac{C^2}{n_v^4}\sum_{i=1}^{n}\left(E_{i,i}^{(\mathrm{HO})}\right)^2 - \frac{2C}{n^2 n_v^2}\sum_{i=1}^{n} E_{i,i}^{(\mathrm{HO})} + \frac{1}{n^3}\right] \\
&= 4\left[\frac{C^2}{n_v^3}\left[\left(\frac{n_v}{n_t}\right)^3 + 1\right] - \frac{2C}{n n_v n_t} + \frac{1}{n^3}\right] \qquad (74) \\
\sum_{1\le i\neq j\le n} \alpha_{i,j}^2 &= 4\left[\frac{C^2}{n_v^4}\sum_{1\le i\neq j\le n}\left(E_{i,j}^{(\mathrm{HO})}\right)^2 - \frac{2C}{n^2 n_v^2}\sum_{1\le i\neq j\le n} E_{i,j}^{(\mathrm{HO})} + \frac{n-1}{n^3}\right] \\
&= 4\left[\frac{C^2}{n_v^3 n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right) + \frac{2C}{n n_v n_t} + \frac{n-1}{n^3}\right] \qquad (75) \\
\sum_{i=1}^{n} \alpha_{i,i} &= 2\left(\frac{Cn}{n_v n_t} - \frac{1}{n}\right) \ . \qquad (76)
\end{aligned}$$

Therefore, by Lemma 16 with $\Lambda_1 = \Lambda_2 = \Lambda_m$ and by Eq. (53), we deduce

$$\begin{aligned}
\mathrm{var}\left(\mathrm{pen}_{\mathrm{HO}}(m) - \mathrm{pen}_{\mathrm{id}}(m)\right) &= 4\left[\frac{C^2}{n_v^3}\left[\left(\frac{n_v}{n_t}\right)^3 + 1\right] - \frac{2C}{n n_v n_t} + \frac{1}{n^3}\right]\zeta\left(\Lambda_m\right) \\
&+ 8\left[\frac{C^2}{n_v^3 n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right) - \frac{2C}{n n_v n_t} + \frac{n-1}{n^3}\right]\beta\left(\Lambda_m\right) \\
&- \frac{8}{n}\left(\frac{Cn}{n_v n_t} - \frac{1}{n}\right)\gamma\left(\Lambda_m\right) + \frac{4}{n}\,\mathrm{var}_P\left(s_m\right) \ .
\end{aligned}$$

Let us now remark that

$$\mathrm{pen}_{\mathrm{HO}}(m) - \mathrm{pen}_{\mathrm{id}}(m) - \mathrm{pen}_{\mathrm{HO}}(m') - \mathrm{pen}_{\mathrm{id}}(m') = Z(\Lambda_m) - Z(\Lambda_{m'}) \ ,$$

where, for all $a \in \{m, m'\}$, $Z(\Lambda_a)$ is defined by Eq. (51) with

$$\forall i,j \in \{1,\dots,n\}, \quad \alpha_{i,j} = 2\left(C\frac{E_{i,j}^{(\mathrm{HO})}}{n_v^2} - \frac{1}{n^2}\right) \qquad \text{and} \qquad \forall \lambda \in \Lambda_a, \quad \beta_\lambda = \frac{-2P(\psi_\lambda)}{n} \ .$$

It comes therefore from Lemma 16 and Eq. (74), (75) and (76) that, for all $a, b \in (m, m')^2$,

$$\begin{aligned}
\mathrm{cov}\left(Z(\Lambda_a), Z(\Lambda_b)\right) &= 4\left[\frac{C^2}{n_v^3}\left[\left(\frac{n_v}{n_t}\right)^3 + 1\right] - \frac{2C}{n n_v n_t} + \frac{1}{n^3}\right]\zeta\left(\Lambda_a, \Lambda_b\right) \\
&+ 8\left[\frac{C^2}{n_v^3 n_t^3}\left(-3n_t^2 + n(3n_t - n_t^2) + n^2(n_t-1)\right) - \frac{2C}{n n_v n_t} + \frac{n-1}{n^3}\right]\beta\left(\Lambda_a, \Lambda_b\right) \\
&- \frac{8}{n}\left(\frac{Cn}{n_v n_t} - \frac{1}{n}\right)\gamma\left(\Lambda_a, \Lambda_b\right) + \frac{4}{n}\,\mathrm{cov}_P\left(\Lambda_a, \Lambda_b\right) \ .
\end{aligned}$$

Eq. (66) follows then from Eq. (54) and the relation

$$\mathrm{var}\left(Z(\Lambda_m) - Z(\Lambda_{m'})\right) = \mathrm{cov}\left(Z(\Lambda_m), Z(\Lambda_m)\right) - 2\,\mathrm{cov}\left(Z(\Lambda_m), Z(\Lambda_{m'})\right) + \mathrm{cov}\left(Z(\Lambda_{m'}), Z(\Lambda_{m'})\right) \ .$$

□

I.2. **More results on the variance.** Theorem 2 and Proposition 15 provide general exact formula that can seem a bit too abstract. This section provides tools for understanding the terms appearing in these results.
Evaluation of the terms in the variance term. First, we give a formulation of the terms appearing in Theorem 2 and Proposition 15 that do not depend of the basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$. We will use the notation of Theorem 2, Proposition 15 and Section C. Recall that by Corollary 6, $\Psi_{\Lambda_m} := \sum_{\lambda\in\Lambda_m}\psi_\lambda^2 = \sup_{t\in\mathbb{B}_m} t^2$, $T_{\Lambda_m} := \sum_{\lambda\in\Lambda_m}\xi_{\lambda,1}^2 = \Psi_{\Lambda_m} - 2s_m + \|s_m\|^2$ are independent of the basis $(\psi_\lambda)_{\lambda\in\Lambda}$.

Proposition 20. *For any $m, m' \in \mathcal{M}_n$, we have*

$$\beta(\Lambda_m,\Lambda_{m'}) = \left\|\sup_{t\in\mathbb{B}_m}\Pi_{\Lambda_{m'}}(ts)\right\|^2 - 2P[s_m s_{m'}] + \|s_m\|^2\|s_{m'}\|^2 \quad , \tag{77}$$

$$\begin{aligned}\gamma(\Lambda_m,\Lambda_{m'}) &= P\left(T_{\Lambda_m}(s_{m'}-Ps_{m'}) + T_{\Lambda_{m'}}(s_m - Ps_m)\right)\\ &= \mathrm{cov}_P(\Psi_{\Lambda_m}, s_{m'}) + \mathrm{cov}_P\left(\Psi_{\Lambda_{m'}}, s_m\right) - 4\,\mathrm{cov}_P(s_m, s_{m'}) \quad .\end{aligned} \tag{78}$$

$$\zeta(\Lambda_m,\Lambda_{m'}) = \mathrm{cov}_P\left(T_{\Lambda_m}, T_{\Lambda_{m'}}\right) = \mathrm{cov}_P\left(\Psi_{\Lambda_m} - 2s_m, \Psi_{\Lambda_{m'}} - 2s_{m'}\right) \quad , \tag{79}$$

$$\mathbf{D}(\Lambda_m,\Lambda_{m'}) = \mathrm{var}_P\left(T_{\Lambda_m} - T_{\Lambda_{m'}}\right) = \mathrm{var}_P\left(\Psi_{\Lambda_m} - \Psi_{\Lambda_{m'}} - 2(s_m - s_{m'})\right) \tag{80}$$

$$\mathbf{C}(\Lambda_m,\Lambda_{m'}) = 2\,\mathrm{cov}_P\left(\Psi_{\Lambda_m} - \Psi_{\Lambda_{m'}}, s_m - s_{m'}\right) - 4\,\mathrm{var}_P(s_m - s_{m'}) \quad . \tag{81}$$

Proof of Proposition 20. **The terms $\zeta(\Lambda_a,\Lambda_b)$:** A direct computation shows that

$$\begin{aligned}\zeta(\Lambda_m,\Lambda_{m'}) &:= \left[\sum_{\lambda\in\Lambda_m,\,\lambda'\in\Lambda_{m'}} C_{\lambda,\lambda'}^{(2,2)} - \left(\sum_{\lambda\in\Lambda_m} v_\lambda\right)\left(\sum_{\lambda\in\Lambda_{m'}} v_\lambda\right)\right]\\ &= \mathbb{E}\left(\sum_{\lambda\in\Lambda_m,\,\lambda'\in\Lambda_{m'}}\xi_{\lambda,1}^2\xi_{\lambda',1}^2\right) - \left(\mathbb{E}\left(\sum_{\lambda\in\Lambda_m}\xi_{\lambda,1}^2\right)\right)\left(\mathbb{E}\left(\sum_{\lambda'\in\Lambda_{m'}}\xi_{\lambda',1}^2\right)\right)\\ &= \mathrm{cov}\left(\sum_{\lambda\in\Lambda_m}\xi_{\lambda,1}^2, \sum_{\lambda'\in\Lambda_{m'}}\xi_{\lambda',1}^2\right) = \mathrm{cov}_P\left(T_{\Lambda_m}, T_{\Lambda_{m'}}\right).\end{aligned}$$

The result for $\mathbf{D}(\Lambda_m,\Lambda_{m'})$ follows then immediately from the definition

$$\mathbf{D}(\Lambda_m,\Lambda_{m'}) = \zeta(\Lambda_m,\Lambda_m) + \zeta(\Lambda_{m'},\Lambda_{m'}) - 2\zeta(\Lambda_m,\Lambda_{m'}) \quad .$$

The terms $\gamma(\Lambda_m,\Lambda_{m'})$: By definition, we have

$$\begin{aligned}\gamma(\Lambda_m,\Lambda_{m'}) &:= \sum_{\lambda\in\Lambda_m,\,\lambda'\in\Lambda_{m'}}\left[\beta_{\lambda'}C_{\lambda,\lambda'}^{(2,1)} + \beta_\lambda C_{\lambda,\lambda'}^{(1,2)}\right]\\ &= \sum_{\lambda\in\Lambda_m,\,\lambda'\in\Lambda_{m'}}\beta_{\lambda'}\mathbb{E}\left((\psi_\lambda - P\psi_\lambda)^2(\psi_{\lambda'} - P\psi_{\lambda'})\right) + \beta_\lambda\mathbb{E}\left((\psi_{\lambda'} - P\psi_{\lambda'})^2(\psi_\lambda - P\psi_\lambda)\right)\\ &= P\left(T_{\Lambda_m}(s_{m'} - Ps_{m'}) + T_{\Lambda_{m'}}(s_m - Ps_m)\right).\end{aligned}$$

We have then, by definition

$$\begin{aligned}\mathbf{C}\left(\Lambda_m,\Lambda_{m'}\right) &:= \gamma\left(\Lambda_m,\Lambda_m\right)+\gamma\left(\Lambda_{m'},\Lambda_{m'}\right)-2\gamma\left(\Lambda_m,\Lambda_{m'}\right)\\ &= 2P\left(T_{\Lambda_m}(s_m-Ps_m)+T_{\Lambda_{m'}}(s_{m'}-Ps_{m'})\right)\\ &- 2P\left(T_{\Lambda_m}(s_{m'}-Ps_{m'})+T_{\Lambda_{m'}}(s_m-Ps_m)\right)\\ &= 2P\left(T_{\Lambda_m}\left((s_m-s_m)-P(s_m-s_{m'})\right)\right)\\ &+ 2P\left(T_{\Lambda_{m'}}\left((s_{m'}-s_m)-P(s_{m'}-s_m)\right)\right)\\ &= 2P\left((T_{\Lambda_m}-T_{\Lambda_{m'}})(s_m-s_{m'}-P(s_m-s_{m'}))\right)\ .\end{aligned}$$

