V-fold cross-validation improved: V-fold penalization

Sylvain Arlot

Submitted to the Annals of Statistics

V-FOLD CROSS-VALIDATION IMPROVED: V-FOLD PENALIZATION

By Sylvain Arlot

Université Paris-Sud

We study the efficiency of V-fold cross-validation (VFCV) for model selection from the non-asymptotic viewpoint, and suggest an improvement on it, which we call "V-fold penalization".

Considering a particular (though simple) regression problem, we prove that VFCV with a bounded V is suboptimal for model selection, because it "overpenalizes" all the more that V is large. Hence, asymptotic optimality requires V to go to infinity. However, when the signal-to-noise ratio is low, it appears that overpenalizing is necessary, so that the optimal V is not always the larger one, despite of the variability issue. This is confirmed by some simulated data.

In order to improve on the prediction performance of VFCV, we define a new model selection procedure, called "V-fold penalization" (penVF). It is a V-fold subsampling version of Efron's bootstrap penalties, so that it has the same computational cost as VFCV, while being more flexible. In a heteroscedastic regression framework, assuming the models to have a particular structure, we prove that penVF satisfies a non-asymptotic oracle inequality with a leading constant that tends to 1 when the sample size goes to infinity. In particular, this implies adaptivity to the smoothness of the regression function, even with a highly heteroscedastic noise. Moreover, it is easy to overpenalize with penVF, independently from the V parameter. A simulation study shows that this results in a significant improvement on VFCV in non-asymptotic situations.

1. Introduction. There are typically two kinds of model selection criteria. On the one-hand, penalized criteria are the sum of an empirical loss and some penalty term, often measuring the complexity of the models. This is the case of AIC (Akaike [Aka73]), Mallows' C_p or C_L (Mallows [Mal73]) and BIC (Schwarz [Sch78]), to name but a few. On the other hand, cross-validation (Allen [All74], Stone [Sto74], Geisser [Gei75]) and related criteria are based on the idea of data splitting. Part of the data (the training set) is used for fitting each model, and the rest of the data (the validation set) is used to measure the performance of the models. There are several versions of cross-validation (CV), *e.g.* leave-one-out (LOO, also called ordinary CV),

AMS 2000 subject classifications: Primary 62G09; secondary 62G08, 62M20

Keywords and phrases: non-parametric statistics, statistical learning, resampling, non-asymptotic, V-fold cross-validation, model selection, penalization, non-parametric regression, adaptivity, heteroscedastic data

leave-p-out (LPO, also called delete-p CV) and generalized CV (Craven and Wahba [CW79]). In practical applications, cross-validation is often computationally very expensive. This is why less greedy CV algorithms have been proposed, among which V-fold cross-validation (VFCV, Geisser [Gei75]) and repeated learning testing methods (Breiman *et al.* [BFOS84]). In this article, we mainly consider VFCV — which seems to be the most widely used nowadays — when the goal of model selection is to be *efficient*, *i.e.* to minimize the prediction risk among a family of estimators. Let us emphasize that this is quite different from picking up the "true model", which is often recalled as the *identification* or *consistency* issue.

The properties of CV (in particular leave-p-out) for prediction and model identification have been widely studied from the asymptotical viewpoint. It typically depends on the splitting ratio, *i.e.* the ratio between the sizes of the validation and training sets ($p/(n-p)$ in the leave-p-out case; $1/(V-1)$ for V-fold cross-validation). This has been shown for instance by Shao [Sha97] (for regression on linear models) and by van der Laan, Dudoit and Keles [vdLDK04] (for density estimation). Asymptotic optimality occurs when this ratio goes to zero at infinity, as shown by Li [Li87] for the leave-one-out, and generalized by Shao [Sha97] for the leave-p-out with $p \ll n$, both in the regression setting, when all the models are linear. Other asymptotic results about CV in regression can be found in the book by Györfi *et al.* [GKKW02], and in the paper of van der Laan, Dudoit and Keles [vdLDK04] for density estimation. Notice that the behaviour of these procedures changes completely when the goal is consistency; we refer to Yang [Yan07] and Sect. 5.4 below for references on this problem.

When it comes to practical application, a major question is how to choose the tuning parameters of CV procedures, since their performance strongly depend on them. In the case of VFCV, this means choosing V. Basically, there are three competing factors. First, the VFCV estimator of the prediction error, $\mathrm{crit}_{\mathrm{VFCV}}$, is biased, and its bias decreases with V. As shown by Burman [Bur89, Bur90], it is possible to correct this bias; otherwise, V should not be taken too small. Second, the variance of $\mathrm{crit}_{\mathrm{VFCV}}$ depends on V: it is always decreasing for small values of V, but then it can either stay decreasing (as in the linear regression case [Bur89]) or start to increase before $V = n$ (as in some classification problems [Bre96, HTF01, MSP05] or in density estimation [CR08]; see Sect. 2.3). Third, the computational cost of VFCV is proportional to V, so that the theoretic optimum (taking only bias and variability into account) can not always be computed. More precisely, it is necessary to understand well how the performance of VFCV depends on V before taking into account the computational cost. This is one of the purposes of this article.

We here aim at providing a better understanding of some CV procedures (including VFCV) from the *non-asymptotic viewpoint*. This may have two major implications. First, non-asymptotic results are made to handle collections of models which may depend on the sample size n: their sizes may typically be a power of n, and they may contain models whose complexities grow with n. Such collections of models are particularly significant for designing adaptive estimators of a function which is only assumed to belong to some hölderian ball, which may require an

arbitrarily large number of parameters. Second, in several practical applications, we are in a "non-asymptotic situation" in the sense that the signal-to-noise ratio is low. We shall see in the following that it should really be taken into account for an optimal tuning of V. It is worth noticing that such a non-asymptotic approach is not common in the literature, since most of the results already mentioned are asymptotic, and none is considering our second point above.

Another important point in our approach is that our framework includes several kinds of heteroscedastic data. We only assume that the observations $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d. with

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \ ,$$

where $s : \mathcal{X} \mapsto \mathbb{R}$ is the (unknown) regression function, $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the (unknown) noise-level, and ϵ_i has a zero mean and a unit variance conditionally to X_i. In particular, the noise-level $\sigma(X)$ can be strongly dependent from X, and the distribution of ϵ can itself depend from X. Such data are generally considered as very difficult to handle, because we have no information on σ, making irregularities of the signal harder to distinguish from noise. Then, simple model selection procedures such as Mallows' C_p may not work (see Chap. 4 of [Arl07] for a theoretical argument), and it is natural to hope that VFCV or other resampling methods may be robust to heteroscedasticity. In this article, both theoretical and simulation results confirm this fact.

In Sect. 2, we provide a non-asymptotic analysis of the performance of VFCV. The aforementioned bias turns out into a non-asymptotic negative result (Thm. 1), showing a rather simple problem for which VFCV can not satisfy an oracle inequality with leading constant smaller than $\kappa(V) - \epsilon_n$, with $\kappa(V) > 1$ for any $V \geq 2$ and $\epsilon_n \to 0$. In particular, VFCV with a bounded V can not be asymptotically optimal. But our analysis also has a major positive consequence in some "non-asymptotic" situations. Indeed, by considering VFCV as a penalization procedure, our previous result can be interpretated as an overpenalization property of VFCV. This should be related to the fact that the efficiency of penalization methods (like Mallows' C_p) is often improved by overpenalization, when the signal-to-noise ratio is small. Then, one can expect the optimal V for VFCV to be smaller than n, even for least-squares regression, which is confirmed by the simulation study of Sect. 4. So, it appears that choosing the optimal V for VFCV may be quite hard. In addition, the optimal choice may not be satisfactory when it corresponds to a highly variable criterion such as the 2-fold CV one. It is likely that there is some room left here to improve on VFCV.

This is why we propose in Sect. 3 another V-fold algorithm, that we call "V-fold penalization" (penVF). It is based upon Efron's resampling heuristics [Efr79], in the same way as Efron's bootstrap penalty [Efr83], but with a V-fold subsampling scheme instead of the bootstrap. It thus has exactly the same computational cost as the classical VFCV, and our results show that is has a similar robustness property, in some heteroscedastic regression framework. In addition, it turns out to be a generalization of Burman's corrected VFCV [Bur89, Bur90] (at least when the splitting into V blocks is regular). The main advance of penVF being that it is straightforward

to overpenalize within any factor when this is required, for instance when the signal-to-noise ratio seems low.

In the least-square regression framework, when we have to select among histogram models (see Sect. 2.2 for an accurate definition), we prove that penVF satisfies a non-asymptotic oracle inequality with a leading constant almost one (Thm. 2). To our knowledge, such a non-asymptotic result is new for any V-fold model selection procedure. One of its strengths is that it requires very few assumptions on the noise, allowing in particular heteroscedasticity. It is a strong result for penVF — which was not built for this particular setting at all — to improve on VFCV for such difficult problems, where VFCV is among the best procedures overall. As a consequence of Thm. 2, one can use penVF with the family of regular histograms in order to obtain an estimator adaptive to the smoothness of the regression function, when the noise is heteroscedastic (while having no information at all on the distribution of the noise). Notice that we only consider this result as a first step towards a more general theorem, without the restriction to histograms, as discussed in Sect. 5.3. The main interest of this toy framework is that we can study it deeply, and then derive general heuristics for practical use.

As an illustration to our theoretical study, we provide the results of a simulation study in Sect. 4. It confirms the good performances of penVF against both VFCV and the simpler Mallows' C_p criterion, in particular for difficult heteroscedastic problems. We also show how useful may be the flexibility of penVF when the signal-to-noise ratio is low. By decoupling V from the overpenalization factor, we allowed a significant improvement of the performance of both VFCV and its bias-corrected version.

Finally, our results are discussed in Sect. 5. The remaining of the paper is devoted to some probabilistic tools (App. A) and proofs (App. B).

2. Performance of V-fold cross-validation. In this section, we provide a non-asymptotic study of V-fold cross-validation (VFCV) in the least-squares regression framework. In order to make explicit computations possible, we focus on the case where each model is an "histogram model", *i.e.* the vector space of piecewise constant functions on some fixed partition of the feature space. This is only a first theoretical step. We use it to derive heuristics, that should help the practical user of VFCV in any framework. Notice also that we do not assume that the regression function itself is piecewise constant.

2.1. *General framework.* First consider the general prediction setting: $\mathcal{X} \times \mathcal{Y}$ is a measurable space, P an unknown probability measure on it and we observe some data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ of common law P. Let $\mathcal{S}$ be the set of predictors (measurable functions $\mathcal{X} \mapsto \mathcal{Y}$) and $\gamma : \mathcal{S} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$ a contrast function. Given a family $(\widehat{s}_m)_{m \in \mathcal{M}_n}$ of data-dependent predictors, our goal is to find the one minimizing the prediction loss $P\gamma(t) := \mathbb{E}_{(X,Y)\sim P}[\gamma(t, (X, Y))]$. Notice that the expectation here is only taken w.r.t. (X, Y), so that $P\gamma(t)$ is random when t is random (*e.g.* data-driven). Assuming that there exists a minimizer $s \in \mathcal{S}$ of the loss (the Bayes predictor), we will often consider the excess loss $l(s, t) = P\gamma(t) - P\gamma(s) \geq 0$ instead of the loss.

We assume that each predictor $\widehat{s}_m$ can be written as a function $\widehat{s}_m(P_n)$ of the empirical distribution of the data $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. The case-example of such a predictor is the empirical risk minimizer $\widehat{s}_m \in \arg\min_{t\in S_m}\{P_n\gamma(t)\}$, where S_m is any set of predictors (called a *model*). In the classical version of VFCV, we first choose some partition $(B_j)_{1\le j\le V}$ of the indexes $\{1,\dots,n\}$. Then, we define

$$P_n^{(j)} = \frac{1}{\mathrm{Card}(B_j)}\sum_{i\in B_j}\delta_{(X_i,Y_i)} \qquad \widehat{s}_m^{(j)} = \widehat{s}_m\left(P_n^{(j)}\right)$$
$$P_n^{(-j)} = \frac{1}{n-\mathrm{Card}(B_j)}\sum_{i\notin B_j}\delta_{(X_i,Y_i)} \qquad \widehat{s}_m^{(-j)} = \widehat{s}_m\left(P_n^{(-j)}\right) \ .$$

The final VFCV estimator is $\widehat{s}_{\widehat{m}_{\mathrm{VFCV}}}(P_n)$ with

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{m\in\mathcal{M}_n}\{\mathrm{crit}_{\mathrm{VFCV}}(m)\} \quad \text{and} \quad \mathrm{crit}_{\mathrm{VFCV}}(m) := \frac{1}{V}\sum_{j=1}^V P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) \ . \tag{1}$$

It is classical to assume that the partition $(B_j)_{1\le j\le V}$ is regular, *i.e.* that $\forall j$, $|\mathrm{Card}(B_j) - n/V| < 1$. In order to understand deeply the properties of VFCV, we have to compare precisely $\mathrm{crit}_{\mathrm{VFCV}}(m)$ to the excess loss $l(s,\widehat{s}_m)$. A crucial point is to compare their expectations, which is quite hard in general. This is why we restrict ourselves to a particular framework, namely the histogram regression one. We describe it in the next subsection.