The terms $\beta\left(\Lambda_m,\Lambda_{m'}\right)$: By definition, we have

$$\begin{aligned}\beta\left(\Lambda_m,\Lambda_{m'}\right) &:= \sum_{\lambda\in\Lambda_m\,,\,\lambda'\in\Lambda_{m'}}\left(C^{(1,1)}_{\lambda,\lambda'}\right)^2=\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\operatorname{cov}\left(\psi_\lambda,\psi_{\lambda'}\right)^2=\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\left(\psi_\lambda\psi_{\lambda'}\right)-P\psi_\lambda P\psi_{\lambda'}\right)^2\\ &= \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\left(\psi_\lambda\psi_{\lambda'}\right)\right)^2-2\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}P\psi_\lambda P\psi_{\lambda'}P\left(\psi_\lambda\psi_{\lambda'}\right)+\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\psi_\lambda P\psi_{\lambda'}\right)^2\\ &= \sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\left(\psi_\lambda\psi_{\lambda'}\right)\right)^2-2P\left[s_ms_{m'}\right]+\|s_m\|^2\|s_{m'}\|^2\,.\end{aligned}$$

By definition of $\Pi_{\Lambda_{m'}}$, using Lemma 5,

$$\begin{aligned}\sum_{\substack{\lambda\in\Lambda_m\\ \lambda'\in\Lambda_{m'}}}\left(P\left(\psi_\lambda\psi_{\lambda'}\right)\right)^2 &= \sum_{\lambda\in\Lambda_m}\left\|\Pi_{\Lambda_{m'}}\left(\psi_\lambda s\right)\right\|^2=\int\sum_{\lambda\in\Lambda_m}\left(\Pi_{\Lambda_{m'}}\left(\psi_\lambda s\right)\right)^2d\mu\\ &= \int\sup_{\sum_{\lambda\in\Lambda_m}a_\lambda^2\le1}\left(\Pi_{\Lambda_{m'}}\left(\sum_{\lambda\in\Lambda_m}a_\lambda\psi_\lambda s\right)\right)^2d\mu=\int\sup_{t\in\mathbb{B}_m}\left(\Pi_{\Lambda_{m'}}\left(ts\right)\right)^2d\mu\ .\quad\square\end{aligned}$$

□

Evaluation of the variance in the regular histogram case. In this section, we fix some integers $d_m,d_{m'}\ge1$ and, for $a\in\{m,m'\}$ consider the model S_a of regular histograms on $\mathbb{R}$ with step size d_a^{-1}. In other words,

$$\Lambda_a:=d_a^{-1}\mathbb{Z}\qquad\text{and}\qquad\forall\lambda\in\Lambda_a\,,\quad I_\lambda=[\lambda,\lambda+d_a^{-1})\,,\quad\psi_\lambda=\sqrt{d_a}\mathbf{1}_{I_\lambda}\ .$$

We also introduce the linear span $S_\star$ of $S_m\cup S_{m'}$ which is the set of histograms on the partition $(I_\lambda\cap I_{\lambda'})_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}$ of $\mathbb{R}$. We denote by $d_\star$ the dimension of $S_\star$. We define, for all $a\in\{\star,m,m'\}$, the orthogonal projection Π_a onto S_a and $s_a:=\Pi_a(s)$. In addition to the general properties of least-squares density estimation, regular histogram satisfy the following: S_a is stable by product,

$$\forall x\in\mathbb{R}\,,\,\forall k\in\mathbb{N}\,,\,\forall a\in\left\{m,m'\right\},\quad\sum_{\lambda\in\Lambda_a}\psi_\lambda^k(x)=d_a^{k/2}\ , \tag{82}$$

$$\forall a\in\left\{m,m'\right\}\,,\,\forall\lambda,\lambda'\in\Lambda_a\,,\quad\psi_\lambda\psi_{\lambda'}=\sqrt{d_a}\mathbf{1}_{\lambda=\lambda'}\psi_\lambda\ , \tag{83}$$

$$\forall a\in\left\{\star,m,m'\right\}\,\forall t\in S_a,\,\forall f\in L^2(\mu)\,,\quad\Pi_a\left(tf\right)=t\Pi_a\left(f\right)\ . \tag{84}$$

Proof of Eq. (84). For any $t\in S_a$,

$$tf=t\Pi_a\left(f\right)+t(f-\Pi_a\left(f\right))\ ,$$

with $t\Pi_a(f) \in S_a$ since S_a is stable by product, and $t(f - \Pi_a(f))$ is orthogonal to S_a since

$$\forall u \in S_a\,, \quad \langle t(f - \Pi_a(f)),\, u\rangle_{L_2(\mu)} = \langle (f - \Pi_a(f)),\, tu\rangle_{L_2(\mu)} = 0$$

since $tu \in S_a$ (using again S_a is stable by product). □

In particular, from Eq. (82), we have $\forall a \in \{m, m'\}$, $\Psi_{\Lambda_a} = d_a$ is constant. We will also use that in general,

$$\sum_{\lambda \in \Lambda_a} (P\psi_\lambda)\, \psi_\lambda(x) = s_a(x) \ .$$

The following proposition gives the orders of magnitude of the terms involved in Proposition 20 for such regular histogram models.

Proposition 21. *For all a, $b \in \{m, m'\}$, we have*

$$\zeta(\Lambda_a, \Lambda_b) = -\gamma(\Lambda_a, \Lambda_b) = 4\,\mathrm{cov}_P(s_a, s_b)$$

$$\text{so that} \qquad \mathbf{D}(\Lambda_m, \Lambda_{m'}) = -\mathbf{C}(\Lambda_m, \Lambda_{m'}) = 4\,\mathrm{var}_P(s_m - s_{m'}) \ . \tag{85}$$

Moreover, assume a constant $L > 0$ exists such that

$$\forall \lambda \in \Lambda_m\,,\ \forall \lambda' \in \Lambda_{m'}\,, \quad I_\lambda \cap I_{\lambda'} \neq \emptyset \ \Rightarrow\ L^{-1} d_\star^{-1} \leq \mu(I_\lambda \cap I_{\lambda'}) \leq L d_\star^{-1} \ . \tag{86}$$

Then, we have

$$\begin{aligned} d_m \|s_m\|^2 + d_{m'} \|s_{m'}\|^2 - 2L\frac{d_m d_{m'}}{d_\star} \|s_\star\|^2 &\leq \mathbf{B}(\Lambda_m, \Lambda_{m'}) + 2P\left((s_m - s_{m'})^2\right) - \left(\|s_m\|^2 - \|s_{m'}\|^2\right)^2 \\ &\leq d_m \|s_m\|^2 + d_{m'} \|s_{m'}\|^2 - 2L^{-1}\frac{d_m d_{m'}}{d_\star} \|s_\star\|^2 \ . \end{aligned} \tag{87}$$

Remark 4. Eq. (87) is of particular interest when $S_m \subset S_{m'}$ since then Eq. (86) holds with $L = 1$ and $d_\star = d_{m'}$, so that

$$\begin{aligned} \mathbf{B}(\Lambda_m, \Lambda_{m'}) &= d_m \|s_m\|^2 + (d_{m'} - 2d_m) \|s_{m'}\|^2 - 2P\left((s_m - s_{m'})^2\right) + \left(\|s_m\|^2 - \|s_{m'}\|^2\right)^2 \\ &= d_m \|s_m\|^2 + (d_{m'} - 2d_m) \|s_{m'}\|^2 - \mathrm{var}_P(s_m - s_{m'}) - P\left((s_m - s_{m'})^2\right) \ . \end{aligned}$$

We have

$$0 \leq \mathrm{var}_P(s_m - s_{m'}) + P\left((s_m - s_{m'})^2\right) \leq 2\|s\|_\infty \|s\|^2 \ .$$

Therefore, when $d_{m'}$ is large, the main term in $\mathbf{B}(\Lambda_m, \Lambda_{m'})$ is given by $d_m \|s_m\|^2 + (d_{m'} - 2d_m)\|s_{m'}\|^2$. Moreover, d_m is an integer dividing $d_{m'}$, hence, $d_m = d_{m'}/2$ or $d_m \leq d_m/3$. In the first case

$$d_m \|s_m\|^2 + (d_{m'} - 2d_m) \|s_{m'}\|^2 = \frac{\|s_m\|}{2} d_{m'} \ ,$$

in the second case

$$\frac{\|s_{m'}\|^2}{3} d_{m'} \leq d_m \|s_m\|^2 + (d_{m'} - 2d_m) \|s_{m'}\|^2 \leq d_{m'} \left(\|s_{m'}\|^2 + \|s_m\|^2\right) \leq 2d_{m'} \|s_{m'}\|^2 \ .$$

In particular, assuming $d_{m'}$ is large enough and $\|s_{m'}\| \geq c\|s\| > 0$ for some constant c, we get that

$$\frac{d_{m'}}{L} \leq \mathbf{B}(\Lambda_m, \Lambda_{m'}) \leq L d_{m'}$$

for some positive constant L.

Proof of Proposition 21. Eq. (85) follows from Proposition 20 and the fact that Ψ_{Λ_m} is constant for histograms. We now prove Eq. (87). Recall that

$$\left\| \sup_{t\in\mathbb{B}_m} \left(\Pi_{\Lambda_m}(ts) \right) \right\|^2 = \sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}} (P\psi_\lambda\psi_{\lambda'})^2 = d_m d_m \sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}} [P(I_\lambda\cap I_{\lambda'})]^2 \ .$$

If $m = m'$, then

$$\left\| \sup_{t\in\mathbb{B}_m} \left(\Pi_{\Lambda_m}(ts) \right) \right\|^2 = d_m \left\| s_m \right\|^2 \ .$$

If $m \neq m'$, then

$$\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}} [P(I_\lambda\cap I_{\lambda'})]^2 = \sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'};I_\lambda\cap I_{\lambda'}\neq\emptyset} \mu(I_\lambda\cap I_{\lambda'}) \frac{[P(I_\lambda\cap I_{\lambda'})]^2}{\mu(I_\lambda\cap I_{\lambda'})} \ .$$

It follows then from the regularity condition (86) that

$$L^{-1}\frac{d_m d_{m'}}{d_\star}\left\|s_\star\right\|^2 \leq \left\| \sup_{t\in\mathbb{B}_m} \left(\Pi_{\Lambda_{m'}}(ts) \right) \right\|^2 \leq L\frac{d_m d_{m'}}{d_\star}\left\|s_\star\right\|^2 \ .$$

□

Sharpness of our concentration inequality. This section shows how our concentration result for the V-fold penalty (Proposition 4) rewrites in the case of regular histogram models, so we can compare the deviation bounds with the variance computations (Propositions 15 and 21).

Proposition 22. *Let S_m be the space of regular histograms introduced in Section I.2, with $1 \leq d_m \leq n$, let s_m be the projection of s onto S_m and let $\widehat{s}_m$ be the projection estimator on S_m. Assume that $\|s\|_\infty \leq B_\star < \infty$. Then, some constant $c = c(B_\star)$ exists such that the following holds: For all $x \leq d_m^{1/2} \wedge (n/\sqrt{d_m})^{1/3}$,*

$$P\left(\left| \left\|\widehat{s}_m - s_m\right\|^2 - \frac{d_m}{n} \right| \leq c\frac{d_m\sqrt{x}}{n}\left(\frac{1}{\sqrt{d_m}} \vee \left(\frac{1}{n}\right)^{1/4} \right) \right) \geq 1 - 2e^{-x} \ .$$

Let U_m, $U_{\mathcal{B},m}$ be the U-statistics defined in Section A on S_m with the basis defined in Section I.2. For all $x \leq d_m^{1/2} \wedge (n/\sqrt{d_m})^{1/3}$, we have

$$P\left(|U_m| \leq c\frac{d_m\sqrt{x}}{n}\left(\frac{1}{\sqrt{d_m}} \vee \left(\frac{1}{n}\right)^{1/4} \right) \right) \geq 1 - 2e^{-x} \ .$$

For all $x \leq \sqrt{d_m} \wedge (n/V)^{1/4}$,

$$P\left(|U_{\mathcal{B},m}| \leq c\frac{d_m x}{n}\left(\frac{1}{\sqrt{d_m V}} \vee \left(\frac{1}{nV}\right)^{1/4} \right) \right) \geq 1 - 2e^{-x} \ .$$

Remark 5. We get that, for $d_m \leq n^{1/2}$, the deviations of the V-fold penalty are of order $\sqrt{d_m x}/n$ and that the dependence in V of the first-order term is proportional to $1 + V^{-1/2}$. This is what is expected by our computations of the variance, so our concentration inequalities are sharp in this case. When $d_m \geq n^{1/2}$, the main term in the concentration inequality is not sharp anymore.

Proof of Proposition 22. By [Ler11] (see Proposition 28), there exists an absolute constant c such that, for all $\epsilon \in (0,1]$, with probability larger than $1 - 2e^{-x}$,

$$\left| \left\|\widehat{s}_m - s_m\right\|^2 - \frac{d_m - \left\|s_m\right\|^2}{n} \right| \leq c\left(\epsilon\frac{d_m - \left\|s_m\right\|^2}{n} + \frac{v_m^2 x}{\epsilon n} + \frac{\Psi_{\Lambda_m} x^2}{\epsilon^3 n^2} \right) \ ,$$

where

$$v_m^2 := \sup_{t\in\mathbb{B}_m} \int t^2 s d\mu \leq B_\star \quad \text{and} \quad \Psi_{\Lambda_m} = d_m \ .$$

If $d_m \leq n^{1/2}$, we choose then $\epsilon = \sqrt{x/d_m}$; if $d_m \geq n^{1/2}$, we choose $\epsilon = \sqrt{x}(1/n)^{1/4}$ to conclude. From Lemma 9,

$$P\left(|U_{\mathcal{B},m}| \leq C\left(\epsilon\frac{d_m}{n\sqrt{V}} + \frac{x^2}{\epsilon n\sqrt{V}} + \frac{\sqrt{V}d_m x^4}{\epsilon^3 n^2}\right)\right) \geq 1 - e^{2-x} \ .$$

Choose $\epsilon = x/\sqrt{d_m}$ if $Vd_m^2 \leq n$ and $\epsilon = x(V/n)^{1/4}$ if $Vd_m^2 \geq n$ to conclude the proof. □

I.3. **Complements on computations questions.** This section gathers the proofs of the statements made in Section 6. First, we state more precisely the naive algorithm briefly discussed there and we prove its complexity. Then, we prove Proposition 2.

I.3.1. *Naive algorithm.*

Algorithm 2.

Input: $\mathcal{B}$ some partition of $\{1,...,n\}$ satisfying (**H5***), $\xi_1,\ldots,\xi_n \in \mathcal{X}$ and $(\psi_\lambda)_{\lambda\in\Lambda_m}$ a finite orthogonal family of $L^2(\mu)$, with $\mathrm{Card}(m) = d_m$.

(1) For $j \in \{1,\ldots,V\}$,
   (a) train $\widehat{s}_m(\cdot)$ with the data set $(\xi_i)_{i\notin\mathcal{B}_j}$, that is, for all $\lambda \in \Lambda_m$, compute $\alpha_{\lambda,j} := P_n^{(-\mathcal{B}_j)}(\psi_\lambda) = \frac{V}{(V-1)n}\sum_{i\notin\mathcal{B}_j}\psi_\lambda(\xi_i)$ so that $\widehat{s}_m^{(-\mathcal{B}_j)} = \sum_{\lambda\in\Lambda_m}\alpha_{\lambda,j}\psi_\lambda$
   (b) compute the norm of $\widehat{s}_m^{(-\mathcal{B}_j)}$: $N_j := \sum_{\lambda\in\Lambda_m}\alpha_{\lambda,j}^2$
   (c) compute $Q_j := P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right) = \frac{V}{n}\sum_{\lambda\in\Lambda_m}\sum_{i\in\mathcal{B}_j}\alpha_{\lambda,j}\psi_\lambda(\xi_i)$
   (d) compute $R_j := P_n^{(-\mathcal{B}_j)}\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right) = \frac{V}{n(V-1)}\sum_{\lambda\in\Lambda_m}\sum_{i\notin\mathcal{B}_j}\alpha_{\lambda,j}\psi_\lambda(\xi_i)$
(2) Compute the V-fold cross-validation criterion: $\mathcal{C} = V^{-1}\sum_{j=1}^V(N_j - 2Q_j)$
(3) Empirical risk:
   (a) Train $\widehat{s}_m(\cdot)$ with the data set $(\xi_i)_{1\leq i\leq n}$, that is, for all $\lambda \in \Lambda_m$, compute $\alpha_\lambda := P_n(\psi_\lambda) = \frac{1}{n}\sum_{i=1}^n\psi_\lambda(\xi_i)$ so that $\widehat{s}_m = \sum_{\lambda\in\Lambda_m}\alpha_\lambda\psi_\lambda$
   (b) compute the norm of $\widehat{s}_m$: $N := \sum_{\lambda\in\Lambda_m}\alpha_\lambda^2$
   (c) compute $R := \frac{1}{n}\sum_{\lambda\in\Lambda_m}\sum_{i=1}^n\alpha_\lambda\psi_\lambda(\xi_i)$
(4) Compute the V-fold penalty: $\mathcal{D} := 2(V-1)V^{-2}\sum_{j=1}^V(Q_j - R_j)$

Output:
Empirical risk: $N - 2R$
V-fold cross-validation estimator of the risk of $\widehat{s}_m$: $\mathrm{crit}_{\mathrm{VFCV}}(m) = \mathcal{C}$
V-fold penalty: $\mathrm{pen}_{\mathrm{VF}}(m) = \mathcal{D}$

Assuming the computational cost of evaluation ψ_λ at some point $\xi \in \Xi$ is of order 1, the computational cost of this naive algorithm 2 is as follows: $n(V-1)d_m$ for step 1, V for steps 2 and 4, nd_m for step 3. So the overall cost of computing the V-fold penalization criterion for m is of order nVd_m

I.3.2. *Proof of Proposition 2.* Let us first note that for every $i \in \{1,\ldots,V\}$ and $\lambda \in \Lambda_m$, $A_{i,\lambda} = P_n^{(\mathcal{B}_i)}(\psi_\lambda)$. So, at step 2, for every $i,j \in \{1,\ldots,V\}$,

$$C_{i,j} = \sum_{\lambda\in\Lambda_m} P_n^{(\mathcal{B}_i)}(\psi_\lambda)P_n^{(\mathcal{B}_j)}(\psi_\lambda) = P_n^{(\mathcal{B}_i)}\left(\sum_{\lambda\in\Lambda_m} P_n^{(\mathcal{B}_j)}(\psi_\lambda)\psi_\lambda\right) = P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_j)})$$

and by symmetry $C_{i,j} = C_{j,i} = P_n^{(\mathcal{B}_j)}(\widehat{s}_m^{(\mathcal{B}_i)})$.
Correctness of Algorithm 1. By assumption (**H5***), we have

$$P_n = \frac{1}{V}\sum_{j=1}^{V} P_n^{(\mathcal{B}_j)} \qquad \widehat{s}_m = \frac{1}{V}\sum_{j=1}^{V} \widehat{s}_m^{(\mathcal{B}_j)} \qquad P_n^{(-\mathcal{B}_i)} = \frac{1}{V}\sum_{\substack{1\le j\le V\\ j\neq i}} P_n^{(\mathcal{B}_j)} \quad \text{and} \quad \widehat{s}_m^{(-\mathcal{B}_i)} = \frac{1}{V}\sum_{\substack{1\le j\le V\\ j\neq i}} \widehat{s}_m^{(\mathcal{B}_j)} \ .$$

Therefore,

$$\|\widehat{s}_m\|^2 = -P_n\gamma(\widehat{s}_m) = P_n(\widehat{s}_m) = \frac{1}{V^2}\sum_{1\le i,j\le V} P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_j)}) = \frac{1}{V^2}\mathcal{S}$$

and

$$\begin{aligned}
\mathrm{crit}_{\mathrm{VFCV}}(m) &= \frac{1}{V}\sum_{j=1}^{V} P_n^{(\mathcal{B}_j)}\gamma\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right)\\
&= \frac{1}{V}\sum_{j=1}^{V}\left(\left\|\widehat{s}_m^{(-\mathcal{B}_j)}\right\|^2 - 2P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(-\mathcal{B}_j)}\right)\right)\\
&= \frac{1}{V}\sum_{j=1}^{V}\left(\frac{1}{(V-1)^2}\sum_{\substack{1\le i,\ell\le V\\ i,\ell\neq j}} P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)}) - \frac{2}{V-1}\sum_{i\neq j} P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(\mathcal{B}_i)}\right)\right)\\
&= \frac{1}{V(V-1)^2}\sum_{1\le i,\ell\le V}\left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)})\sum_{j=1}^{V}\mathbf{1}_{i\neq j\,,\,\ell\neq j}\right) - \frac{2}{V(V-1)}\sum_{1\le i\neq j\le V} P_n^{(\mathcal{B}_j)}\left(\widehat{s}_m^{(\mathcal{B}_i)}\right)\\
&= \frac{1}{V(V-1)^2}\sum_{1\le i,\ell\le V}\left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)})(V-1-\mathbf{1}_{i\neq\ell})\right) - \frac{2}{V(V-1)}\left(\mathcal{S}-\mathcal{T}\right)\\
&= \frac{1}{V(V-1)}\sum_{1\le i\le V}\left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_i)})\right) + \frac{V-2}{V(V-1)^2}\sum_{1\le i\neq\ell\le V}\left(P_n^{(\mathcal{B}_i)}(\widehat{s}_m^{(\mathcal{B}_\ell)})\right) - \frac{2}{V(V-1)}\left(\mathcal{S}-\mathcal{T}\right)\\
&= \frac{1}{V(V-1)}\mathcal{T} + \frac{V-2}{V(V-1)^2}\left(\mathcal{S}-\mathcal{T}\right) - \frac{2}{V(V-1)}\left(\mathcal{S}-\mathcal{T}\right)\\
&= \frac{1}{V(V-1)}\mathcal{T} - \frac{1}{(V-1)^2}\left(\mathcal{S}-\mathcal{T}\right) \ ,
\end{aligned}$$