2.2. *The histogram regression case.* In the regression framework, the data $(X_i,Y_i)\in\mathcal{X}\times\mathbb{R}$ are i.i.d. of common law P. Denoting by s the regression function, we have

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{2}$$

where $\sigma:\mathcal{X}\mapsto\mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with mean 0 and variance 1 conditionally to X_i. In order to simplify the theory, we will make two main assumptions on the data throughout this paper:

$$\sigma(X)\ge\sigma_{\min}>0 \quad \text{a.s.} \qquad \text{and} \qquad \|Y\|_\infty \le A < +\infty \ .$$

Notice that we do not assume $\sigma_{\min}$ and A to be known from the statistician. Moreover, those two assumptions can be relaxed, as shown by Chap. 6 and Sect. 8.3 of [Arl07]. The feature space $\mathcal{X}$ is typically a compact subset of $\mathbb{R}^d$. We use the least-squares contrast $\gamma:(t,(x,y))\mapsto(t(x)-y)^2$ to measure the quality of a predictor $t:\mathcal{X}\mapsto\mathcal{Y}$. As a consequence, the Bayes predictor is the regression function s, and the excess loss is $l(s,t) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2$. To each model S_m, we associate the *empirical risk minimizer*

$$\widehat{s}_m := \widehat{s}_m(P_n) = \arg\min_{t\in S_m}\{P_n\gamma(t)\}$$

(when it exists and is unique). Define also $s_m := \arg\min_{t \in S_m} P\gamma\,(t)$.

We now focus on histograms. Each model in $(S_m)_{m \in \mathcal{M}_n}$ is the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda \in \Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for any probability measure μ on $\mathcal{X}$, we can make explicit computations. The following notations will be useful throughout this article.

$$p_\lambda := P(X \in I_\lambda) \qquad \widehat{p}_\lambda := P_n(X \in I_\lambda) \qquad \sigma_\lambda^2 := \mathbb{E}\left[(Y - s(X))^2 \,\middle|\, X \in I_\lambda \right] \ .$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i, *i.e.* $\min_{\lambda \in \Lambda_m} \{\widehat{p}_\lambda\} > 0$. Prop. 1 below compares the V-fold criterion and the ideal criterion $P\gamma(\widehat{s}_m)$ in expectation.

Proposition 1. *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$ and $(B_j)\, 1 \leq j \leq V$ some "almost regular" partition of $\{1, \dots, n\}$,* i.e. *such that*

$$\max_j \left\{ \frac{\mathrm{Card}(B_j)}{n} \right\} \leq c_B < 1 \qquad \textit{and} \qquad \sup_j \left\{ \left| \frac{\mathrm{Card}(B_j)}{n} - \frac{1}{V} \right| \right\} \leq \epsilon_n^{reg} \xrightarrow[n \to \infty]{} 0 \ .$$

Then, the expectation of the ideal and V-fold criteria are respectively equal to

$$\mathbb{E}\left[P\gamma(\widehat{s}_m) \right] = P\gamma(s_m) + \frac{1}{n} \sum_{\lambda \in \Lambda_m} (1 + \delta_{n,p_\lambda})\, \sigma_\lambda^2 \tag{3}$$

$$\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m) \right] = P\gamma(s_m) + \frac{V}{V-1} \times \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(1 + \delta_{n,p_\lambda}^{(VF)} \right) \sigma_\lambda^2 \tag{4}$$

where $\delta_{n,p}$ only depends on (n,p), $\delta_{n,p}^{(VF)}$ depends on (n,p) and the partition $(B_j)_{1 \leq j \leq V}$, but both are small when the product np is large:

$$|\delta_{n,p}| \leq L_1 \qquad \textit{and} \qquad \left| \delta_{n,p}^{(VF)} \right| \leq L_2 \left[\epsilon_n^{reg} + \max\left((np)^{-1/4}, e^{-np(1-c_B)} \right) \right] \ ,$$

where L_1 is a numerical constant ant L_2 only depends on c_B.

Remark 1. Since we deal with histograms, $\widehat{s}_m$ is not defined when $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda = 0$, which occurs with positive probability. We then have to take a convention for $P\gamma\,(\widehat{s}_m)$ (on the event $\min_{\lambda \in \Lambda_m} \{\widehat{p}_\lambda\} = 0$, which has generally a very small probability) so that it has a finite expectation. The same kind of problem occur with $\mathrm{crit}_{\mathrm{VFCV}}$. See the proof of Prop. 1.

Prop. 1 is consistent with Burman's asymptotic estimate of the bias of VFCV [Bur89]. The major advance here is that it is non-asymptotic, and we have explicit upper bounds on the

remainder terms (see the proof of Prop. 1 in App. B.4). It shows that the classical V-fold cross-validation overestimates the variance term $n^{-1}\sum_{\lambda\in\Lambda_m}\sigma_\lambda^2$, because it estimates the generalization ability of $\widehat{s}_m^{(-j)}$, which is built upon less data than $\widehat{s}_m$. This interpretation is consistent with the results of Shao [Sha97] on linear regression, and van der Laan, Dudoit and Keles [vdLDK04] in the density estimation framework.

When V stays bounded as n grows to infinity, it is then natural to think that VFCV is underfitting, and thus be suboptimal for prediction. Since Prop. 1 is non-asymptotic and quite accurate, we are now in position to prove such a result.

THEOREM 1. *Let $n\in\mathbb{N}$, $(X_i,Y_i)_{1\le i\le n}$ be i.i.d. random variables, with $X\sim\mathcal{U}([0,1])$ and $Y=X+\sigma\epsilon$ with $\sigma>0$ and $\|\epsilon\|_\infty<+\infty$. Let $\mathcal{M}_n=\{1,\dots,n\}$ and $\forall m\in\mathcal{M}_n$, S_m be the model of regular histograms with $D_m=m$ pieces on $\mathcal{X}=[0,1]$. Let $V\in\{2,\dots,n\}$ and $(B_j)_{1\le j\le V}$ be some partition of $\{1,\dots,n\}$ such that for every j, $\left|\mathrm{Card}(B_j)-nV^{-1}\right|<1$.*

Then, there is an event of probability at least $1-K_1n^{-2}$ on which

$$l(s,\widehat{s}_{\widehat{m}_{\mathrm{VFCV}}})\ge(1+\kappa(V)-\ln(n)^{-1/5})\inf_{m\in\mathcal{M}_n}\{l(s,\widehat{s}_m)\}\quad, \tag{5}$$

for some constant $\kappa(V)>0$ depending only on V (and decreasing as a function of V), and a constant K_1 which depends on σ, A and V.

We now make a few comments:

- In the same framework, using similar arguments, we can prove an upper bound on $l(s,\widehat{s}_{\widehat{m}_{\mathrm{VFCV}}})$ showing that the constant $1+\kappa(V)$ is exact (up to the $\ln(n)^{-1/5}$ term). In particular,

$$\frac{l(s,\widehat{s}_{\widehat{m}_{\mathrm{VFCV}}})}{\inf_{m\in\mathcal{M}_n}\{l(s,\widehat{s}_m)\}}\xrightarrow[n\to+\infty]{a.s.}1+\kappa(V)=1+\frac{2^{2/3}}{3}\left[1-\left(\frac{V-1}{V}\right)^{1/3}\right]^2>1\ .$$

- When $(B_j)_{1\le j\le V}$ is not assumed regular, the proof of Prop. 1 shows that the factor $V/(V-1)$ becomes $\sum_{j=1}^V n/(n-\mathrm{Card}(B_j))$ which is always larger, because $x\mapsto(n-x)^{-1}$ is convex. On the other hand, if one chooses a $(X_i)_{1\le i\le n}$-dependent partition such that for every $\lambda\in\Lambda_m$, $\mathrm{Card}\{X_i\in I_\lambda\ \text{ and }\ i\in B_j\}$ is (almost) independent from j, then a similar proof shows that $\delta_{n,p}^{(VF)}$ is made much smaller than the previous upper bound. In a nutshell, it seems that the best performance of VFCV corresponds in general to the regular partition case, for which (5) holds.
- Although we restrict in Thm. 1 to a very particular problem, a similar result stays valid much more generally, possibly with a different value for the constant $\kappa(V)$. The only purpose of our assumptions is to compare very precisely $\mathrm{crit}_{\mathrm{VFCV}}(m)$ and $P\gamma(\widehat{s}_m)$ as functions of m. Since $D_{\widehat{m}_{\mathrm{VFCV}}}$ is smaller than the optimum from a multiplicative factor independent from n only, this analysis strongly depends on how $P\gamma(\widehat{s}_m)$ varies with m.

- One can easily extend this result to any cross-validation like method, when two conditions are satisfied. First, the ratio between the size of the training set and n has to be upper-bounded by $1 - V^{-1} < 1$ (uniformly in n). Second, the number of training sets considered has to be bounded by $B_{\max}$ (from which K_1 may depend). This includes for instance the hold-out case, and repeated learning-testing methods. Notice that the second assumption is mainly technical; if we were able to prove the corresponding concentration inequalities, the leave-p-out with $p \sim n/V$ should have approximately the same properties.

2.3. *How to choose V.*

2.3.1. *Classical analysis.* There are three well-known factors to take into account in order to choose V:

- *bias*: when V is too small, $\mathrm{crit}_{\mathrm{VFCV}}$ overestimates the variance term in $P\gamma(\widehat{s}_m)$, which leads to underfitting and suboptimal model selection (Thm. 1).
- *variability*: the variance of $\mathrm{crit}_{\mathrm{VFCV}}(m)$ is a decreasing function of V, at least in the linear regression framework (see Burman [Bur89] for an asymptotic expansion of this variance). In general, $V = 2$ is known to be quite variable because of the single split. When the prediction algorithm $(X_i, Y_i)_{1 \leq i \leq n} \mapsto \widehat{s}_m$ is unstable (*e.g.* classification with CART, as noticed by Hastie, Tibshirani and Friedman [HTF01]; see also Breiman [Bre96]), the leave-one-out criterion (*i.e.* $V = n$) is also known to be quite variable, but this phenomenon seems to disappear when $\widehat{s}_m$ is more stable (Molinaro, Simon and Pfeiffer [MSP05]). In particular, in the least-squares regression framework, the variance of $\mathrm{crit}_{\mathrm{VFCV}}(m)$ should decrease with V.
- *computational complexity*: V-fold cross-validation needs to compute at least V empirical risk minimizers for each model.

In the least-squares regression setting, V has to be chosen large in order to improve accuracy (by reducing bias and variability); on the contrary, computational issues arise when V is too big. This is why $V = 5$ and $V = 10$ are very classical and popular choices.

2.3.2. *The non-asymptotic need for overpenalization.* We now come to some particularity of the non-asymptotic viewpoint. Indeed, our proof of Thm. 1 shows that the asymptotic behaviour of hold-out and cross-validation criterions only depend on their bias, because all these criterions are sufficiently close to their expectations asymptotically. However, this is not true when the sample size is fixed, and even the less variable criterions are far from being deterministic. As a consequence, using an unbiased estimator is no longer a guarantee of being optimal, since it can still lead to choosing a very poor model with a positive probability.

In order to analyze this phenomenon, it is useful to take the penalization viewpoint. The idea of penalization for model selection is to define

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \right\} \quad , \tag{6}$$

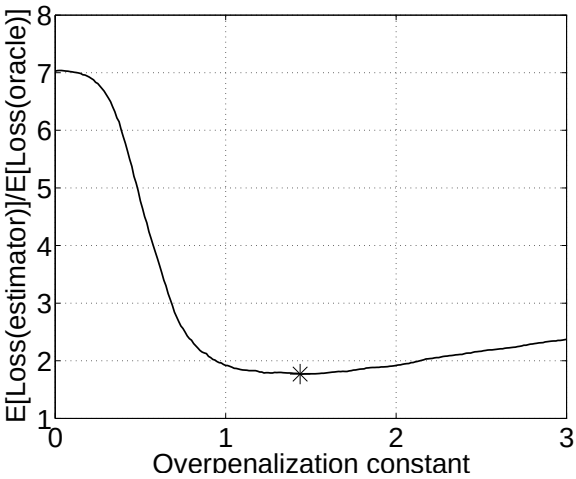

FIG 1. *The non-asymptotic need for overpenalization: the prediction performance C_{or} (defined in Sect. 4.1) of the model selection procedure* (6) *with* $\mathrm{pen}(m) = C_{\mathrm{ov}}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ *is represented as a function of* C_{ov}. *Data and models are the ones of experiment (S1):* $n = 200$, $\sigma \equiv 1$, $s(x) = \sin(\pi x)$. *See Sect. 4 for more details.*

where $\mathrm{pen}(m)$ is chosen so that $P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$ is close to the prediction error $P\gamma\left(\widehat{s}_m\right)$. In other words, the "ideal penalty" is

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma(\widehat{s}_m) \ . \tag{7}$$

According to Prop. 1 and (38) (which follows its proof), in the histogram regression case, we can compute the expectation of the ideal penalty:

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(2 + \delta_{n,p_\lambda}\right)\sigma_\lambda^2 \ , \tag{8}$$

which is close to Mallows' C_p penalty $2\sigma^2 D_m n^{-1}$ in the homoscedastic case. The point is that overpenalization (that is, taking pen larger than $\mathrm{pen}_{\mathrm{id}}$, even in expectation) can improve the prediction performance of $\widehat{s}_{\widehat{m}}$ when the signal-to-noise ratio is small. This can be seen on Fig. 1, according to which the optimal overpenalization constant $C_{\mathrm{ov}}^\star$ seems to be between 1.2 and 1.7 for this particular model selection problem. See also [Arl07] for a longer discussion of this problem.