so the formula for $\mathrm{crit}_{\mathrm{VFCV}}$ is correct. Lemma 1 implies the formula for $\mathrm{pen}_{\mathrm{VF}}$ is also correct.
Computational cost of Algorithm 1. Step 1 has a cost of order $V \times d_m \times (n/V) = nd_m$. Step 2 has a cost of order $V^2 d_m$. Step 3 has a cost of order V^2. Summing the three steps yields the result.
Computational cost for histograms. In the histogram case, step 1 can be performed with a cost of order $Vd_m + n$. Indeed, one can initialize the $V \times d_m$ matrix A with zeros (cost: Vd_m), and then go sequentially through the data set: for $j = 1, \ldots, n$, find the unique $i(j) \in \{1, \ldots, V\}$ such that $j \in \mathcal{B}_{i(j)}$, the unique $\lambda(j) \in \Lambda_m$ such that $\xi_j \in \lambda(j)$, and add $(V/n)\psi_\lambda(\xi_j)$ to $A_{(i(j),\lambda(j))}$. Since the partitions $\mathcal{B}$ and Λ_m can be coded so that finding $i(j)$ and $\lambda(j)$ has a cost of order 1, the resulting cost of step 1 is $Vd_m + n$, hence the overall cost is of order $V^2 d_m + n$. □

I.4. **Proofs of technical results of the main paper.**

I.4.1. *Proof of Lemma 8.* Let us first recall that, from (**H5***), for all $K \in \{1, ..., V\}$, we have

$$1 - \frac{n_K}{n} = 1 - \frac{1}{V}, \qquad \text{hence} \qquad \frac{1}{1 - n_K/n} = \frac{1}{1 - V^{-1}} = \frac{V}{V-1} \ .$$

It follows that

$$\mathbb{E}(W_i) = \frac{1}{V} \sum_{K=1}^{V} \frac{1}{1 - n_K/n} \mathbf{1}_{i \notin \mathcal{B}_K} = 1 \ .$$

When $K_i = K_j$, we have

$$\mathbb{E}(W_i W_j) = \frac{1}{V} \sum_{K=1}^{V} \frac{1}{(1 - n_K/n)^2} \mathbf{1}_{i \notin \mathcal{B}_K} \mathbf{1}_{j \notin \mathcal{B}_K} = \frac{V}{(V-1)^2} \sum_{\substack{K \neq K_i \\ 1 \leq K \leq V}} 1 = 1 + \frac{1}{V-1} \ .$$

When $K_i \neq K_j$, we have

$$\mathbb{E}(W_i W_j) = \frac{1}{V} \sum_{K=1}^{V} \frac{1}{(1 - n_K/n)^2} \mathbf{1}_{i \notin \mathcal{B}_K} \mathbf{1}_{j \notin \mathcal{B}_K} = \frac{V}{(V-1)^2} \sum_{\substack{1 \leq K \leq V \\ K \neq K_i \, , \, K \neq K_j}} 1 = 1 - \frac{1}{(V-1)^2} \ .$$

Hence,

$$\mathbb{E}\left((W_i - 1)(W_j - 1)\right) = \mathbb{E}(W_i W_j) - \mathbb{E}(W_i) - \mathbb{E}(W_j) + 1 = \frac{1}{V-1} - \frac{V \mathbf{1}_{K_i \neq K_j}}{(V-1)^2} \ .$$

I.4.2. *Proof of Lemma 13.* From Lemma 6.8 in [Ler12b], for all m, $m' \in \mathcal{M}_n$ and all $u, \eta > 0$, with probability larger than $1 - e^{-u}$,

$$(P_n - P)(s_m - s_{m'}) \leq \frac{\eta}{2} \|s_m - s_{m'}\|^2 + \frac{2uv_{m,m'}^2 + u^2 b_{m,m'}^2/(9n)}{\eta n} \ .$$

For all $x \leq y$, (**H2g**) ensures that

$$v_{m,m'}^2 \leq \Phi \frac{R_m \vee R_{m'}}{\sqrt{R_n^\star}}, \ (x + x_m + x_{m'}) \frac{b_{m,m'}^2}{n} \leq 9\Phi \frac{R_m \vee R_{m'}}{\sqrt{R_n^\star}} \ .$$

The triangular inequality gives

$$\|s_m - s_{m'}\|^2 \leq 2\|s - s_m\|^2 + 2\|s - s_{m'}\|^2 \leq 2 \frac{R_m + R_{m'}}{n} \ .$$

Take $u = x + x_m + x_{m'}$ and $\eta = \sqrt{3\Phi}\sqrt{x + x_m + x_{m'}}(R_n^\star)^{-1/4}$. We have $\sum_{m,m' \in \mathcal{M}_n} e^{-x - x_m - x_{m'}} = e^{-x}$, hence, a union bound gives, for $L = 2\sqrt{3}$,

$$\mathbb{P}\left(\exists m, m' \in \mathcal{M}_n, \ (P_n - P)(s_m - s_{m'}) > L\sqrt{\Phi} \frac{\sqrt{x + x_m + x_{m'}}}{(R_n^\star)^{1/4}} \left(\frac{R_m}{n} + \frac{R_{m'}}{n} \right) \right) \leq e^{-x} \ .$$

I.4.3. *Proof of Lemma 16.* For every $\Lambda_a \subset \Lambda$, we define

$$Z_1(\Lambda_a) := \sum_{1 \leq i,j \leq n} \sum_{\lambda \in \Lambda_a} (\alpha_{i,j} \xi_{\lambda,i} \xi_{\lambda,j}) \quad \text{and} \quad Z_2(\Lambda_a) := \sum_{1 \leq i \leq n} \sum_{\lambda \in \Lambda_a} (\beta_\lambda \xi_{\lambda,i}) \ .$$

First, since the $\xi_{\lambda,i}$ are centered, and the random vectors $(\xi_{\lambda,i})_{\lambda\in\Lambda}$ are independent,

$$\mathbb{E}\left[Z_1(\Lambda_a)\right] = \sum_{i=1}^{n}\sum_{\lambda\in\Lambda_a}(\alpha_{i,i})\,\mathbb{E}\left[\xi_{\lambda,i}^2\right] = \left(\sum_{i=1}^{n}\alpha_{i,i}\right)\left(\sum_{\lambda\in\Lambda_a}\overline{v}_\lambda\right) \tag{88}$$

$$\text{and}\quad \mathbb{E}\left[Z_2(\Lambda_a)\right] = 0\ . \tag{89}$$

Second, using repeatedly that the $\xi_{\lambda,i}$ are centered, and the random vectors $(\xi_{\lambda,i})_{\lambda\in\Lambda}$ are independent, we get: for every $\Lambda_1, \Lambda_2 \subset \Lambda$,

$$\begin{aligned}
&\mathbb{E}\left[Z_1(\Lambda_1)\,Z_1(\Lambda_2)\right] = \sum_{1\le i,j,k,\ell\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,j}\alpha_{k,\ell}\mathbb{E}\left[\xi_{\lambda,i}\xi_{\lambda,j}\xi_{\lambda',k}\xi_{\lambda',\ell}\right]\right)\\
&=\sum_{i=1}^{n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,i}^2\mathbb{E}\left[\xi_{\lambda,i}^2\xi_{\lambda',i}^2\right]\right) + \sum_{1\le i\ne k\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,i}\alpha_{k,k}\mathbb{E}\left[\xi_{\lambda,i}^2\xi_{\lambda',k}^2\right]\right)\\
&\quad+\sum_{1\le i\ne j\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,j}^2\mathbb{E}\left[\xi_{\lambda,i}\xi_{\lambda,j}\xi_{\lambda',i}\xi_{\lambda',j}\right]\right) + \sum_{1\le i\ne j\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,j}^2\mathbb{E}\left[\xi_{\lambda,i}\xi_{\lambda,j}\xi_{\lambda',j}\xi_{\lambda',i}\right]\right)
\end{aligned} \tag{90}$$

$$\begin{aligned}
&=\left(\sum_{i=1}^{n}\alpha_{i,i}^2\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\overline{C}_{\lambda,\lambda'}^{(2,2)}\right) + \left[\left(\sum_{1\le i\le n}\alpha_{i,i}\right)^2 - \left(\sum_{1\le i\le n}\alpha_{i,i}^2\right)\right]\left(\sum_{\lambda\in\Lambda_1}\overline{v}_\lambda\right)\left(\sum_{\lambda\in\Lambda_2}\overline{v}_\lambda\right)\\
&\quad+2\left(\sum_{1\le i\ne j\le n}\alpha_{i,j}^2\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\overline{C}_{\lambda,\lambda'}^{(1,1)}\right)^2\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[Z_1(\Lambda_1)\,Z_2(\Lambda_2)\right] &= \sum_{1\le i,j,k\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,j}\beta_{\lambda'}\mathbb{E}\left[\xi_{\lambda,i}\xi_{\lambda,j}\xi_{\lambda',k}\right]\right)\\
&=\sum_{i=1}^{n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\alpha_{i,i}\beta_{\lambda'}\overline{C}_{\lambda,\lambda'}^{(2,1)}\right) = \left(\sum_{i=1}^{n}\alpha_{i,i}\right)\left(\sum_{\lambda\in\Lambda_a\,,\,\lambda'\in\Lambda_b}\left[\beta_{\lambda'}\overline{C}_{\lambda,\lambda'}^{(2,1)}\right]\right)
\end{aligned} \tag{91}$$

$$\mathbb{E}\left[Z_2(\Lambda_1)\,Z_2(\Lambda_2)\right] = \sum_{1\le i,j\le n}\ \sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\beta_\lambda\beta_{\lambda'}\mathbb{E}\left[\xi_{\lambda,i}\xi_{\lambda',j}\right]\right) = n\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\beta_\lambda\beta_{\lambda'}\overline{C}_{\lambda,\lambda'}^{(1,1)}\right)\ . \tag{92}$$

The result follows from the combination of (88), (89), (90), (91) and (92) since

$$
\begin{aligned}
&\operatorname{cov}\left(Z_1(\Lambda_1)+Z_2(\Lambda_1)\,,\,Z_1(\Lambda_2)+Z_2(\Lambda_2)\right)\\
&=\mathbb{E}\left[Z_1(\Lambda_1)Z_1(\Lambda_2)\right]+\mathbb{E}\left[Z_1(\Lambda_1)Z_2(\Lambda_2)\right]+\mathbb{E}\left[Z_2(\Lambda_1)Z_1(\Lambda_2)\right]\\
&\quad+\mathbb{E}\left[Z_2(\Lambda_1)Z_2(\Lambda_2)\right]-\mathbb{E}\left[Z_1(\Lambda_1)+Z_2(\Lambda_1)\right]\mathbb{E}\left[Z_1(\Lambda_2)+Z_2(\Lambda_2)\right]\\
&=\left(\sum_{i=1}^n\alpha_{i,i}^2\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\overline{C}^{(2,2)}_{\lambda,\lambda'}\right)+\left[\left(\sum_{1\le i\le n}\alpha_{i,i}\right)^2-\left(\sum_{1\le i\le n}\alpha_{i,i}^2\right)\right]\left(\sum_{\lambda\in\Lambda_1}\overline{v}_\lambda\right)\left(\sum_{\lambda\in\Lambda_2}\overline{v}_\lambda\right)\\
&\quad+2\left(\sum_{1\le i\ne j\le n}\alpha_{i,j}^2\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\overline{C}^{(1,1)}_{\lambda,\lambda'}\right)^2\right)+\left(\sum_{i=1}^n\alpha_{i,i}\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left[\beta_{\lambda'}\overline{C}^{(2,1)}_{\lambda,\lambda'}\right]\right)\\
&\quad+\left(\sum_{i=1}^n\alpha_{i,i}\right)\left(\sum_{\lambda\in\Lambda_2\,,\,\lambda'\in\Lambda_1}\left[\beta_{\lambda'}\overline{C}^{(2,1)}_{\lambda,\lambda'}\right]\right)+n\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\beta_\lambda\beta_{\lambda'}\overline{C}^{(1,1)}_{\lambda,\lambda'}\right)\\
&\quad-\left(\sum_{i=1}^n\alpha_{i,i}\right)^2\left(\sum_{\lambda\in\Lambda_1}\overline{v}_\lambda\right)\left(\sum_{\lambda\in\Lambda_2}\overline{v}_\lambda\right)\\
&=\left(\sum_{i=1}^n\alpha_{i,i}^2\right)\left[\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\overline{C}^{(2,2)}_{\lambda,\lambda'}-\left(\sum_{\lambda\in\Lambda_1}\overline{v}_\lambda\right)\left(\sum_{\lambda\in\Lambda_2}\overline{v}_\lambda\right)\right]\\
&\quad+2\left(\sum_{1\le i\ne j\le n}\alpha_{i,j}^2\right)\left(\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\overline{C}^{(1,1)}_{\lambda,\lambda'}\right)^2\right)\\
&\quad+\left(\sum_{i=1}^n\alpha_{i,i}\right)\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left[\beta_{\lambda'}\overline{C}^{(2,1)}_{\lambda,\lambda'}+\beta_\lambda\overline{C}^{(1,2)}_{\lambda,\lambda'}\right]+n\sum_{\lambda\in\Lambda_1\,,\,\lambda'\in\Lambda_2}\left(\beta_\lambda\beta_{\lambda'}\overline{C}^{(1,1)}_{\lambda,\lambda'}\right)\ .
\end{aligned}
$$

I.5. **Extension of the results on V-fold penalties to general pseudo-regular partitions.** Extending Theorem 1 to partitions $\mathcal{B}$ satisfying (**H5**) instead of (**H5***) essentially requires to extend Propositions 3 and 4 to pseudo-regular partitions. Then, the proof of Theorem 1 straightforwardly yields an oracle inequality under assumption (**H5**).

I.5.1. *Exact formula for V-fold penalties: the general case.* In this section, we extend Proposition 3 to partitions $\mathcal{B}$ satisfying (**H5**) instead of (**H5***).

Lemma 23. *Let $V\ge 2$, $n\ge 4$ and $\mathcal{B}$ satisfying* (**H5**). *A function $\delta:\{1,\dots,V\}^2\to[-32,32]$ (defined by Eq. (95)) exists such that for every $m\in\mathcal{M}_n$,*

(93)

$$\frac{V-1}{2C}\operatorname{pen}_{\mathrm{VF}}(m)=\|\widehat{s}_m-s_m\|^2-\frac{V}{V-1}\left(U_m-U_{\mathcal{B},m}\right)+R_{\operatorname{pen}_{\mathrm{VF}}}(m,\mathcal{B}),\quad \textit{where}$$

(94)

$$R_{\operatorname{pen}_{\mathrm{VF}}}(m,\mathcal{B}):=\frac{1}{n^3}\sum_{\lambda\in\Lambda_m}\sum_{K,K'=1}^{V}\delta(K,K')\left(\sum_{i\in\mathcal{B}_K}(\psi_\lambda(X_i)-P\psi_\lambda)\right)\left(\sum_{j\in\mathcal{B}_{K'}}(\psi_\lambda(X_j)-P\psi_\lambda)\right)\ .$$

Furthermore, if (**H5***) *holds, $\delta\equiv 0$.*

Proving Lemma 23 requires the following lemma about the "covariance" of the weights W_i.

Lemma 24. *Assume that $V \geq 2$ and $n \geq 4$ and that* (**H5**) *holds. For all $i \in \{1, ..., n\}$, let K_i be the index of the block $\mathcal{B}_K$ such that $i \in B_{K_i}$. Then, a function $\delta : \{1, \dots, V\}^2 \to \mathbb{R}$ exists such that for all $i, j \in \{1, \dots, n\}$,*

$$E_{i,j}^{(\mathrm{VF})} := \mathbb{E}\left((W_i - 1)(W_j - 1)\right) = \frac{1}{V-1} - \frac{V}{(V-1)^2}\mathbf{1}_{K_i \neq K_j} + \frac{\delta(K_i, K_j)}{n(V-1)} \quad \text{and} \quad |\delta(K_i, K_j)| \leq 32 \ . \tag{95}$$

Furthermore, if (**H5***) *holds, $\delta \equiv 0$.*

Proof of Lemma 24. Let us first recall that, from (**H5**), for all $K \in \{1, ..., V\}$, we have

$$1 - \frac{n_K}{n} = 1 - \frac{1}{V} - \frac{\eta_K}{n}, \text{ where } |\eta_K| \leq 1.$$

Hence,

$$\frac{1}{1 - n_K/n} = \frac{1}{1 - V^{-1} - n^{-1}\eta_K} = \frac{V}{V-1}\frac{1}{1 - \frac{V\eta_K}{n(V-1)}} = \frac{V}{V-1}\left(1 + r_K\right)$$

with $r_K := \frac{V\eta_K}{n(V-1) - V\eta_K}$. Since $V \geq 2$ and $n \geq 4$,

$$|r_K| \leq \frac{V}{n(V-1) - V} = \frac{1}{n\frac{(V-1)}{V} - 1} \leq \frac{1}{\frac{n}{2} - 1} \leq \frac{4}{n} \ .$$

Moreover, $r_K = 0$ when $n_K = n/V$. We have

$$\mathbb{E}(W_i) = \frac{1}{V}\sum_{K=1}^{V} \frac{1}{1 - n_K/n}\mathbf{1}_{i \notin \mathcal{B}_K} = 1 + \frac{1}{V-1}\sum_{\substack{K \neq K_i \\ 1 \leq K \leq V}} r_K \ .$$

When $K_i = K_j$, we have

$$\begin{aligned}\mathbb{E}(W_i W_j) &= \frac{1}{V}\sum_{K=1}^{V} \frac{1}{(1 - n_K/n)^2}\mathbf{1}_{i \notin \mathcal{B}_K}\mathbf{1}_{j \notin \mathcal{B}_K} \\ &= \frac{V}{(V-1)^2}\sum_{\substack{K \neq K_i \\ 1 \leq K \leq V}} \left(1 + r_K\right)^2 = 1 + \frac{1}{V-1} + \frac{V}{(V-1)^2}\sum_{\substack{K \neq K_i \\ 1 \leq K \leq V}} \left(2r_K + r_K^2\right) \ .\end{aligned}$$

When $K_i \neq K_j$, we have

$$\begin{aligned}\mathbb{E}(W_i W_j) &= \frac{1}{V}\sum_{K=1}^{V} \frac{1}{(1 - n_K/n)^2}\mathbf{1}_{i \notin \mathcal{B}_K}\mathbf{1}_{j \notin \mathcal{B}_K} \\ &= \frac{V}{(V-1)^2}\sum_{\substack{1 \leq K \leq V \\ K \neq K_i\,,\, K \neq K_j}} \left(1 + r_K\right)^2 = 1 - \frac{1}{(V-1)^2} + \frac{V}{(V-1)^2}\sum_{\substack{1 \leq K \leq V \\ K \neq K_i\,,\, K \neq K_j}} \left(2r_K + r_K^2\right) \ .\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}\left((W_i - 1)(W_j - 1)\right) = \mathbb{E}(W_i W_j) - \mathbb{E}(W_i) - \mathbb{E}(W_j) + 1 \\ &= \frac{1}{V-1} - \frac{V\mathbf{1}_{K_i \neq K_j}}{(V-1)^2} + \frac{1}{(V-1)^2}\sum_{\substack{1 \leq K \leq V \\ K \neq K_i\,,\, K \neq K_j}} \left(2r_K + Vr_K^2\right) - \frac{1}{V-1}\left(r_{K_i} + r_{K_j}\right)\mathbf{1}_{K_i \neq K_j} \ .\end{aligned}$$

Since $|r_K| \le 4n^{-1}$, $V|r_K| \le 4$, hence $|2r_K + Vr_K^2| \le 6|r_K|$. We also have

$$\text{if } K_i = K_j, \qquad \frac{1}{V-1}\sum_{K=1}^{V} 6\,|r_K| \le \frac{24}{n}\ .$$

$$\text{if } K_i \neq K_j, \qquad \frac{1}{V-1}\left[\sum_{\substack{1\le K\le V\\ K\neq K_i\,,\,K\neq K_j}} 6\,|r_K|\right] + |r_{K_i}| + \left|r_{K_j}\right| \le \frac{24(V-2)}{(V-1)n} + \frac{8}{n} \le \frac{32}{n}\ .$$

Hence, if

$$\delta(K_i,K_j) = n(V-1)\left[\frac{1}{(V-1)^2}\sum_{\substack{1\le K\le V\\ K\neq K_i\,,\,K\neq K_j}} \left(2r_K + Vr_K^2\right) - \frac{1}{V-1}\left(r_{K_i}+r_{K_j}\right)\mathbf{1}_{K_i\neq K_j}\right]\ ,$$

$|\delta(K_i,K_j)| \le 32$. Furthermore, if $n_K = n/V$ for every K, then $r_K = 0$ so that $\delta(K_i,K_j) = 0$. □

Proof of Lemma 23. From Eq. (37), we have

$$\mathrm{pen}_{\mathrm{VF}}(m) = \frac{2C}{n^2}\sum_{\lambda\in\Lambda_m}\sum_{1\le i,j\le n}\mathbb{E}\left[(W_i-1)(W_j-1)\right]\left(\psi_\lambda(X_i)-P\psi_\lambda\right)\left(\psi_\lambda(X_j)-P\psi_\lambda\right)\ .$$