2.3.3. *Choosing V in the non-asymptotic framework.* Since V-fold cross-validation is choosing the model $\widehat{m}_{\mathrm{VFCV}}$ which minimizes some criterion $\mathrm{crit}_{\mathrm{VFCV}}$, it can be written as a penalization procedure: it satisfies (6) with

$$\mathrm{pen}_{\mathrm{VFCV}}(m) := \mathrm{crit}_{\mathrm{VFCV}}(m) - P_n\gamma\left(\widehat{s}_m\right) \ .$$

Using again Prop. 1 and (38), we can compute its expectation:

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{VFCV}}(m)\right] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left[1 + \frac{V}{V-1} \left(1 + \delta_{n,p_\lambda}^{(VF)} \right) \right] \sigma_\lambda^2 .$$

Compared to (8), this shows that V-fold cross-validation is overpenalizing within a factor $1 + 1/(2(V-1))$.

We can now revisit the question of choosing V for optimal prediction, in such a non-asymptotic situation:

- the *overpenalization* factor is $1 + 1/(2(V-1))$.
- the *variance* of $\mathrm{crit}_{\mathrm{VFCV}}$ roughly decreases with V.
- the *computational complexity* of computing $\mathrm{crit}_{\mathrm{VFCV}}$ is roughly proportional to V.

First, take only the prediction performance into account. The variability question should be less crucial than overpenalization, because the variance of $\mathrm{crit}_{\mathrm{VFCV}}$ depends only on V through second order terms, according to the asymptotic computations of Burman [Bur89]. Since the optimal overpenalization constant is $C_{\mathrm{ov}}^\star > 1$, the performance of V-fold cross-validation should be optimal for some $V^\star < n$. This analysis is confirmed by the simulation study of Sect. 4, where $V = 2$ provides better performance than $V = 5$ and $V = 10$ for several different experiments.

Now, if computational cost comes into the balance, or if we consider less stable prediction algorithms than least-squares regression estimators, the optimal V may be even smaller. Whatever the framework, it seems quite difficult to find the optimal V, even if $C_{\mathrm{ov}}^\star$ was known (which is far from being the case in general). It would be at least necessary to understand well how the variance of $\mathrm{crit}_{\mathrm{VFCV}}$ depends on V in the non-asymptotic framework. This is a difficult practical problem, since "there is no universal (valid under all distributions) unbiased estimator of the variance of V-fold cross-validation" (Bengio and Grandvalet [BG04]). In the density estimation framework, this question has been tackled recently by Celisse and Robin [CR08].

The conclusion of this section is that choosing V for V-fold is a very complex issue in practice, even independently from the cost of computing $\mathrm{crit}_{\mathrm{VFCV}}$. Moreover, it seems unsatisfactory to select a model according to a criterion as variable as the 2-fold cross-validation one when $V^\star = 2$ because of the need for overpenalization. Finally, when the signal-to-noise ratio is large, we would like to obtain a nearly unbiased procedure without having to take V very large, which can be computationally too heavy.

In other words, we would like to decouple the choice of an overpenalization factor from the variability issue (which is essentially linked with complexity). The drawback of V-fold cross-validation is that they both depend on the V parameter. As we shall see in the next section, such a decoupling can be naturally obtained through the use of penalization.

3. An alternative V-fold algorithm: V-fold penalties. There are several ways to define V-fold cross-validation like penalization procedures with a tunable overpenalization factor,

independent from the V parameter. A first idea may be to multiply $\mathrm{pen}_{\mathrm{VFCV}}(m)$ by a constant *i.e.* to use (6) with the penalty

$$\mathrm{pen}(m) = C_{\mathrm{ov}} \left(1 + \frac{1}{2(V-1)}\right)^{-1} \left(\mathrm{crit}_{\mathrm{VFCV}}(m) - P_n \gamma\left(\widehat{s}_m\right)\right) \quad .$$

From the proof of Thm. 1 (see also the one of Thm. 2 below), it is clear that when $C_{\mathrm{ov}} \sim 1$, this procedure satisfies with large probability a non-asymptotic oracle inequality with leading constant $1 + \epsilon_n$, and more generally an oracle inequality with leading constant $K(C_{\mathrm{ov}}) \geq 1$. However, this may seem a little artificial, and strongly dependent from the histogram regression framework in which the computations of Prop. 1 work.

In this section, we consider another approach, that we call "V-fold penalization", which seems more natural to us. We shall see below that it is closely related to an idea of Burman [Bur89, Bur90] for correcting the bias of V-fold cross-validation. However, Burman did not consider his method as a penalization one. His goal was only to obtain an unbiased estimate of the prediction error, so that it is not straightforward to choose an overpenalization factor different from 1 with his method. This is a major difference with our approach.

3.1. *Definition of V-fold penalties.*

3.1.1. *General framework.* We come back to the general setting of Sect. 2.1. Recall that each predictor $\widehat{s}_m$ can be written as a function $\widehat{s}_m(P_n)$ of the empirical distribution of the data $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. We want to build a penalization method, *i.e.* choose $\widehat{m}$ according to (6), so that the prediction error of $\widehat{s}_{\widehat{m}}$ is as small as possible. This could be done exactly if we knew the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m(P_n))$, but this quantity depends on the unknown distribution P. Following a heuristics due to Efron [Efr79], we propose to define pen as the resampling estimate of $\mathrm{pen}_{\mathrm{id}}$, according to a V-fold subsampling scheme. We first recall the general form of this heuristics.

Basically, the *resampling heuristics* tells that one can mimic the relationship between P and P_n by building a n-sample of common distribution P_n (the "resample"). P_n^W denoting the empirical distribution of the resample, the pair (P, P_n) should be close (in distribution) to the pair (P_n, P_n^W) (conditionally to P_n for the latter distribution). Then, the expectation of any quantity of the form $F(P, P_n)$ can be estimated by $\mathbb{E}_W\left[F(P_n, P_n^W)\right]$, where $\mathbb{E}_W[\cdot]$ denotes expectation w.r.t. the resampling randomness. In the case of $\mathrm{pen}_{\mathrm{id}}$, this leads to Efron's bootstrap penalty [Efr83]. Later on, this heuristics has been generalized to other resampling schemes, with the exchangeable weighted bootstrap (Mason and Newton [MN92], Præstgaard and Wellner [PW93]). The empirical distribution of the resample then has the general form

$$P_n^W := \frac{1}{n}\sum_{i=1}^n W_i \delta_{(X_i,Y_i)} \qquad \text{with} \quad W \in \mathbb{R}^n \quad \text{an exchangeable weight vector,}$$

independent from the data (W is said to be *exchangeable* when its distribution is invariant by any permutation of its coordinates). Fromont [Fro07] used it successfully (with a particular

upper bound on $\mathrm{pen}_{\mathrm{id}}$) to build global penalties in the classification framework. Exchangeable resampling penalties (generalizing Efron's bootstrap penalty) have also been recently proposed, and studied in the regression framework [Arl07]. The idea of V-fold penalties is to use a V-fold subsampling scheme instead, *i.e.* take $W_i = \frac{V}{V-1}\mathbb{1}_{i \notin B_J}$ with $J \sim \mathcal{U}(\{1,\ldots,V\})$ independent from the data ($\mathcal{U}(E)$ denotes the uniform distribution over the set E). Then, $P_n^W = P_n^{(-J)}$ and we obtain the following algorithm.

Algorithm 1 (V-fold penalization).

1. Choose a partition $(B_j)_{1 \le j \le V}$ of $\{1,\ldots,n\}$, as regular as possible.
2. Choose a constant $C \ge C_{W,\infty} = V-1$.
3. Compute the following resampling penalty for each $m \in \mathcal{M}_n$:

$$\mathrm{pen}(m) = \mathrm{pen}_{\mathrm{VF}}(m) := \frac{C}{V}\sum_{j=1}^{V}\left[P_n\gamma\left(\widehat{s}_m\left(P_n^{(-j)}\right)\right) - P_n^{(-j)}\gamma\left(\widehat{s}_m\left(P_n^{(-j)}\right)\right)\right] \ .$$

4. Choose $\widehat{m}$ according to (6).

Remark 2 (About the constant C). Contrary to Efron's resampling heuristics, we have to put a constant $C \neq 1$ in front of the penalty (pen being an unbiased estimator of $\mathrm{pen}_{\mathrm{id}}$ when $C = C_{W,\infty}$). This is because each W_i has a variance $(V-1)^{-1} \neq 1$ (we only normalized W so that $\mathbb{E}[W_i] = 1$ for every i). According to Lemma 8.4 of [Arl07], the right normalizing constant can be derived from the exchangeable case. As a consequence, from Theorem 3.6.13 in [vdVW96],

$$C_{W,\infty} \sim_{n\to\infty} \left(n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(W_i - 1\right)^2\right)^{-1} \sim_{n\to\infty} V-1 \ .$$

The asymptotic value of $C_{W,\infty}$ can also be derived from the computations of Burman [Bur89] in the linear regression framework. Indeed, with our notations, Burman's criterion (formula (2.3) in [Bur89]) is

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

If all the blocks of the partition have the same size n/V, then $P_n^{(j)} - P_n = (V-1)(P_n - P_n^{(-j)})$, so that Burman's corrected VFCV coincides exactly with V-fold penalization when $C = V-1$. Since $\mathrm{crit}_{\mathrm{corr.VF}}(m)$ is an asymptotically unbiased estimator of $P\gamma\left(\widehat{s}_m\right)$ (at least for linear regression), the result follows. From the non-asymptotic viewpoint, we prove in Sect. 3.2 below that $V-1$ also leads to an unbiased estimator of $\mathrm{pen}_{\mathrm{id}}$ in the histogram regression case.

Notice also that we do not assume that $C = C_{W,\infty}$, but only $C \geq C_{W,\infty}$. This is a major quality of V-fold penalization (penVF): it is straightforward to choose any overpenalization factor, independently from V. Further comments about the choice of C and V are made in Sect. 5.

3.1.2. *The histogram regression case.* We now come back to the framework of Sect. 2.2, in which we can analyze deeper Algorithm 1. Remind that histograms are not our final goal, but only a convenient setting from which we can derive heuristics for practical use of penVF in any framework. From now on, $(S_m)_{m\in\mathcal{M}_n}$ is a collection of histogram models and $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ the associated collection of least-squares estimators. We first introduce some more notations:

$$s_m = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \quad \text{and} \quad \widehat{s}_m = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \quad \text{with} \quad \beta_\lambda = \mathbb{E}\left[Y \mid X \in I_\lambda\right] \quad \text{and} \quad \widehat{\beta}_\lambda = \frac{1}{n\widehat{p}_\lambda} \sum_{X_i\in I_\lambda} Y_i$$

$$\widehat{p}_\lambda^W := P_n^W(X \in I_\lambda) = \widehat{p}_\lambda W_\lambda \quad \text{with} \quad W_\lambda := \frac{1}{n\widehat{p}_\lambda} \sum_{X_i\in I_\lambda} W_i$$

$$\text{and} \qquad \widehat{s}_m^W := \arg\min_{t\in S_m} P_n^W \gamma(t) = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda^W \mathbb{1}_{I_\lambda} \quad \text{with} \quad \widehat{\beta}_\lambda^W := \frac{1}{n\widehat{p}_\lambda^W} \sum_{X_i\in I_\lambda} W_i Y_i \ .$$

Assuming that $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda > 0$ (otherwise, the model m should clearly not be chosen), we can compute the ideal penalty (see (37) and (38) in Sect. B.4) and its resampling estimate:

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) = \sum_{\lambda\in\Lambda_m} (p_\lambda + \widehat{p}_\lambda)\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 + (P - P_n)\gamma(s_m)$$

$$\mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m^W)\right] = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\left(\widehat{p}_\lambda + \widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] \ , \tag{9}$$

since $\sum_i \mathbb{E}[W_i] = 1$ implies that $\mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right] = 0$. The penalty (9) is well-defined if and only if $\widehat{s}_m^W$ is a.s. uniquely defined, *i.e.* $W_\lambda > 0$ for every $\lambda \in \Lambda_m$ a.s. This is why we modified the definition of the weights in algorithm 1, so that this problem does not occur.

Algorithm 2 (V-fold penalization for histograms).

1. Replace $\mathcal{M}_n$ by $\widehat{\mathcal{M}}_n = \{m \in \mathcal{M}_n \text{ s.t. } \min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq 3\}$.
2. Choose a constant $C \geq C_{W,\infty} = V - 1$.
3. For every $m \in \widehat{\mathcal{M}}_n$, choose a partition $(B_j)_{1\leq j\leq V}$ of $\{1,\dots,n\}$ such that
$$\forall \lambda \in \Lambda_m, \forall 1 \leq j \leq V, \quad \left|\mathrm{Card}\left(B_j \cap \{i \text{ s.t. } X_i \in I_\lambda\}\right) - \frac{n\widehat{p}_\lambda}{V}\right| < 1 \ .$$
4. Compute the following resampling penalty for each $m \in \mathcal{M}_n$:
$$\mathrm{pen}(m) = \mathrm{pen}_{\mathrm{VF}}(m) := \frac{C}{V}\sum_{j=1}^{V}\left[P_n\gamma\left(\widehat{s}_m^{(-j)}\right) - P_n^{(-j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \ . \tag{10}$$

5. Choose $\widehat{m}$ according to (6).

At step 3, we choose a different partition for each model m. Our choice is consistent with the proposal of Breiman *et al.* [BFOS84] (see also Burman [Bur90], Sect. 2) to stratify the data and choose a partition which respects the stratas. In the histogram case, natural stratas are the sets $\{i \text{ s.t. } X_i \in I_\lambda\}$. In particular, steps 1 and 3 of Algorithm 2 ensure that $\min_{\lambda\in\Lambda_m} W_\lambda > 0$ for every $m \in \widehat{\mathcal{M}}_n$, so that (10) is well-defined.

Other modifications of algorithm 1 are possible. For instance, keep the same regular partition $(B_j)_{1\le j\le V}$ for all the models, and take

$$\mathrm{pen}_{\mathrm{VF}}(m) = C \sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \middle| \, W_\lambda > 0 \right] + \mathbb{E}_W \left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right] \right) \tag{11}$$

instead of (9). This is what we did in the simulations of Sect. 4, and a short theoretical study of this method is done in Sect. 8.4.1 of [Arl07]. It confirms that the two algorithms should have very similar performances in practical applications.

3.2. *Expectations.* We now come to the expectation of V-fold penalties, in the histogram regression framework.

Proposition 2. *Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ and* $\mathrm{pen} = \mathrm{pen}_{\mathrm{VF}}$ *be defined as in Algorithm 2. Then, if* $\min_{\lambda\in\Lambda_m} \{ n\widehat{p}_\lambda \} \ge 3$,

$$\mathbb{E}^{\Lambda_m} \left[\mathrm{pen}_{\mathrm{VF}}(m) \right] = \frac{1}{n} \sum_{\lambda\in\Lambda_m} \left(\frac{2C}{V-1} + \frac{C}{V-1} \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})} \right) \sigma_\lambda^2 \tag{12}$$

with $\mathbb{E}^{\Lambda_m}[\cdot] = \mathbb{E}^{\Lambda_m}\left[\cdot \,\middle|\, (\mathbb{1}_{X_i\in I_\lambda})_{1\le i\le n,\, \lambda\in\Lambda_m} \right]$ *and* $\frac{2}{n\widehat{p}_\lambda - 2} \ge \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})} \ge 0$.

Comparing (12) with (8), it appears that $\mathrm{pen}_{\mathrm{VF}}$ is an (almost) unbiased estimator of $\mathrm{pen}_{\mathrm{id}}$ when $C = V - 1$. Indeed, when $\min_{\lambda\in\Lambda_m} \{ np_\lambda \}$ goes to infinity faster than some constant times $\ln(n)$, so does $\min_{\lambda\in\Lambda_m} \{ n\widehat{p}_\lambda \}$ with a large probability. Moreover, following the proof of Lemma 3, we can show that

$$\mathbb{E}\left[\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penV})} \mathbb{1}_{n\widehat{p}_\lambda \ge 3} \right] \le \kappa \min\left(1, (np_\lambda)^{-1/4} \right) \xrightarrow[np_\lambda\to\infty]{} 0$$

for some absolute constant $\kappa > 0$. This is consistent with the asymptotic computations of Burman [Bur89]. The main novelty of Prop. 2 is that we have an explicit non-asymptotic upperbound on the remainder term. This is crucial to derive oracle inequalities for Algorithm 2.

3.3. *Oracle inequalities and asymptotic optimality.* We are now in position to state the main result of this section: V-fold penalties (Algorithm 2) satisfy a non-asymptotic oracle inequality with a leading constant close to 1, on a large probability event. This implies the asymptotic optimality of Algorithm 2 in terms of excess loss. For this, we assume the existence of some non-negative constants $\alpha_{\mathcal{M}}$, $c_{\mathcal{M}}$, c_{rich}, η such that:

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\text{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.
(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}; c_{\text{rich}}\sqrt{n}]$.
(**P3**) The constant C is well chosen: $\eta(V-1) \geq C \geq V-1$.

THEOREM 2. *Assume that the (X_i, Y_i)'s satisfy the following:*

(**Ab**) *Bounded data:* $\|Y_i\|_\infty \leq A < \infty$.
(**An**) *Noise-level bounded from below:* $\sigma(X_i) \geq \sigma_{\min} > 0$ *a.s.*
(**Ap**) *Polynomial decreasing of the bias: there exists* $\beta_1 \geq \beta_2 > 0$ *and* $C_{\text{b}}^+, C_{\text{b}}^- > 0$ *such that*

$$C_{\text{b}}^- D_m^{-\beta_1} \leq l(s, s_m) \leq C_{\text{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{L}(X)$*:* $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\text{r},\ell}^X > 0$.

Let $\widehat{m}$ be the model chosen by algorithm 2 (under restrictions (**P1** − **3**)*, with $\eta = 1$). Then, there exists a constant K_2 and a sequence ϵ_n converging to zero at infinity such that*

$$l(s, \widehat{s}_{\widehat{m}}) \leq (1 + \epsilon_n) \inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) \} \tag{13}$$

with probability at least $1 - K_2 n^{-2}$. Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right] \leq (1 + \epsilon_n)\, \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \{ l(s, \widehat{s}_m) \} \right] + \frac{A^2 K_2}{n^2} \quad . \tag{14}$$

The constant K_2 may depend on V and constants in (**Ab**)*,* (**An**)*,* (**Ap**)*,* ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *and* (**P1** − **3**)*, but not on n. The term ϵ_n is smaller than $\ln(n)^{-1/5}$ for instance; it can also be taken smaller than $n^{-\delta}$ for any $0 < \delta < \delta_0(\beta_1, \beta_2)$, at the price of enlarging K_2.*

We first make a few comments on our assumptions.

1. When assumption (**P3**) is satisfied with $\eta > 1$, the same result holds with a leading constant $2\eta - 1 + \epsilon_n$ instead of $1 + \epsilon_n$ in (13) and (14).
2. In Thm. 2, we assume that V is fixed when n grows. A careful look at the proof shows that we only need $V \leq \ln(n)$ for n large enough. With a few more work, we could go up to V of order n^δ for some $\delta > 0$ depending on the assumptions of Thm. 2, but we can not handle the leave-one-out case ($V = n$). This is probably a technical restriction, since a similar result for several exchangeable weights (including leave-one-out) is proven in Chap. 6 of [Arl07].

3. (**Ab**) and (**An**) are rather mild (and neither A nor $\sigma_{\min}$ need to be known from the statistician). In particular, they allow quite general heteroscedastic noises. They can even be relaxed, for instance thanks to results proven in Chap. 6 and Sect. 8.3 of [Arl07], allowing the noise to vanish or to be unbounded.
4. ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied for "almost regular" histograms when X has a lower bounded density w.r.t. Leb, as for instance all the simulation experiments of Sect. 4.
5. The upper bound in (**Ap**) holds when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and s α-hölderian with $\alpha \in (0,1]$. The lower bound may seem more surprising, since it means that s is not too well approximated by the models S_m. However, it is classical to assume that $l(s,s_m) > 0$ for every $m \in \mathcal{M}_n$ for proving the asymptotic optimality of Mallows' C_p (*e.g.* by Shibata [Shi81], Li [Li87] and Birgé and Massart [BM06]). We here make a stronger assumption because we need a non-asymptotic lower bound on the dimension of both the oracle and selected models. The reason why it is not too restrictive is that non-constant α-hölderian functions satisfy (**Ap**) with

$$\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha^{-1} \qquad \text{and} \qquad \beta_2 = 2\alpha k^{-1} \ ,$$

when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$ (*cf.* Sect. 8.10 in [Arl07] for more details). Notice also that Stone [Sto85] and Burman [Bur02] used the same assumption in the density estimation framework.

Theorem 2 has at least two major consequences. First, V-fold penalties provide an *asymptotically optimal* model selection procedure, at least in the histogram regression framework, as soon as $C \sim V-1$. This should be compared to Thm. 1, where we proved that V-fold cross-validation is suboptimal for a rather mild homoscedastic problem. Notice that a slight modification of the proof of Thm. 2 shows that several other cross-validation like methods (even with the same computational cost) have similar theoretical properties. We discuss this point in Sect. 5.

Second, Thm. 2 can handle several kinds of *heteroscedastic noises*, while Algorithm 2 does not need any knowledge about σ, $\|Y\|_\infty$ or the smoothness of s. Even the tuning of C and V can be made (at least at first order) without any information on the distribution P of the data. This shows that V-fold penalization is a *naturally adaptive algorithm*, as long as $\mathcal{M}_n$ allows adaptation. The point here is that when s belongs to some hölderian ball $\mathcal{H}(\alpha,R)$ (with $\alpha \in (0,1]$ and $R > 0$), we can choose $\mathcal{M}_n$ as the family of regular histograms on $\mathcal{X} \subset \mathbb{R}^k$ to obtain such an adaptivity result. Then, from Thm. 2, we can build an estimator adaptive to (α,R) in a heteroscedastic framework (see [Arl07] for more details). If moreover the noise-level σ satisfies some regularity assumption, we can show that this estimator attains the minimax estimation rate, up to some numerical constant, when $\alpha = k = 1$.

Notice also that a similar adaptation result could be obtained with V-fold cross-validation, which also satisfies (13) and (14) with leading constants $K(V) > 1$, under similar assumptions. The advance with V-fold penalization is that we have simultaneously the adaptivity property of

V-fold cross-validation, its mild computational cost (when V is chosen small), and asymptotic optimality (contrary to VFCV).

Finally, we would like to emphasize that building such estimators is not the final goal of penVF. As a matter of fact, there are several procedures that are adaptive to the smoothness of s and the heteroscedasticity of the noise (*e.g.* by Efromovich and Pinsker [EP96] or Galtchouk and Pergamenshchikov [GP05]), and they may have better performances than both VFCV and penVF in this particular framework. Contrary to these *ad hoc* procedures, particulary built for dealing with heteroscedasticity, VFCV and penVF are general-purpose devices. What our theoretical results show is that they behave quite well in this framework, for which they were not built in particular.

4. Simulation study. As an illustration of the results of the two previous sections, we compare the performances of VFCV, penVF (for several values of V) and Mallows' C_p on some simulated data.