Thanks to Lemma 24, we deduce that

$$\begin{aligned}
\frac{\mathrm{pen}_{\mathrm{VF}}(m)}{2C} &= \frac{1}{n^2(V-1)}\sum_{\lambda\in\Lambda_m}\sum_{1\le i,j\le n}\left(\psi_\lambda(X_i)-P\psi_\lambda\right)\left(\psi_\lambda(X_j)-P\psi_\lambda\right)\\
&- \frac{V}{(V-1)^2}\frac{1}{n^2}\sum_{\lambda\in\Lambda_m}\sum_{K\neq K'=1}^{V}\sum_{i\in\mathcal{B}_K,\,j\in\mathcal{B}_{K'}}\left(\psi_\lambda(X_i)-P\psi_\lambda\right)\left(\psi_\lambda(X_j)-P\psi_\lambda\right)\\
&+ \frac{1}{n^3(V-1)}\sum_{\lambda\in\Lambda_m}\sum_{1\le i,j\le n}\delta(K_i,K_j)\left(\psi_\lambda(X_i)-P\psi_\lambda\right)\left(\psi_\lambda(X_j)-P\psi_\lambda\right)\\
&= \frac{1}{V-1}\left\|\widehat{s}_m - s_m\right\|^2 - \frac{V}{(V-1)^2}\left(U_m - U_{\mathcal{B},m}\right)\\
&+ \frac{1}{n^3(V-1)}\sum_{\lambda\in\Lambda_m}\sum_{K,K'=1}^{V}\delta(K,K')\left(\sum_{i\in\mathcal{B}_K}\left(\psi_\lambda(X_i)-P\psi_\lambda\right)\right)\left(\sum_{j\in\mathcal{B}_{K'}}\left(\psi_\lambda(X_j)-P\psi_\lambda\right)\right).
\end{aligned}$$

□

I.5.2. *Concentration of V-fold penalties : the general case.* In this section, we extend Proposition 4 to partitions $\mathcal{B}$ satisfying (**H5**) instead of (**H5***).

Proposition 25. *Let $V \ge 2$, $n \ge 4$ and $\mathcal{B}$ satisfying* (**H5**). *For every $m \in \mathcal{M}_n$, let $\mathrm{pen}_{\mathrm{VF}}(m)$ be the V-fold penalty defined by Eq.* (4) *on a linear S_m satisfying Eq.* (**H1**) *and* (**H4**). *Let $2 \le x \le \frac{\sqrt{R_n^\star}}{C_\star^2} \wedge V^{1/6}\,(R_n^\star)^{1/4}$ and let*

$$\varepsilon_4(n,V,x) = C_\star\left\{\left(\frac{x^3}{n}\right) \vee \left(\frac{\sqrt{x}}{(R_n^\star)^{1/4}}\left[1 + \frac{x}{V^{1/3}} + \left(\frac{x^2\,(R_n^\star)^{1/4}}{\sqrt{n}}\right)\right]\right)\right\} \tag{96}$$

There exists an absolute constant L such that,

$$\mathbb{P}\left(\left|\frac{V-1}{C}\operatorname{pen}_{\mathrm{VF}}(m,\mathcal{B}) - 2\left\|\widehat{s}_m - s_m\right\|^2\right| > L\varepsilon_4(n,V,x)\frac{R_m}{n}\right) \leq 2e^{2-x} \ . \tag{97}$$

Proof of Proposition 25. Proposition 25 follows from Lemma 23 and concentration results for all the terms appearing in Lemma 23: Lemma 11 for U_m, Lemma 26 for $U_{\mathcal{B},m}$, and finally, for $R_{\operatorname{pen}_{\mathrm{VF}}}(m,\mathcal{B})$, Lemmas 27 and Eq. (35) imply that

$$\mathbb{P}\left(|R_{\operatorname{pen}_{\mathrm{VF}}}(m,\mathcal{B})| > L\frac{C_\star x^3}{n}\frac{R_m}{n}\right) \leq e^{2-x} \ .$$

□

Under (**H5**), Lemma 9 implies the following concentration inequality for $U_{\mathcal{B},m}$.

Lemma 26. *Let $\xi_{1:n}$ be i.i.d random variables and let S_m be a linear space satisfying Assumptions* (**H1**), (**H4**). *For all $m \in \mathcal{M}_n$, let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of S_m. Let $(\mathcal{B}_K)_{K=1,\dots,V}$ be a partition of $\{1,\dots,n\}$ satisfying* (**H5**). *Let*

$$\varepsilon_5(n,V,x) = \frac{x}{V^{1/3}} \vee \left\{\frac{x^2\,(R_n^\star)^{1/4}}{\sqrt{n}}\right\} \vee \frac{R_n^\star}{n}.$$

An absolute constant $L > 0$ exists such that, for all $m \in \mathcal{M}_n$ and $2 \leq x \leq C_\star^{-2}\,(R_n^\star)^{1/2} \wedge V^{1/6}\,(R_n^\star)^{1/4}$,

$$\mathbb{P}\left(|U_{\mathcal{B},m}| > L\varepsilon_5(n,V,x)\frac{C_\star x^{1/2}}{(R_n^\star)^{1/4}}\frac{R_m}{n}\right) \leq 2e^{-x} \ .$$

Proof of Lemma 26. Under the pseudo regularity assumption (**H5**), we have $n_K \leq n/V + 1$, hence

$$\sum_{K=1}^{V} n_K^2 \leq \left(\frac{n}{V}+1\right)\sum_{K=1}^{V} n_K = n\left(\frac{n}{V}+1\right) \ .$$

Therefore,

$$\frac{1}{n}\sqrt{\sum_{K=1}^{V} n_K^2} \leq \sqrt{\frac{1}{V}+\frac{1}{n}} \leq \sqrt{\frac{2}{V}} \ .$$

From Lemma 9 with $\nu = \sqrt{x}\,(R_n^\star)^{-1/4}$, $\mathcal{R}_m = C_\star R_m$, we deduce that there exists an absolute constant L such that, for all $\epsilon > 0$ satisfying $\epsilon\sqrt{x} \leq 1$,

$$P\left(|U_{\mathcal{B},m}| \leq C_\star L x\frac{R_m}{n}\left(\left(\epsilon+\frac{\nu^2}{\epsilon}\right)\frac{1}{\sqrt{V}} + \frac{\sqrt{V}}{n\epsilon^3}\right)\right) \geq 1 - 2e^{-x} \ .$$

Let $\epsilon = V^{1/6}\nu \wedge x^{-1/2}$, we have

$$\frac{\nu^2}{\epsilon\sqrt{V}} \vee \frac{\epsilon}{\sqrt{V}} \leq \frac{\nu}{V^{1/3}}, \qquad \frac{\sqrt{V}}{n\epsilon^3} \leq \frac{1}{n\nu^3} \vee \left\{\frac{x\,(R_n^\star)^{1/4}}{\sqrt{n}}\nu\right\} \leq \left(\frac{1}{x}\frac{R_n^\star}{n} \vee \left\{\frac{x\,(R_n^\star)^{1/4}}{\sqrt{n}}\right\}\right)\nu \ .$$

□

The concentration of the remainder term follows from the following lemma.

Lemma 27. *Let $\xi_1, ..., \xi_n$ be i.i.d. Let S_m be a linear space of function and let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of S_m. Let $(\mathcal{B}_K)_{K=1}^V$ be a partition of $\{1, ..., n\}$ and for all $K = 1, ..., V$, let $n_K = \mathrm{Card}(\mathcal{B}_K)$. Let $(\delta(K, K'))_{K,K'=1,...,V}$ be a family of real numbers, bounded by $\delta_\star$ and let*

$$Z_1 := \sum_{\lambda\in\Lambda_m} \sum_{K\neq K'=1}^{V} \delta(K, K') \left(\sum_{i\in\mathcal{B}_K} (\psi_\lambda(X_i) - P\psi_\lambda) \right) \left(\sum_{i\in\mathcal{B}_{K'}} (\psi_\lambda(X_i) - P\psi_\lambda) \right)$$

$$Z_2 := \sum_{\lambda\in\Lambda_m} \sum_{K=1}^{V} \delta(K) \left(\sum_{i\in\mathcal{B}_K} (\psi_\lambda(X_i) - P\psi_\lambda) \right)^2 .$$

Some absolute constants C_1, C_2 exist such that, for all $x > 0$, we have

$$\mathbb{P}\left(|Z_1| \leq C_1 n\delta_\star \left(D_m x + v_m^2 x^2 + \frac{V}{n} b_m^2 x^3 \right) \right) \geq 1 - e^{2-x} \tag{98}$$

$$\mathbb{P}\left(|Z_2| \leq C_2 n\delta_\star \left(D_m \sqrt{x} + v_m^2 x^{3/2} + \frac{V}{n} b_m^2 x^{5/2} \right) \right) \geq 1 - e^{2-x} . \tag{99}$$

Proof of Lemma 27. For all $\lambda \in \Lambda_m$, for all $K = 1, ..., V$, let

$$Z_{\lambda,K} = \sum_{i\in\mathcal{B}_K} (\psi_\lambda(X_i) - P\psi_\lambda) \quad \text{and} \quad Z_K = (Z_{\lambda,K})_{\lambda\in\Lambda_m} .$$

Proof of Eq. (98). The random variables $(Z_K)_{K=1,..,V}$ are independent. From [BBLM05, Theorem 2] (recalled by Lemma 32), for all $q \geq 2$, we have

$$\|Z_1\|_q \leq 2\sqrt{c}\sqrt{q} \sqrt{\left\| \sum_{K=1}^{V} \left(Z_1 - \mathbb{E}\left[Z_1 \,|\, (Z_{K'})_{K'\neq K} \right] \right)^2 \right\|_{q/2}} .$$

As the $Z_{\lambda,K}$ are centered, we deduce that

$$Z_1 - \mathbb{E}\left[Z_1 \,|\, (Z_{K'})_{K'\neq K} \right] = \sum_{\lambda\in\Lambda_m} \sum_{K'=1,\ K'\neq K}^{V} \delta(K, K') Z_{\lambda,K} Z_{\lambda,K'} .$$

Hence, from the triangular inequality and Cauchy-Schwarz inequality

$$\|Z_1\|_q \leq 2\sqrt{c}\sqrt{q} \sqrt{\sum_{K=1}^{V} \left\| \left(\sum_{K'=1,\ K'\neq K}^{V} |\delta(K, K')| \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 \right)^{1/2} \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right)^{1/2} \right)^2 \right\|_{q/2}} .$$

Using the inequality $2ab \leq \eta a^2 + \eta^{-1} b^2$ and the triangular inequality, we deduce that

$$\|Z_1\|_q \leq \sqrt{2cq} \sqrt{\sum_{K=1}^{V} \eta \left\| \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 \right)^2 \right\|_{q/2} + \eta^{-1} \left\| \left(\sum_{K'=1,\ K'\neq K}^{V} |\delta(K, K')| \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right)^{1/2} \right)^4 \right\|_{q/2}}$$

$$\leq \sqrt{2cq} \sqrt{\sum_{K=1}^{V} \eta \left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 \right\|_q^2 + \eta^{-1} \left\| \sum_{K'=1,\ K'\neq K}^{V} |\delta(K, K')| \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right)^{1/2} \right\|_{2q}^4} . \tag{100}$$