4.1. *Experimental setup.* We consider four experiments, called S1, S2, HSd1 and HSd2. Data are generated according to

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i$$

with X_i i.i.d. uniform on $\mathcal{X} = [0;1]$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independent from X_i. The experiments differ from the regression function s (smooth for S, see Fig. 2; smooth with jumps for HS, see Fig. 3), the noise type (homoscedastic for S1 and HSd1, heteroscedastic for S2 and HSd2) and the number n of data. Instances of data sets are given by Fig. 4 to 7. Their last difference lies in the families of models. Defining

$$\forall k, k_1, k_2 \in \mathbb{N}\backslash\{0\}, \quad (I_\lambda)_{\lambda\in\Lambda_k} = \left(\left[\frac{j}{k};\frac{j+1}{k}\right)\right)_{0\le j\le k-1} \text{ and}$$
$$(I_\lambda)_{\lambda\in\Lambda_{(k_1,k_2)}} = \left(\left[\frac{j}{2k_1};\frac{j+1}{2k_1}\right)\right)_{0\le j\le k_1-1} \cup \left(\left[\frac{1}{2}+\frac{j}{2k_2};\frac{1}{2}+\frac{j+1}{2k_2}\right)\right)_{0\le j\le k_2-1} ,$$

the four model families are indexed by $m \in \mathcal{M}_n \subset (\mathbb{N}\backslash\{0\}) \cup (\mathbb{N}\backslash\{0\})^2$:

S1 regular histograms with $1 \le D \le n(\ln(n))^{-1}$ pieces, *i.e.*

$$\mathcal{M}_n = \left\{1, \dots, \left\lfloor \frac{n}{\ln(n)} \right\rfloor\right\} .$$

S2 histograms regular on $[0;1/2]$ (resp. on $[1/2;1]$), with D_1 (resp. D_2) pieces, $1 \le D_1, D_2 \le n(2\ln(n))^{-1}$. The model of constant functions is added to $\mathcal{M}_n$, *i.e.*

$$\mathcal{M}_n = \{1\} \cup \left\{1, \dots, \left\lfloor \frac{n}{2\ln(n)} \right\rfloor\right\}^2 .$$

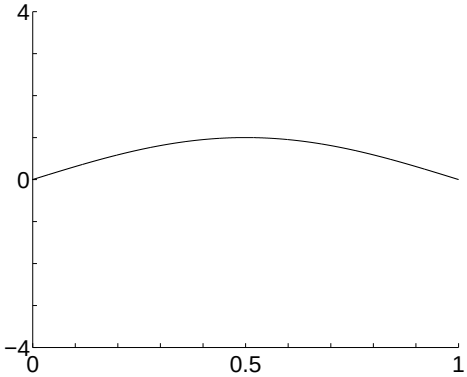

FIG 2. $s(x) = \sin(\pi x)$

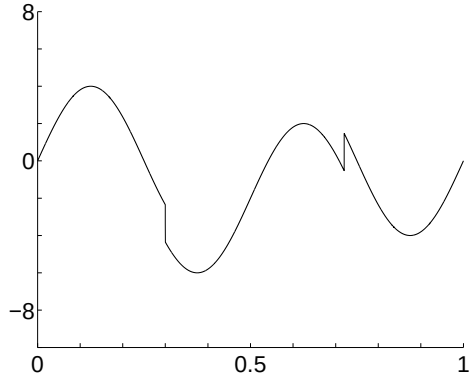

FIG 3. $s(x) = \text{HeaviSine}(x)$ *(see [DJ95])*

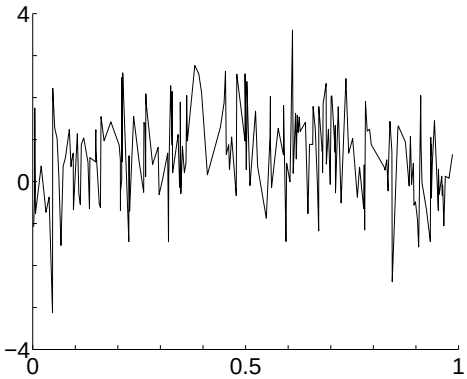

FIG 4. *S1:* $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$

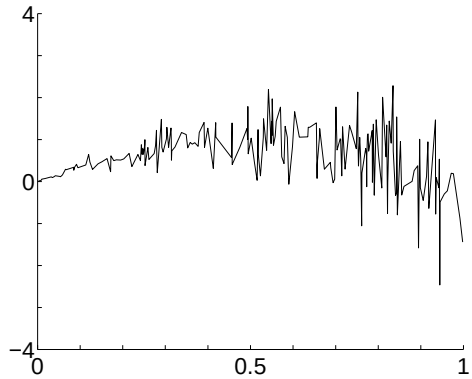

FIG 5. *S2:* $s(x) = \sin(\pi x)$, $\sigma(x) = x$, $n = 200$

HSd1 dyadic regular histograms with 2^k pieces, $0 \leq k \leq \ln_2(n) - 1$, *i.e.*

$$\mathcal{M}_n = \left\{ 2^k \text{ s.t. } 0 \leq k \leq \ln_2(n) - 1 \right\} \quad .$$

HSd2 dyadic regular histograms with bin sizes 2^{-k_1} and 2^{-k_2}, $0 \leq k_1, k_2 \leq \ln_2(n) - 2$ (dyadic version of S2). The model of constant functions is added to $\mathcal{M}_n$, *i.e.*

$$\mathcal{M}_n = \{1\} \cup \left\{ 2^k \text{ s.t. } 0 \leq k \leq \ln_2(n) - 2 \right\}^2 \quad .$$

Notice that we choose models that can approximately fit the true shape of $\sigma(x)$ in experiments S2 and HSd2. This choice makes the oracle model even more efficient, hence the model selection problem more challenging.

We compare the following algorithms:

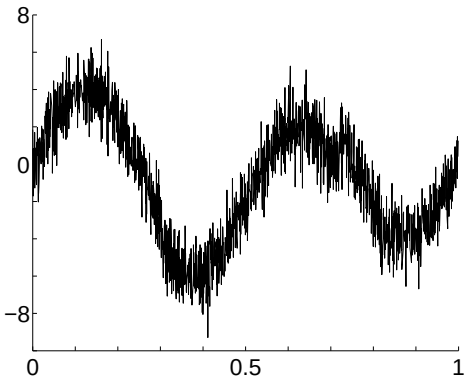

FIG 6. *HSd1: HeaviSine,* $\sigma \equiv 1$, $n = 2048$

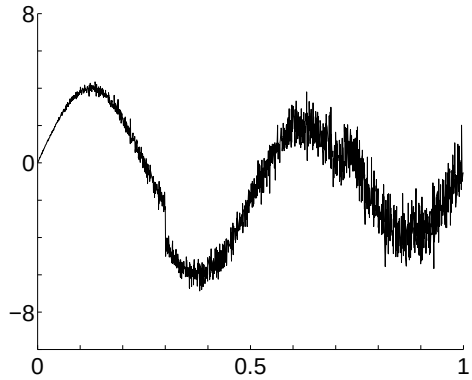

FIG 7. *HSd2: HeaviSine,* $\sigma(x) = x$, $n = 2048$

VFCV Classical V-fold cross-validation, defined by (1), with $V \in \{2, 5, 10, 20\}$.

LOO Classical Leave-one-out (*i.e.* VFCV with $V = n$).

penVF V-fold penalty, with $V \in \{2, 5, 10, 20\}$. $C = C_{W,\infty} = V - 1$. The partition (B_j) is chosen once, as in Algorithm 1, and $\mathrm{pen}_{\mathrm{VF}}$ is defined by (11). In practice, this is almost the same as Algorithm 2.

penLoo V-fold penalty, with $V = n$. $C = C_{W,\infty} = n - 1$.

Mal Mallows' C_p penalty: $\mathrm{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$, where $\widehat{\sigma}^2 = 2n^{-1} d^2 \left(Y_{1\ldots n}, S_{n/2} \right)$ is the classical variance estimator (d being the Euclidean distance on $\mathbb{R}^n$, $S_{n/2}$ any vector space of dimension $n/2$ of $\mathbb{R}^n$ and $Y_{1\ldots n} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$). The non-asymptotic validity of this procedure for model selection in homoscedastic regression has been assessed by Baraud [Bar00].

$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$ Ideal deterministic penalty: $\mathrm{pen}(m) = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$. We use it as a witness of what is a good performance in each experiment.

For each penalization procedure, we also consider the same penalty multiplied by 5/4 (denoted by a + symbol added after its shortened name). This intends to test for overpenalization (the choice of the factor 5/4 being arbitrary and certainly not optimal).

In each experiment, for each simulated data set, we replace $\mathcal{M}_n$ by $\widehat{\mathcal{M}}_n$ as in step 1 of Algorithm 2. Then, we compute the least-squares estimators $\widehat{s}_m$ for each $m \in \widehat{\mathcal{M}}_n$. Finally, we select $\widehat{m} \in \widehat{\mathcal{M}}_n$ using each algorithm and compute its true excess loss $l(s, \widehat{s}_{\widehat{m}})$ (and the excess loss $l(s, \widehat{s}_m)$ for every $m \in \widehat{\mathcal{M}}_n$). We simulate $N = 1000$ data sets, from which we can estimate the model selection performance of each procedure, through the two following benchmarks:

$$C_{\mathrm{or}} = \frac{\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m) \right]} \qquad \text{and} \qquad C_{\mathrm{path-or}} = \mathbb{E}\left[\frac{l(s, \widehat{s}_{\widehat{m}})}{\inf_{m \in \mathcal{M}_n} l(s, \widehat{s}_m)} \right] .$$

Basically, C_{or} is the constant that should appear in an oracle inequality like (14), and $C_{\mathrm{path-or}}$ corresponds to a pathwise oracle inequality like (13). As C_{or} and $C_{\mathrm{path-or}}$ approximatively give

the same rankings between algorithms, we only report C_{or} in Tab. 1.

4.2. *Results and comments.* First of all, our experiments show the interest of both penVF and VFCV in several difficult framework, with relatively small sample sizes. Although it can not compete with simple procedures such as Mallows' C_p from the computational viewpoint, it is much more efficient when the noise is heteroscedastic (S2 and HSd2). In these hard frameworks, the performances of penVF and VFCV are comparable to those of the "ideal deterministic penalty" $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$. On the other hand, they perform slighlty worse than Mallows' for the easier problems (S1 and HSd1), which we interpretate as the unavoidable price for robustness.

Secondly, in the four experiments, the best procedures are always the overpenalizing ones: many of them even beat the perfectly unbiased $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$, showing the crucial need to overpenalize. This is mainly due to the small sample size compared to the high noise-level, since it is no the case when σ is smaller, and less obvious when n is larger (see respectively experiments S0.1 and S1000 in Chap. 5 of [Arl07]). We would like to insist on the importance of this phenomenon, which is seldom mentioned because it it vanishes in the asymptotic framework, and it is quite hard to find from theoretical results.

We can now come back to the discussion of Sect. 2.3 on the choice of V for VFCV, which is enlightened by the results of Tab. 1. In the first three experiments, and more clearly in HSd1, $V = 2$ has comparable or better performances than $V \in \{5, 10, 20, n\}$. This is highly non intuitive, unless we consider the need for overpenalization in those experiments where the signal-to-noise ratio is quite low. It appears that the variability issue is less important in those three cases. This is not because the variance of $\mathrm{crit}_{\mathrm{VFCV}}$ is negligible in front of its bias, but mainly because its dependence on V is only mild. Hence, whatever V, it has to be compensate by overpenalizing. On the contrary, the best choices are $V = 20$ and $V = n$ in experiment HSd2, where overpenalization seems to be less needed. The main conclusion here should be that one really has to take into account both overpenalization and variance for choosing an optimal V. The larger V is not always the better one, so that a larger computation time does not always improve the accuracy. The main difficulty here is that it does not seem straightforward to choose V from the data only.

Finally, let us compare the performances of V-fold cross-validation and V-fold penalization in Tab. 1. At first glance, it seems that penVF with $V < 20$ performs worse than VFCV in the first three experiments, and not clearly better in the last one. The point is that it matches exactly with the experiments for which overpenalization is crucial. But looking at the performance of penVF+, we have evidence for the advantage conferred to penVF by its flexibility. In three over four experiments, penVF+ with any $V \in \{5, 10, 20, n\}$ does better than VFCV with any choice of V; and it is almost the case for HSd1. This comes from the overpenalizing ability of V-fold penalization, which is crucial in such non-asymptotic situations.

Moreover, choosing the optimal V for penVF or penVF+ is much simpler than for VFCV: it is always the largest V. Remark that $V = n$ does not always perform significantly better than

TABLE 1

*Accuracy indexes C_{or} for each algorithm in four experiments, $\pm$ a rough estimate of uncertainty of the value reported (*i.e. *the empirical standard deviation divided by $\sqrt{N}$). In each column, the more accurate algorithms (taking the uncertainty into account; $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$ and $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]+$ are not taken into account there) are bolded.*

Experiment	S1	S2	HSd1	HSd2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (sample size)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.919 ± 0.03	2.296 ± 0.05	1.028 ± 0.004	1.102 ± 0.004
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]+$	1.792 ± 0.03	2.028 ± 0.04	1.003 ± 0.003	1.089 ± 0.004
Mal	1.928 ± 0.04	3.687 ± 0.07	1.015 ± 0.003	1.373 ± 0.010
Mal+	$\mathbf{1.800 \pm 0.03}$	3.173 ± 0.07	$\mathbf{1.002 \pm 0.003}$	1.411 ± 0.008
2−FCV	2.078 ± 0.04	2.542 ± 0.05	$\mathbf{1.002 \pm 0.003}$	1.184 ± 0.004
5−FCV	2.137 ± 0.04	2.582 ± 0.06	1.014 ± 0.003	1.115 ± 0.005
10−FCV	2.097 ± 0.05	2.603 ± 0.06	1.021 ± 0.003	1.109 ± 0.004
20−FCV	2.088 ± 0.04	2.578 ± 0.06	1.029 ± 0.004	1.105 ± 0.004
LOO	2.077 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2−F	2.578 ± 0.06	3.061 ± 0.07	1.038 ± 0.004	$\mathbf{1.103 \pm 0.005}$
pen5−F	2.219 ± 0.05	2.750 ± 0.06	1.037 ± 0.004	1.104 ± 0.004
pen10−F	2.121 ± 0.05	2.653 ± 0.06	1.034 ± 0.004	1.104 ± 0.004
pen20−F	2.085 ± 0.04	2.639 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penLoo	2.080 ± 0.05	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2−F+	2.175 ± 0.05	2.748 ± 0.06	1.011 ± 0.003	1.106 ± 0.004
pen5−F+	1.913 ± 0.03	2.378 ± 0.05	$\mathbf{1.006 \pm 0.003}$	$\mathbf{1.102 \pm 0.004}$
pen10−F+	1.872 ± 0.03	$\mathbf{2.285 \pm 0.05}$	$\mathbf{1.005 \pm 0.003}$	$\mathbf{1.098 \pm 0.004}$
pen20−F+	1.898 ± 0.04	$\mathbf{2.254 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.098 \pm 0.004}$
penLoo+	$\mathbf{1.844 \pm 0.03}$	$\mathbf{2.215 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.096 \pm 0.004}$

$V = 20$ or $V = 10$, which can be considered as almost optimal choices. For the practical user, the choice of V thus reduces to a trade-off between computational complexity and performance (the latter being governed by the variability of the V-fold penalties). Then, once V is chosen, C has to be taken equal to $(V-1)$ times the overpenalization factor (and estimating it from the data remains an open question).