The random variables $|\delta(K,K')|\,(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2)^{1/2}$ being independent, we can use [BBLM05, Theorem 2] (see Lemma 33) to obtain

$$\left\| \sum_{K'=1,\ K'\neq K}^{V} |\delta(K,K')| \left(\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right)^{1/2} \right\|_{2q}^4 \leq 16c^2q^2 \left(\sum_{K'=1,\ K'\neq K}^{V} \delta(K,K')^2 \left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right\|_q \right)^2 .$$

Taking $\eta = 4cq \sum_{K'=1,\ K'\neq K}^{V} \delta(K,K')^2 \left\|\sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2\right\|_q / \left\|\sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2\right\|_q$ in Eq. (100), we deduce that

$$\|Z_1\|_q \leq 4cq \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2 \left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K'}^2 \right\|_q \left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 \right\|_q} . \tag{101}$$

By definition of $Z_{\lambda,K}^2$, we have

$$\sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 = n_K^2 \left(\sup_{t\in\mathbb{B}_m} (P_n^{(\mathcal{B}_K)} - P)t \right)^2 .$$

From [Ler11] (recalled by Proposition 28), we have then, for all $x>0$,

$$\mathbb{P}\left(\left| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 - n_K D_m \right| > c\left(\epsilon n_K D_m + n_K \frac{v_m^2 x}{\epsilon} + \frac{b_m^2 x^2}{\epsilon^3} \right) \right) \leq 2e^{-x} .$$

It comes by integration (see Lemma 31 for detailed computations) that

$$\left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 - n_K D_m \right\|_q \leq c\left(\epsilon n_K D_m + 2e^4 n_K \frac{v_m^2 q}{\epsilon} + \frac{b_m^2 q^2}{\epsilon^3} \right) \tag{102}$$

Plugging this inequality in Eq. (101), we obtain that there exists an absolute constant C such that

$$\begin{aligned} \|Z_1\|_q \leq{}& 4cqD_m \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2 n_K n_{K'}} \\ &+ C\left(\left(\epsilon D_m q + \frac{v_m^2 q^2}{\epsilon} \right) \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2 n_K n_{K'}} + \frac{b_m^2 q^3}{\epsilon^3} \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2} \right) . \end{aligned}$$

Using [Arl07, Lemma 8.10] (recalled by Lemma 30), we obtain that there exists an absolute constant C_1 such that, with probability $1-e^{2-x}$,

$$|Z_1| \leq C_1 \left(\left(D_m x + v_m^2 x^2 \right) \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2 n_K n_{K'}} + b_m^2 x^3 \sqrt{\sum_{K\neq K'=1}^{V} \delta(K,K')^2} \right) .$$

Proof of Eq. (99). By independence of the random variables $\delta(K)\sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2$, from [BBLM05, Theorem 2] (see Lemma 33), we have

$$\|Z_2 - \mathbb{E}(Z_2)\|_q \leq 2\sqrt{c}\sqrt{q} \sqrt{\sum_{K=1}^{V} \delta(K)^2 \left\| \sum_{\lambda\in\Lambda_m} Z_{\lambda,K}^2 \right\|_q^2}$$

From Eq. (102), we have

$$\left\| \sum_{\lambda \in \Lambda_m} Z_{\lambda,K}^2 \right\|_q \leq c \left(n_K \left(D_m + v_m^2 q \right) + b_m^2 q^2 \right).$$

Hence,

$$\|Z_2 - \mathbb{E}(Z_2)\|_q \leq C\sqrt{q} \left(\sqrt{\sum_{K=1}^V \delta(K)^2 n_K^2} \left(D_m + v_m^2 q \right) + b_m^2 q^2 \sqrt{\sum_{K=1}^V \delta(K)^2} \right)$$

Using [Arl07, Lemma 8.10] (recalled by Lemma 30), we obtain that an absolute constant $C_2 > 0$ exists such that, with probability $1 - e^{2-x}$,

$$|Z_2 - \mathbb{E}(Z_2)| \leq C_2\sqrt{x} \left(\sqrt{\sum_{K=1}^V \delta(K)^2 n_K^2} \left(D_m + v_m^2 x \right) + b_m^2 x^2 \sqrt{\sum_{K=1}^V \delta(K)^2} \right) .$$

We conclude the proof with the inequality $|\mathbb{E}(Z_2)| \leq \sum_{K=1}^V n_K |\delta(K)| D_m \leq \delta_\star D_m n$. □

I.6. **Probabilistic Tools.** This section recalls several probabilistic tools (most of them classical) that are used several times in the proofs. First, we recall a concentration inequality obtained in Theorem 4.1 in the supplementary material of [Ler11].

Proposition 28 ([Ler11])**.** *Let $\xi_{1:N}$ be iid random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common distribution P. Let S be a symmetric class of functions bounded by b. For all $t \in S$, let $P_N t = N^{-1} \sum_{i=1}^N t(\xi_i)$, $v^2 = \sup_{t \in S} P[(t - Pt)^2]$, $Z = \sup_{t \in S} (P_N - P)t$, $D = N\mathbb{E}(Z^2)$. For all $x > 0$ and all $\epsilon \in (0, 1]$, with probability larger than $1 - 2e^{-x}$,*

$$\left| Z^2 - \frac{D}{N} \right| \leq L \left(\epsilon \frac{D}{N} + \frac{1}{\epsilon} \left(\frac{v^2 x}{N} + \left(\frac{bx}{\epsilon N} \right)^2 \right) \right).$$

The constant $L = (16(\ln 2)^{-2} + 8)$ works. In particular, if $R > 0$ and $\eta > 0$ satisfy

$$D \leq R,\ v^2 \leq \eta^2 R,\ \frac{b^2}{N} \leq \eta^4 R,$$

taking $\epsilon = \eta\sqrt{x}$ in the previous inequality yields, for all x such that $\eta\sqrt{x} \leq 1$,

$$P\left(\left| Z^2 - \frac{D}{N} \right| \leq 3L\eta\sqrt{x} \frac{R}{N} \right) \geq 1 - 2e^{-x}.$$

In particular, Proposition 28 implies the following.

Corollary 29 (Corollary 4.3 in the supplementary material of [Ler11])**.** *Let $\xi_{1:N}$ be i.i.d random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(t_\lambda)_{\lambda \in \Lambda_m}$ be a set of functions in $L^2(\mu)$. Let*

$$B = \left\{ t = \sum_{\lambda \in \Lambda_m} a_\lambda t_\lambda,\ \sum_{\lambda \in \Lambda_m} a_\lambda^2 \leq 1 \right\},\ D = \mathbb{E}\left(\sup_{t \in B} (t(\xi_1) - Pt)^2 \right),$$

$$v^2 = \sup_{t \in B} P[(t - Pt)^2],\ b = \sup_{t \in B} \|t\|_\infty .$$

Let U be the following U-statistics

$$U = \frac{1}{N(N-1)} \sum_{i \neq j = 1}^N (t_\lambda(\xi_i) - Pt_\lambda)(t_\lambda(\xi_j) - Pt_\lambda).$$

For all $x > 0$ and all $\epsilon \in (0,1]$, with probability larger than $1 - 4e^{-x}$,

$$|U| \leq L' \left(\epsilon \frac{D}{N} + \frac{1}{\epsilon} \left(\frac{v^2 x}{N} + \left(\frac{bx}{\epsilon N} \right)^2 \right) \right).$$

The constant $L' = 2(L + 4(\ln 2)^{-1})$ works, where L is defined in Proposition 28. In particular, if $R > 0$ and $\eta > 0$ satisfy

$$D \leq R,\ v^2 \leq \eta^2 R,\ \frac{b^2}{p} \leq \eta^4 R,$$

taking $\epsilon = \eta\sqrt{x}$ in the previous inequality yields, for all x such that $\eta\sqrt{x} \leq 1$,

$$P\left(|U| \leq 3L'\eta\sqrt{x}\frac{R}{N} \right) \geq 1 - 4e^{-x}.$$

We now recall some lemmas proved in [Arl07], about the links between moment and concentration inequalities.

Lemma 30 (Lemma 8.10 in [Arl07])**.** *Let $\lambda_1, ..., \lambda_N \geq 0$, $\mu_1, ..., \mu_N > 0$ and ξ be a random variable such that for all $q \geq q_0 > 0$,*

$$\|\xi\|_q \leq \sum_{i=1}^N \lambda_i q^{\mu_i} \ .$$

Then, for every $y \geq 0$,

$$\mathbb{P}\left(|\xi| \geq \sum_{i=1}^N \left[\lambda_i \left(\frac{ey}{\min_j \mu_j} \right)^{\mu_i} \right] \right) \leq e^{q_0 \min_j\{\mu_j\}} e^{-y} \ .$$

We give a little generalization of Lemma 8.12 in [Arl07]

Lemma 31. *Let $a_1, ..., a_N \geq 0$, $\alpha_1, ..., \alpha_N > 0$, $b \geq 0$ and ξ be a random variable such that*

$$\mathbb{P}\left(|\xi| \geq \sup_{i=1,...,N} a_i y^{\alpha_i} \right) \leq b \exp(-y) \ .$$

Then, for every $q \geq \max(\alpha_i^{-1}) \vee 1$,

$$\|\xi\|_q \leq be^4 \sum_{i=1}^N a_i \left(\frac{\alpha_i}{e} \right)^{\alpha_i + 3/2} q^{\alpha_i} \ .$$

Proof of Lemma 31.

$$\begin{aligned} \|\xi\|_q^q &= \int_0^\infty \mathbb{P}(|\xi| > y^{1/q}) dy \leq b \int_0^\infty e^{-\min_{i=1,...,N}(y^{1/q}/a_i)^{1/\alpha_i}} dy \\ &\leq b \sum_{i=1}^N \int_0^\infty e^{-(y^{1/q}/a_i)^{1/\alpha_i}} dy = b \sum_{i=1}^N a_i^q q\alpha_i \int_0^\infty t^{q\alpha_i - 1} e^{-t} dt. \end{aligned}$$

Now, from the inequality

$$\forall \beta \geq 1,\ \int_0^\infty t^{\beta-1} e^{-t} dt \leq e \left(\frac{\beta}{e} \right)^\beta \sqrt{\beta} \ ,$$

since, for all i, $q\alpha_i \geq 1$, we deduce

$$\|\xi\|_q^q \leq be \sum_{i=1}^N a_i^q q\alpha_i \left(\frac{q\alpha_i}{e} \right)^{q\alpha_i} \sqrt{q\alpha_i}$$

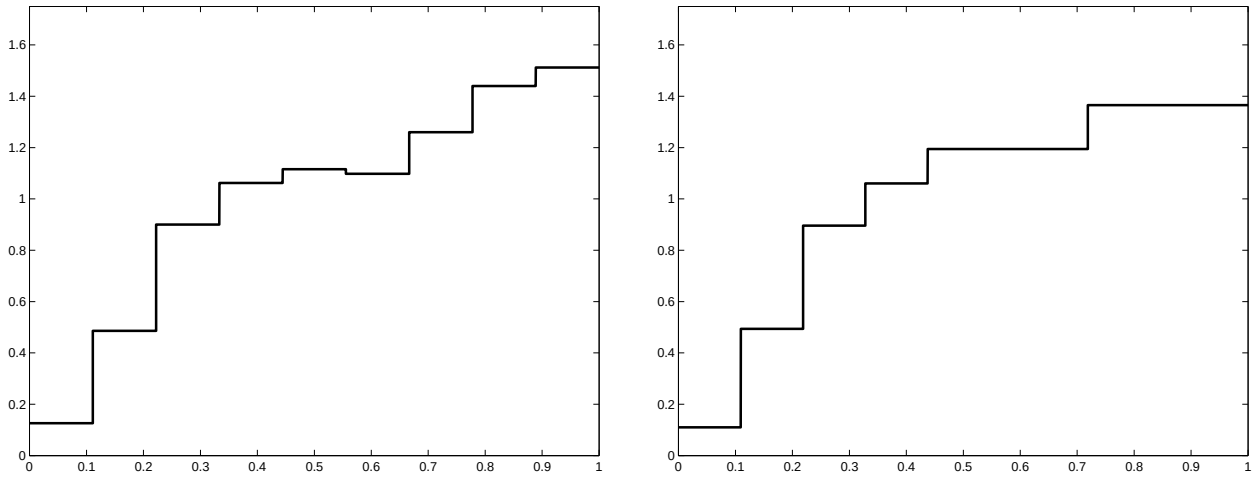

FIGURE 4. Oracle model for some sample generated according to L. Left: Regu. Right: Dya2.

Hence, since $q \geq 1$, and $q^{1/q} \leq e$,

$$\|\xi\|_q \leq (be)^{1/q} \sum_{i=1}^{N} a_i (q\alpha_i)^{3/(2q)} \left(\frac{q\alpha_i}{e}\right)^{\alpha_i} \leq be^4 \sum_{i=1}^{N} a_i \left(\frac{\alpha_i}{e}\right)^{\alpha_i + 3/2} q^{\alpha_i}.$$

□

Finally, we recall two moment inequalities that are corollaries of [BBLM05, Theorem 2] (see also Lemmas 8.17 and 8.18 in [Arl07], respectively).

Lemma 32 (Corollary of Theorem 2 of [BBLM05])**.** *Let $\xi_{1:N}$ be N independent random variables, let f be a measurable function $\mathbb{R}^N \to \mathbb{R}$ and*

$$Z = f(\xi_{1:N}) \ .$$

There exists $c \leq 1.271$ such that, for every $q \geq 2$,

$$\|Z - \mathbb{E}(Z)\|_q \leq 2\sqrt{c}\sqrt{q} \sqrt{\left\| \sum_{i=1}^{N} \left(Z - \mathbb{E}\left[Z | (\xi_j)_{j \neq i} \right] \right)^2 \right\|_{q/2}} \ .$$

Lemma 33 (Corollary of Theorem 2 of [BBLM05])**.** *Let $\xi_{1:N}$ be N independent random variables admitting q-th moments for some $q \geq 2$. Let $S = \sum_{i=1}^{N} \xi_i$. Then, there exists $c \leq 1.271$ such that*

$$\|S\|_q \leq 2\sqrt{c}\sqrt{q} \sqrt{\sum_{i=1}^{N} \|\xi_i\|_q^2} \ .$$

I.7. **Additional simulation results.** This section provides simulation results in addition to the ones of Section 5. First, Table 3 is an extended version of Table 2, with more procedures compared and two additional settings (L-Regu and S-Regu). Second, Figure 4 is an analogous of Figure 2 in setting L, that illustrates the difference between the model collections Regu and Dya2. Third, Figure 5 is an analogous of Figure 3 in setting L, that illustrates how the variance of V-fold criteria depends on V.

TABLE 3. Simulation results: settings L and S. The best procedures (up to standard-deviations) are bolded, where the data-driven procedures are considered separately from the procedures using the knowledge of $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$.

Experiment	L–Dya2	L–Regu	S–Dya2	S–Regu
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	6.62 ± 0.18	2.35 ± 0.05	2.09 ± 0.03	1.76 ± 0.02
$1.25 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	4.78 ± 0.12	2.04 ± 0.03	$\mathbf{1.95 \pm 0.02}$	$\mathbf{1.63 \pm 0.01}$
$1.5 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	4.13 ± 0.09	$\mathbf{1.92 \pm 0.02}$	$\mathbf{1.93 \pm 0.02}$	1.66 ± 0.01
$2 \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	$\mathbf{3.66 \pm 0.06}$	1.97 ± 0.02	2.02 ± 0.02	1.83 ± 0.01
C_p	8.52 ± 0.24	2.35 ± 0.05	3.26 ± 0.04	1.76 ± 0.02
$1.25 \times C_p$	6.10 ± 0.17	2.03 ± 0.03	3.04 ± 0.04	1.64 ± 0.01
$1.5 \times C_p$	4.97 ± 0.12	$\mathbf{1.92 \pm 0.02}$	3.01 ± 0.04	1.66 ± 0.01
$2 \times C_p$	4.38 ± 0.09	1.97 ± 0.02	3.18 ± 0.03	1.83 ± 0.01
$\mathrm{pen}_{\mathrm{LOO}}$	6.41 ± 0.18	2.35 ± 0.05	2.08 ± 0.03	1.76 ± 0.02
$1.25 \times \mathrm{pen}_{\mathrm{LOO}}$	4.65 ± 0.12	2.03 ± 0.03	$\mathbf{1.93 \pm 0.02}$	1.64 ± 0.01
$1.5 \times \mathrm{pen}_{\mathrm{LOO}}$	4.01 ± 0.09	$\mathbf{1.92 \pm 0.02}$	$\mathbf{1.91 \pm 0.02}$	1.66 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{LOO}}$	$\mathbf{3.61 \pm 0.06}$	1.97 ± 0.02	1.99 ± 0.02	1.83 ± 0.01
$\mathrm{pen}_{\mathrm{VF}}$ (V=10)	6.76 ± 0.17	2.44 ± 0.05	2.14 ± 0.03	1.78 ± 0.02
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	4.96 ± 0.12	2.05 ± 0.04	1.96 ± 0.02	$\mathbf{1.62 \pm 0.01}$
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	4.28 ± 0.10	$\mathbf{1.93 \pm 0.02}$	$\mathbf{1.91 \pm 0.02}$	1.64 ± 0.01
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=10)	$\mathbf{3.66 \pm 0.06}$	$\mathbf{1.92 \pm 0.02}$	1.95 ± 0.02	1.77 ± 0.01
$\mathrm{pen}_{\mathrm{VF}}$ (V=5)	7.53 ± 0.19	2.60 ± 0.06	2.21 ± 0.03	1.80 ± 0.02
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	5.50 ± 0.13	2.15 ± 0.04	2.00 ± 0.02	$\mathbf{1.63 \pm 0.01}$
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	4.65 ± 0.11	$\mathbf{1.96 \pm 0.03}$	1.95 ± 0.02	$\mathbf{1.61 \pm 0.01}$
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=5)	3.80 ± 0.07	$\mathbf{1.94 \pm 0.02}$	1.98 ± 0.02	1.72 ± 0.01
$\mathrm{pen}_{\mathrm{VF}}$ (V=2)	10.27 ± 0.24	3.22 ± 0.09	2.46 ± 0.03	2.02 ± 0.03
$1.25 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	7.77 ± 0.19	2.41 ± 0.05	2.23 ± 0.03	1.73 ± 0.02
$1.5 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	6.41 ± 0.16	2.18 ± 0.04	2.10 ± 0.02	$\mathbf{1.63 \pm 0.01}$
$2 \times \mathrm{pen}_{\mathrm{VF}}$ (V=2)	5.12 ± 0.12	$\mathbf{1.94 \pm 0.03}$	2.06 ± 0.02	1.64 ± 0.01
LOO	6.41 ± 0.18	2.35 ± 0.05	2.08 ± 0.03	1.76 ± 0.02
10-fold CV	6.25 ± 0.16	2.34 ± 0.05	2.07 ± 0.03	1.71 ± 0.02
5-fold CV	6.27 ± 0.16	2.28 ± 0.05	2.09 ± 0.03	1.68 ± 0.02
2-fold CV	6.41 ± 0.16	2.18 ± 0.04	2.10 ± 0.02	$\mathbf{1.63 \pm 0.01}$
Oracle: $10^{-3}\times$	5.49 ± 0.06	13.28 ± 0.16	43.91 ± 0.29	62.66 ± 0.39
Best: $10^{-3}\times$	19.82 ± 0.33	25.45 ± 0.28	83.70 ± 0.67	101.00 ± 0.76

CNRS ; SIERRA PROJECT-TEAM, LABORATOIRE D'INFORMATIQUE DE L'ECOLE NORMALE SUPÉRIEURE, (CNRS/ENS/INRIA UMR 8548), INRIA - 23 AVENUE D'ITALIE - CS 81321, 75214 PARIS CEDEX 13 - FRANCE

E-mail address: sylvain.arlot@ens.fr

LABORATOIRE J.A. DIEUDONNÉ, CNRS UMR 6621, UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS, 06108 NICE CEDEX 0 FRANCE

E-mail address: mlerasle@unice.fr

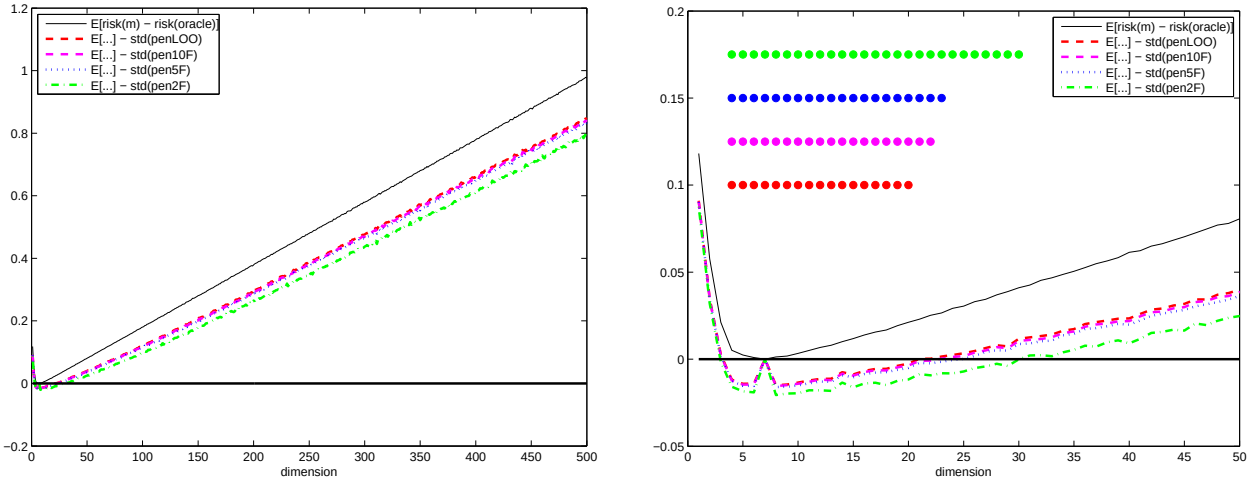

FIGURE 5. Visualization of variances as a function of V in experiment L-Regu. Right: zoom of the left part and "selectable models".