We conclude this section by some additional remarks, concerning some particular points of our simulation study.

- We also performed Mallows' C_p (and its overpenalized version Mal+) with the true mean variance $\mathbb{E}\left[\sigma^2(X)\right]$ instead of $\widehat{\sigma}^2$ (which would not be possible on a real data set). It gave worse performance for all experiments but S2, in which $C_{\mathrm{or}}(\mathrm{Mal}) = 2.657 \pm 0.06$ and $C_{\mathrm{or}}(\mathrm{Mal+}) = 2.437 \pm 0.05$. This shows that overpenalization is really crucial in experiment S2, even more than the shape of the penalty itself. But once we overpenalize, penVF+ remains significantly better than Mallows' C_p ($\mathrm{crit}_{\mathrm{VFCV}}$ being too variable for small V to do better than Mallows). The ability to overpenalize with penVF while keeping the variability low (*i.e.* V large) thus appears to be crucial in this case. In addition, it can be proved that Mallows' C_p penalty (and, more generally, any penalty of the form $\widehat{K}D_m$) leads to suboptimal model selection in some heteroscedastic framework. See [Arl07], Chap. 4. This should be compared to Thm. 2, which can be applied in that framework.
- In experiment HSd1, 2-fold cross-validation appears to be among the best model selection procedures overall. This should be linked with the fact that $\mathcal{M}_n$ only consists on histograms on dyadic partitions of $[0,1]$, so that the assumptions of Thm. 1 are not fulfilled. More precisely, our computations may show that the model which minimize $\mathbb{E}\left[\mathrm{crit}_{\mathrm{VFCV}}(m)\right]$ with $V = 2$ is the oracle model for arbitrarily large values of n. This emphasizes the fact that VFCV is not universally suboptimal for model selection for prediction. It is only unable to make the right choice among estimators whose excess losses are within a constant factor smaller than some $K(V) > 1$.
- Eight additional experiments are reported in Chap. 5 of [Arl07], showing similar results with various n, σ and s (the assumptions of Thm. 2 not being always satisfied). Notice that overpenalization is not always necessary, in particular when the signal-to-noise ratio is larger. In such situations, $V = 20$ or $V = n$ is generally optimal for VFCV.

5. Discussion.

5.1. *V-fold cross-validation* vs. *V-fold penalties.* Time has come for us to give an accurate answer to this practical (but quite hard) question: how to use V-fold?

Firstly, the classical V-fold cross-validation is biased and asymptotically suboptimal for prediction in some "easy framework" (*i.e.* with a smooth regression function and an homoscedastic Gaussian noise). It thus has to be corrected, and we suggest a V-fold penalization algorithm that provides such a correction. This algorithm is asymptotically optimal in theory, quite efficient on some simulated data, and has the same computational cost as VFCV.

Secondly, a non-asymptotic phenomenon is likely to arise, that make the problem harder: when the sample size is small and the noise-level large, overpenalizing procedures are more efficient than unbiased ones. Then, our V-fold penalization method allows to choose an overpenalizing factor, whereas VFCV imposes it (through V) and a corrected VFCV forbids it. This flexibility is the main reason why we suggest to use penVF instead of VFCV or Burman's corrected VFCV. Otherwise, V has to be chosen very carefully, taking into account variability, bias and the possible need for some bias.

We shall now explain how to use V-fold penalties. It depends on two tuning parameters: the number V of folds and the overpenalization factor $C/(V-1)$. The choice of V depends on the trade-off between variability and computational complexity. If the latter one does not matter, the optimal choice is close to $V = n$ (at least for least-squares regression). Otherwise, the choice has to be done by the final user. We refer to asymptotic computations of Burman [Bur89, Bur90] (in linear regression) and the recent work of Celisse and Robin [CR08] (in density estimation) for quantitative measures of variability according to V. Further research in that direction would be very useful for practical use of V-fold model selection criteria.

The question of choosing the overpenalization factor is probably harder to solve. According to our simulation study, the optimal one depends at least on the sample size, the noise level and the smoothness of the regression function. Since the first criterion is that the penalty almost never underestimates the ideal one, a wise choice of C depends on the fluctuations of both the V-fold penalty and the ideal penalty. We thus need a better understanding of the variability of penVF. Another idea would be to replace the conditional expectation in (7) by a quantile, in order to build a simultaneous confidence region for the prediction errors $(P\gamma(\widehat{s}_m))_{m\in\mathcal{M}_n}$. Then, we could deduce a confidence set, to which the oracle model should belong. Defining $\widehat{m}$ as the more parcimonious model in this confidence set, we would have done the work of overpenalization by choosing the probability coverage of the confidence region. We refer to [Arl07] (Sect. 6.6 and 11.3.3) for further discussions about overpenalization.

5.2. *Other cross-validation methods.* In this paper, we focused on VFCV and penVF, among many other cross-validation like methods: hold-out, repeated learning-testing methods [BFOS84], leave-p-out, *etc.* However, it follows from our proofs that the asymptotic performances of these methods mainly depends on their bias, which is itself a function of the ratio between the size of the learning set and the sample size. It is thus possible to have asymptotic optimality with any complexity cost, even without using penVF.

Let us fix for instance the computational complexity to the one of 2-fold cross-validation. We may use 2-fold cross-validation, Burman's corrected 2-fold CV, 2-fold penalization or repeated learning-testing methods (with 2 splits of the data and a learning set of size equivalent to the sample size n). Asymptotically, the first one is suboptimal (Thm. 1), while the three other ones are optimal (Thm 2 and the proof of Thm. 1). We have already seen in Sect. 5.1 that Burman's corrected 2-fold can not overpenalize when needed, which can be a serious drawback in non-asymptotic situations. Repeated learning-testing does not have this drawback, since it is

possible to overpenalize within any factor $C \geq 1$ by choosing a learning set of size $\sim n/(2C-1)$.

However, there remains a strong argument in favour of 2-fold penalization. When C has to be taken close to 1 (which is the asymptotic situation), repeating learning-testing requires the size of the learning set to be very close to n. Hence, if we can only make two splits, most of the data remains in both learning sets. This makes the final criterion much variable, since it strongly depends on the few data which belong to the union of the two training sets. On the contrary, with 2-fold penalization (as well as 2-fold cross-validation and its corrected version), each data point belongs is used once for learning and once for training.

Finally, it seems to us that V-fold penalization should be preferred, because of its versatility: it is asymptotically optimal, quite flexible (for non-asymptotic situations) and makes use of all the data for both learning and training.

5.3. *Prediction in other frameworks.* In order to make theoretical computations feasible, we restricted ourselves to the histogram regression framework in this article. Of course, this is only a first step towards a more general study of V-fold methods for model selection. Although all our proofs strongly rely on some particular features of histograms (in particular for computing expectations), we conjecture than most of our conclusions stay valid much more generally. The main argument supporting this claim is that part of our concentration inequalities are still valid in a general framework, including bounded regression and binary classification. Accurate statements and proofs are to be found in Chap. 7 of [Arl07]. In addition, penVF is built upon the same general heuristics as VFCV, and was never designed particularly for the heteroscedastic histogram regression problem. Hence, it should have at least the same robustness and adaptivity properties as VFCV, while its flexibility should allow better performance in terms of multiplicative constants (which may be crucial, when the sample size is small).

Let us now point out some expected changes in our analysis in the general case. First, the no-overpenalization constant $C_{W,\infty}$ may not stay equal to $V-1$. Although me mentioned an asymptotic theoretical argument, it may break down when one considers models with a large number of parameters (that is, dependent from n). If this occurs, we suggest to use a data-dependent procedure for estimating $C_{W,\infty}$, based upon the so-called "slope heuristics" [BM06, AM08]. Basically, it states that $C_{W,\infty}$ is twice the constant under which $D_{\widehat{m}}$ blows up dramatically. We refer to the above papers for a detailed statement of this algorithm, as well as theoretical insights.

Second, the influence of V on variability may also be quite different. For instance, in classification, it is often noticed that the leave-one-out is much more variable than VFCV with smaller values of V [HTF01]. According to Molinaro, Simon and Pfeiffer [MSP05], this seems to disappear when the algorithm producing $\widehat{s}_m$ is stable. In addition, in the density estimation framework, Celisse and Robin [CR08] also report that the variance of $\mathrm{crit}_{\mathrm{VFCV}}$ increases for large V. We believe that an extensive study of this variability issue in all those frameworks should be made, considering that it is a crucial point for choosing V for VFCV. It would also be quite interesting to determine whether the variability of penVF depends on V in the same way or not.

5.4. *Consistency.* We focused in this article on prediction, but one often uses model selection for identification. In this framework, one assumes that $s \in S_{m^\star}$ (and maybe also to some more complex models), and the goal of a model selection procedure is to catch $m^\star$ as often as possible, whatever the prediction risk of $\widehat{s}_{m^\star}$. Asymptotic optimality there become consistency, *i.e.*

$$\mathbb{P}\left(\widehat{m} = m^\star\right) \xrightarrow[n\to\infty]{} 1 \ .$$

There is a huge amount of papers about model selection for identification; we refer to the introduction of papers by Yang [Yan06, Yan07] for references about the consistency of cross-validation in the regression and classification settings.

The main point for consistency is that overpenalization is needed, even from the asymptotic viewpoint. This is the main reason why BIC is roughly the AIC criterion multiplied by a constant times $\ln(n)$. See also Aerts, Claeskens and Hart [ACH99] about this question. Our penalization interpretation of VFCV (and more generally, any cross-validation like method) then enlightens several theoretical and empirical results about the consistency issue.

With VFCV, the overpenalization factor is bounded from above by 3/2 (which corresponds to $V = 2$). Hence, V-fold cross-validation may be inconsistent in general for any V (although it can sometimes be used, when one compares sufficiently different models, see Yang [Yan07]). Moreover, the better choice is often $V = 2$ as remarked by Zhang [Zha93], Dietterich [Die98] and Alpaydin [Alp99]. On the contrary, V-fold penalties could work, by choosing $C \propto (V-1)\ln(n)$ (for instance). We conjecture that such a method would be consistent, whatever V.

More generally, it has been noticed several times that the consistency of cross-validation requires the size of the learning set to be chosen negligible in front of the sample size. In the linear regression framework, this has be shown by Shao [Sha93, Sha97]. In the classification setting, this is called the "cross-validation paradox" by Yang [Yan06]. With penVF, we believe that we may have proposed a way of solving this paradox, by allowing to choose the overpenalization factor independently from the size of the learning set.

APPENDIX A: PROBABILISTIC TOOLS

In this section, we give some probability theory results that we need to prove our main result, while being of self-interest. In the rest of the paper, for any $a, b \in \mathbb{R}$, we denote by $a \wedge b$ the minimum of a and b, and by $a \vee b$ the maximum of a and b.

A.1. Expectations of inverses of binomials. For any non-negative random variable Z, define

$$e_Z^+ = e_{\mathcal{L}(Z)}^+ := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\,\middle|\, Z > 0\right] \ .$$

Non-asymptotic bounds on this quantity when Z has a binomial distribution are required in the proof of Prop. 1, which is at the core of our main results. Former results concerning e_Z^+ can be found in papers by Lew [Lew76] (for general Z) or Znidaric [Žni05] (for the binomial case), but they are either asymptotic or not accurate enough. The following lemma solves this issue.

LEMMA 3. *For any $n \in \mathbb{N}\setminus\{0\}$ and $p \in (0;1]$, $\mathcal{B}(n,p)$ denotes the binomial distribution with parameters (n,p), $\kappa_3 = 5.1$ and $\kappa_4 = 3.2$. Then, if $np \geq 1$,*

$$\kappa_4 \wedge \left(1 + \kappa_3 (np)^{-1/4}\right) \geq e^+_{\mathcal{B}(n,p)} \geq 1 - e^{-np} \quad . \tag{15}$$

In particular, $e^+_{\mathcal{B}(n,p)} \to 1$ when $np \to \infty$, which can be derived from [Žni05].

A.2. Concentration of inverses of multinomials. Let $(X_\lambda)_{\lambda\in\Lambda_m} \sim \mathcal{M}(n;(p_\lambda)_{\lambda\in\Lambda_m})$ be a multinomial random vector, $(a_\lambda)_{\lambda\in\Lambda_m}$ a family of non-negative real numbers, and define for every $T \in (0,1]$

$$Z_{m,T} := \sum_{\lambda\in\Lambda_m} a_\lambda \min\left(T, X_\lambda^{-1}\right) \quad .$$

Such a quantity naturally appears in our setting, mainly because of the randomness of the design. Unfortunately, classical concentration inequalities for sums of random variables can not be applied to $Z_{m,T}$ because the X_λ are not independent. Using that they are negatively associated [JDP83], we can use the Cramér-Chernoff method [DR98] to obtain the following lemma. Its complete proof can be found in Sect. 8.8 of [Arl07].

LEMMA 4. *Assume that $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n \geq 1$ and $T \in (0,1]$. Define $c_1 = 0.184$, $c_2 = 0.28$, $c_3 = 9.6$, $c_4 = 0.09$, $c_5 = 10.5$, and for every $t \geq 0$, $\varphi_1(t) = \max(t,1)e^{-\max(t,1)}$.*

1. *Lower deviations: for every $x \geq 0$, with probability at least $1 - e^{-x}$,*

$$\mathbb{E}[Z_{m,1}] - Z_{m,1} \leq \frac{\varphi_1(c_1 B_n)}{c_1} \sum_{\lambda\in\Lambda_m} \frac{a_\lambda}{np_\lambda} + 3\sqrt{2}\sqrt{\sum_{\lambda\in\Lambda_m} \frac{a_\lambda^2}{(np_\lambda)^2}}\sqrt{4D_m \exp(-c_1 B_n) + x} \tag{16}$$

2. *Upper deviations: for every $x \geq 0$, with probability at least $1 - e^{-x}$,*

$$\begin{aligned} Z_{m,T} - \mathbb{E}[Z_{m,T}] &\leq \frac{\varphi_1(c_2 B_n)}{c_2} \sum_{\lambda\in\Lambda_m} \left(\frac{a_\lambda}{np_\lambda}\right) \\ &+ \sqrt{\sum_{\lambda\in\Lambda_m} \left(\frac{a_\lambda}{np_\lambda}\right)^2 (D_m e^{-c_4 B_n} + x)} \times c_3 \vee \left[\frac{c_5 T\sqrt{x + e^{-c_4 B_n}}}{n \min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{a_\lambda}\right\}\sqrt{\sum_{\lambda\in\Lambda_m}\left(\frac{a_\lambda}{np_\lambda}\right)^2}}\right] \end{aligned} \tag{17}$$

A.3. Moment inequalities for some U-statistics. There are several papers about concentration or moment inequalities for U-statistics, *e.g.* [GLZ00, Ada05]. It appears that our main results strongly rely on concentration properties for a particular kind of U-statistics of order 2, which are given by the following lemma. It can be derived either from the aforementioned papers, or from [BBLM05], as we did in Sect. 8.9 of [Arl07].

LEMMA 5. *Let $(a_\lambda)_{\lambda\in\Lambda_m}$ and $(b_\lambda)_{\lambda\in\Lambda_m}$ be two families of real numbers, $(r_\lambda)_{\lambda\in\Lambda_m}$ a family of integers. For all $\lambda\in\Lambda_m$, let $(\xi_{\lambda,i})_{1\le i\le r_\lambda}$ be independent centered random variables admitting $2q$-th moments $m_{2q,\lambda,i}$ for some $q\ge 2$. We define $S_{\lambda,1}$, $S_{\lambda,2}$ and Z as follows:*

$$Z=\sum_{\lambda\in\Lambda_m}\left(a_\lambda S_{\lambda,2}+b_\lambda S_{\lambda,1}^2\right)\quad \textit{with}\quad S_{\lambda,1}=\sum_{i=1}^{r_\lambda}\xi_{\lambda,i}\ \textit{and}\ S_{\lambda,2}=\sum_{i=1}^{r_\lambda}\xi_{\lambda,i}^2\ . \tag{18}$$

Then, there is a numerical constant $\kappa\le 1.271$ such that, for every $q\ge 2$,

$$\|Z-\mathbb{E}[Z]\|_q\le 4\sqrt{\kappa}\sqrt{q}\sqrt{\sum_{\lambda\in\Lambda_m}\left((a_\lambda+b_\lambda)^2\sum_{i=1}^{r_\lambda}m_{2q,\lambda,i}^4\right)}+8\sqrt{2}\kappa q\sqrt{\sum_{\lambda\in\Lambda_m}\left(b_\lambda^2\sum_{1\le i\ne j\le r_\lambda}m_{2q,\lambda,i}^2m_{2q,\lambda,j}^2\right)}\ .$$

APPENDIX B: PROOFS

B.1. Notations. Before starting the proofs, we introduce some notations or conventions:

- The letter L will be used to design "some positive numerical constant, possibly different from some place to another". In the same way, a constant which depends on $c_1,\dots,c_k$ will be denoted $L_{c_1,\dots,c_k}$, and if $(\mathbf{A})$ denotes a set of assumptions, $L_{(\mathbf{A})}$ will be any constant that depends on the parameters appearing in $(\mathbf{A})$.
- For any non-negative random variable Z, we define $e^0_{\mathcal{L}(Z)}:=\mathbb{E}[Z]\,\mathbb{E}[Z^{-1}\mathbb{1}_{Z>0}]$.
- For every model $m\in\mathcal{M}_n$, and every $j\in\{1,\dots,V\}$,

$$\begin{aligned} p_1(m)&:=P\left(\gamma(\widehat{s}_m)-\gamma(s_m)\right) & p_2(m)&:=P_n\left(\gamma(s_m)-\gamma(\widehat{s}_m)\right)\\ p_1^{(-j)}(m)&:=P\left(\gamma(\widehat{s}_m^{(-j)})-\gamma(s_m)\right) & p_2^{(-j)}(m)&:=P_n^{(-j)}\left(\gamma(s_m)-\gamma(\widehat{s}_m^{(-j)})\right)\\ \overline{\delta}(m)&:=(P_n-P)\left(\gamma(s_m)-\gamma(s)\right) & \overline{\delta}^{(j)}(m)&:=(P_n^{(j)}-P)\left(\gamma(\widehat{s}_m^{(-j)})-\gamma(s)\right)\ . \end{aligned}$$

- Histograms-specific notations: for any random variable Z, $q>0$, $m\in\mathcal{M}_n$ and $\lambda\in\Lambda_m$:

$$\mathbb{E}^{\Lambda_m}[Z]:=\mathbb{E}\left[Z\,|\,(\mathbb{1}_{X_i\in I_\lambda})_{1\le i\le n,\,\lambda\in\Lambda_m}\right]\qquad \|Z\|_q^{(\Lambda_m)}:=\mathbb{E}^{\Lambda_m}\left[|Z|^q\right]^{1/q}$$
$$S_{\lambda,1}:=\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\qquad\text{and}\qquad S_{\lambda,2}:=\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2\ .$$

- Conventions for p_1 and p_2 when $\widehat{s}_m$ is not well-defined (in the histogram framework):

$$\widetilde{p_1}(m)=\widetilde{p_1}^{(0)}(m)+\sum_{\lambda\in\Lambda_m}p_\lambda(\sigma_\lambda)^2\mathbb{1}_{\widehat{p}_\lambda=0}\qquad\text{with}\qquad \widetilde{p_1}^{(0)}(m)=\sum_{\lambda\in\Lambda_m}\frac{p_\lambda\mathbb{1}_{\widehat{p}_\lambda>0}}{(n\widehat{p}_\lambda)^2}S_{\lambda,1}^2\ . \tag{19}$$

$$\widetilde{p}_2(m):=p_2(m)+\frac{1}{n}\sum_{\lambda\in\Lambda_m}(\sigma_\lambda)^2\mathbb{1}_{n\widehat{p}_\lambda=0}$$

Notice that whatever the convention we choose (and even if we keep their original definition), p_1 and p_2 have the same value when $\widehat{s}_m$ is uniquely defined, and we will always remove from $\mathcal{M}_n$ the other models. The choice we make here is only important when writing expectations, so it is merely technical. In the following, we will often write simply p_1 (resp. p_2) instead of $\widetilde{p_1}$ (resp. $\widetilde{p}_2$).

B.2. Proof of Thm. 1. The idea of the proof is to show that $\mathrm{crit}_1(m) = P\gamma(\widehat{s}_m)$ and $\mathrm{crit}_2(m) = \mathrm{crit}_{\mathrm{VFCV}}(m) - \widehat{c}$ (for some random quantity $\widehat{c}$ independent from m) satisfy the assumptions of Lemma 6 below, on an event of large probability. To this aim, we will use Prop. 1 as well as concentration inequalities of Sect. B.5.

First, we have to be more precise about what we do with models m such that $\widehat{s}_m^{(-j)}$ is not well defined for at least one $j \in \{1, \dots, V\}$. Denote $E_n(m)$ this event. By (56) in Lemma 12, $E_n(m)$ has a probability smaller than n^{-2} as soon as $D_m \leq Ln(\ln(n))^{-1}$, so that all the reasonable conventions will have the same effect. For the sake of simplicity, we choose in this proof is to eliminate such models from $\mathcal{M}_n$. Notice that this removes automatically models such that $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} \leq 1$, in particular all models of dimension strictly larger than $n/2$.

Denote $\widehat{c} = V^{-1} \sum_{j=1}^V P_n^{(j)} \gamma(s)$. Then, for every $m \in \mathcal{M}_n$,

$$\mathrm{crit}_2(m) := \mathrm{crit}_{\mathrm{VFCV}}(m) - \widehat{c} = l(s, s_m) + \frac{1}{V} \sum_{j=1}^V \left(p_1^{(-j)}(m) + \overline{\delta}^{(j)}(m) \right) + \infty \mathbb{1}_{E_n(m)} \ . \tag{20}$$

First, notice that for every j, conditionally to $(X_i, Y_i)_{i \notin B_j}$, $\widehat{s}_m^{(-j)}$ is deterministic. In addition, $\|Y\|_\infty \leq A := 1 + \sigma \|\epsilon\|_\infty < \infty$ by assumption. So, Lemma 10 can be applied with $t = \widehat{s}_m^{(-j)}$ and n changed into $\mathrm{Card}(B_j) \geq Ln/V$. More precisely, for every $m \in \mathcal{M}_n$ such that $E_n(m)$ does not hold, for every $j \in \{1, \dots, V\}$, taking $x = 4\ln(n)$ and $\eta = \ln(n)^{-1}$, there is an event of probability $1 - Ln^{-4}$ on which

$$\left| \overline{\delta}^{(j)}(m) \right| \leq \frac{l(s, \widehat{s}_m^{(-j)})}{\ln(n)} + \frac{LVA^2 \ln(n)^2}{n} \ . \tag{21}$$

A union bound shows that these inequalities hold uniformly over j and m on an event of probability at least $1 - Ln^{-2}$. Combined with (20), this gives

$$\mathrm{crit}_2(m) \geq \left(1 - \ln(n)^{-1} \right) \left[l(s, s_m) + \frac{1}{V} \sum_{j=1}^V p_1^{(-j)}(m) \right] - \frac{LVA^2 \ln(n)^2}{n} + \infty \mathbb{1}_{E_n(m)} \tag{22}$$

and a similar upper bound.

A second key remark is that for every j, $p_1^{(-j)}$ has the distribution of p_1 with a sample size $n - \mathrm{Card}(B_j)$ instead of n. We can then apply Prop. 9 (with $\gamma = 4$) to get that on an event of

probability $1 - Ln^{-2}$, for every $j \in \{1, \dots, V\}$ and $m \in \mathcal{M}_n$ such that $E_n(m)$ does not hold,

$$p_1^{(-j)}(m) \leq \mathbb{E}\left[p_1^{(-j)}(m)\right] + L_{A,\sigma}\left[\ln(n)^2 D_m^{-1/2} + \sqrt{D_m} e^{-LnD_m^{-1}}\right] \mathbb{E}\left[p_2^{(-j)}(m)\right] \tag{23}$$

$$p_1^{(-j)}(m) \geq \mathbb{E}\left[p_1^{(-j)}(m)\right] - L_{A,\sigma}\left[\ln(n)^2 D_m^{-1/2} + e^{-LnD_m^{-1}}\right] \mathbb{E}\left[p_2^{(-j)}(m)\right] \tag{24}$$

$$p_1^{(-j)}(m) \geq \left(L \ln(n)^{-1} - L_{A,\sigma} \ln(n)^2 D_m^{-1}\right) \mathbb{E}\left[p_2^{(-j)}(m)\right] \quad . \tag{25}$$

Finally, since $s(x) = x$, X is uniform and the models are regular histograms on $\mathcal{X} = [0,1]$, we can compute exactly for each model the bias and the variance term (when the sample size is n):

$$l(s, s_m) = \frac{1}{12 D_m^2} \qquad \text{and} \qquad \mathbb{E}\left[p_2(m)\right] = \frac{\sigma^2 D_m}{n} + \frac{1}{12 D_m n} \quad . \tag{26}$$

We now explain how this can be used to check the assumptions of Lemma 6. Let c_1 and κ_1 be positive constants to be chosen later.

Small models. First, assume that $D_m < \ln(n)^{\kappa_1}$. Combining (22), (24), (26) and using that $\mathbb{E}\left[p_1^{(-j)}(m)\right] \geq 0$, $\mathrm{crit}_2(m)$ is roughly of the order of the bias term. Hence, condition (29) holds with $c_3 = L$ and $\kappa_3 = 2\kappa_1$ when $n \geq L_{A,\sigma,V,\kappa_1}$. Notice that this holds for every $\kappa_1 > 0$.

Intermediate models. We now consider models of dimension $\ln(n)^{\kappa_1} \leq D_m \leq c_1 n (\ln(n))^{-1}$. As already noticed, $E_n(m)$ does not hold true for any of them, with a large probability.

From (22) (and the similar upper bound), (24), (23) and (26), it follows that condition (28) holds with $a = 1/12$, $b = \sigma^2$, $C = V/(V-1)$, $c_2 = L_{A,V,\sigma}$ and $\kappa_2 = 1$, as soon as $n \geq L_{A,\sigma,V}$, $c_1 \leq L$ and $\kappa_1 \geq 6$. Very similar (and somehow simpler) arguments prove that the condition (27) holds with the same parameters.

Large models. Finally, let $m \in \mathcal{M}_n$ be such that $D_m > c_1 n (\ln(n))^{-1}$. Combining (22), (25) and (26), $\mathrm{crit}_2(m)$ is roughly of the order of the variance term $L\mathbb{E}\left[p_2^{(-j)}(m)\right]$ when $n \geq L_{A,\sigma,V,c_1}$. As a result, condition (30) holds with $c_4 = Lc_1\sigma^2$ and $\kappa_4 = 2$, for $n \geq L_{A,\sigma,V,c_1}$.

Choosing now $c_1 \leq L$ and $\kappa_1 = 6$, the conclusion directly follows from Lemma 6 below. Notice that we have assumed several times that $n \geq n_0 = L_{A,\sigma,V}$. These conditions can be dropped by choosing $K_1 \geq n_0^2$. □

LEMMA 6. *Let $a, b, (c_i)_{1 \leq i \leq 4}, (\kappa_i)_{1 \leq i \leq 4}, c_{\mathrm{rich}} > 0$ and $C > 1$ be some constants, $n \in \mathbb{N}$ and $\mathcal{M}_n$ a set of indexes. Assume that for every $m \in \mathcal{M}_n$, $D_m \in [1, n]$, and moreover that $\forall x \in [1, n - c_{\mathrm{rich}}]$, $\exists m \in \mathcal{M}_n$ such that $D_m \in [x, x + c_{\mathrm{rich}}]$. Let crit_1 and crit_2 be some functions $\mathcal{M}_n \mapsto \mathbb{R}$ satisfying the following conditions:*

(i) for every $m \in \mathcal{M}_n$,

$$\mathrm{crit}_1(m) = \left(\frac{a}{D_m^2} + \frac{bD_m}{n} \right) (1 + \epsilon_{1,m}) \tag{27}$$

$$\mathrm{crit}_2(m) = \left(\frac{a}{D_m^2} + \frac{CbD_m}{n} \right) (1 + \epsilon_{2,m}) \tag{28}$$

with $\max_{i=1,2} \sup_{m \in \mathcal{M}_n \ \ s.t. \ \ \ln(n)^{\kappa_1} \leq D_m \leq \frac{c_1 n}{\ln(n)}} |\epsilon_{i,m}| \leq c_2 \ln(n)^{-\kappa_2}$.

(ii) for every $m \in \mathcal{M}_n$ *such that* $D_m < \ln(n)^{\kappa_1}$,

$$\mathrm{crit}_2(m) \geq c_3 \left(\ln(n) \right)^{-\kappa_3} \quad . \tag{29}$$

(iii) for every $m \in \mathcal{M}_n$ *such that* $D_m \geq \frac{c_1 n}{\ln(n)}$,

$$\mathrm{crit}_2(m) \geq c_4 \left(\ln(n) \right)^{-\kappa_4} \quad . \tag{30}$$

Then, there is some constant $K(C) = 2^{2/3} \times 3^{-1} \left(C^{-1/3} - 1 \right)^2 > 0$ *and some* $n_0 > 0$ *(depending on* a, b, $(c_i)_{1 \leq i \leq 4}$, $(\kappa_i)_{1 \leq i \leq 4}$, c_{rich} *and* C*) such that, if* $n \geq n_0$, *for every* $\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \mathrm{crit}_2(m)$,

$$\mathrm{crit}_1(\widehat{m}) \geq \left(1 + K(C) - \ln(n)^{-\kappa_2/5} \right) \inf_{m \in \mathcal{M}_n} \left\{ \mathrm{crit}_1(m) \right\} \quad . \tag{31}$$

SKETCH OF THE PROOF OF LEMMA 6. We skip this proof which is only technical. The main arguments are the following. First, there is a model m_1 of dimension close to $(2an)^{1/3} b^{-1/3}$, so that $\mathrm{crit}_1(m_1)$ is close to $3 \times 2^{-2/3} a^{1/3} b^{2/3} n^{-2/3}$. Second, any model $\widehat{m}$ which minimizes $\mathrm{crit}_2(m)$ must have a dimension close to $(2an)^{1/3} (bC)^{-1/3}$. This implies that $\mathrm{crit}(\widehat{m})$ is larger than $(1 + K(C) - \ln(n)^{-\kappa_2/5}) \mathrm{crit}_1(m_1)$, and the result follows. □

B.3. Proof of Thm. 2. In this section, $L_{(\mathbf{pVF})}$ denotes a constant that depends only on the set of assumptions of Thm. 2, including V. For every $m \in \mathcal{M}_n$, define $\mathrm{pen}'_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}_{\mathrm{id}}(m) + (P - P_n)\gamma(s)$. Then, by definition of $\mathrm{pen}_{\mathrm{id}}$ and $\widehat{m}$, we have for every $m \in \widehat{\mathcal{M}}_n$,

$$l(s, \widehat{s}_{\widehat{m}}) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m}) \right) \leq l(s, \widehat{s}_m) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m) \right) \quad . \tag{32}$$

The idea of the proof is to show that $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ is negligible in front of $l(s, \widehat{s}_m)$ for "reasonable" models (*i.e.*, those which are likely to be either selected by penVF, or an oracle model) with a large probability. We will prove it by using Prop. 1 and 2, as well as the concentration inequalities of Sect. B.5.

For every $m \in \mathcal{M}_n$, define $A_n(m) = \min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \}$ and $B_n(m) = \min_{\lambda \in \Lambda_m} \{ np_\lambda \}$. We now define the event Ω_n on which the concentration inequalities of Prop. 9 and 11 and Lemma 10

and 12, hold with $\gamma = \alpha_{\mathcal{M}} + 2$ (or similarly $x = (\alpha_{\mathcal{M}} + 2)\ln(n)$), for every $m \in \mathcal{M}_n$. Using assumption (**P1**), the union bound gives $\mathbb{P}(\Omega_n) \geq 1 - L_{c_{\mathcal{M}}} n^{-2}$.

First, let $c > 0$ be a constant to be chosen later, and consider $\widetilde{\mathcal{M}}_n$, the set of models $m \in \mathcal{M}_n$ such that $\ln(n)^6 \leq D_m \leq cn(\ln(n))^{-1}$. According to ($\mathbf{Ar}_\ell^{\mathbf{X}}$), this implies $B_n(m) \geq c_{\mathrm{r},\ell}^X c^{-1}\ln(n)$, so that (56) ensures that $A_n(m) \geq \ln(n)$ if $c \leq L_{c_{\mathrm{r},\ell}^X, \alpha_{\mathcal{M}}}$. In particular, $m \in \widetilde{\mathcal{M}}_n$ on Ω_n. Now, using both bounds on D_m, by construction of Ω_n,

$$\max\left\{ \left|\widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m)\right]\right|, \left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right|, \left|\overline{\delta}(m)\right|, \left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \right\}$$

is smaller than $L_{(\mathbf{pVF})}\ln(n)^{-1}\left(l(s,s_m) + \mathbb{E}\left[p_2(m)\right]\right)$ on this event, at least if $c \leq L_{c_{\mathrm{r},\ell}^X}$ (to ensure that $B_n(m)$ is large enough). We now fix $c = L_{c_{\mathrm{r},\ell}^X, \alpha_{\mathcal{M}}}$ that satisfies those two conditions. Using Prop. 2, Lemma 7 and the lower bound on $B_n(m)$, we have for every $m \in \widetilde{\mathcal{M}}_n$

$$\frac{-L_{(\mathbf{pVF})}}{\ln(n)^{1/4}} l(s,\widehat{s}_m) \leq (\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}})(m) \leq \left[2(\eta - 1) + \frac{L_{(\mathbf{pVF})}}{\ln(n)^{1/4}}\right] l(s,\widehat{s}_m) \ .$$

as soon as $n \geq L_{(\mathbf{pVF})}$ (this restriction is necessary because the bounds are in terms of excess loss of $\widehat{s}_m$ instead of $l(s,s_m) + \mathbb{E}\left[p_2\right]$). Combined with (32), this gives: if $n \geq L_{(\mathbf{pVF})}$ and $c \leq L_{c_{\mathrm{r},\ell}^X, \alpha_{\mathcal{M}}}$,

$$l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\widehat{m}\in\widetilde{\mathcal{M}}_n} \leq \left[2\eta - 1 + \frac{L_{(\mathbf{pVF})}}{\ln(n)^{1/4}}\right] \times \inf_{m\in\widetilde{\mathcal{M}}_n}\left\{l(s,\widehat{s}_m)\right\} \ . \tag{33}$$

Second, we prove that any minimizer $\widehat{m}$ of crit belongs to $\widetilde{\mathcal{M}}_n$ on the event Ω_n. Define, for every $m \in \mathcal{M}_n$, $\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s)$, which has the same minimizers over $\widehat{\mathcal{M}}_n$ as crit. According to (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If $n \geq L_{(\mathbf{pVF})}$, $m_0 \in \widetilde{\mathcal{M}}_n$, from which we deduce (using ($\mathbf{Ap}$))

$$\mathrm{crit}'(m_0) \leq l(s,s_{m_0}) + \left|\overline{\delta}(m_0)\right| + \mathrm{pen}(m_0) \leq L_{(\mathbf{pVF})}\left(n^{-\beta_2/2} + n^{-1/2}\right) \ . \tag{34}$$

On the other hand, if $D_m < \ln(n)^6$, we have

$$\mathrm{crit}'(m) \geq l(s,s_m) - \left|\overline{\delta}(m)\right| - p_2(m) \geq C_{\mathrm{b}}^{-}\left(\ln(n)\right)^{-6\beta_1} - L_A\sqrt{\frac{\ln(n)}{n}} - L_{(\mathbf{pVF})}\frac{\ln(n)^7}{n} \tag{35}$$

on Ω_n. In addition, if $D_m > cn(\ln(n))^{-1}$ and $m \in \widehat{\mathcal{M}}_n$, by Prop. 2, $\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) - p_2(m)\right] \geq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right]$. As a consequence, by construction of Ω_n, we have $\mathrm{pen}(m) - p_2(m) \geq (1 - L_{(\mathbf{pVF})}n^{-1/4})\mathbb{E}\left[p_2(m)\right]$ on it, so that

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) - p_2(m) - \left|\overline{\delta}(m)\right| \geq L_{(\mathbf{pVF})}\ln(n)^{-1} \tag{36}$$

when $n \geq L_{(\mathbf{pVF})}$. Comparing (34), (35) and (36), it follows that $\widehat{m} \in \widetilde{\mathcal{M}}_n$ on Ω_n, provided that $n \geq L_{(\mathbf{pVF})}$.

Finally, we show that the infimum can be extended to $\mathcal{M}_n$ in the right-hand side of (33), with the convention $l(s, \widehat{s}_m) = +\infty$ if $A_n(m) = 0$. Using similar arguments as above (as well as the definition of Ω_n, in particular (45) for large models), we have $l(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{pVF})} \left(n^{-\beta_2/2} + n^{-1/2} \right)$ on Ω_n. On the other hand, for every $m \in \mathcal{M}_n$, if $D_m < \ln(n)^6$, $l(s, \widehat{s}_m) \geq l(s, s_m) \geq L_{(\mathbf{pVF})} \ln(n)^{-6\beta_1}$ while if $D_m > cn(\ln(n))^{-1}$, $l(s, \widehat{s}_m) \geq L_{(\mathbf{pVF})} \ln(n)^{-2}$ on Ω_n as soon as $n \geq L_{(\mathbf{pVF})}$. Hence, if $n \geq L_{(\mathbf{pVF})}$, no model $m \notin \widetilde{\mathcal{M}}_n$ can contribute to the infimum in the right-hand side of (33).

To conclude the proof of (13), we notice that $L_{(\mathbf{pVF})} \ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ if $n \geq L_{(\mathbf{pVF})}$. All the conditions of the kind $n \geq n_0$ can finally be removed by enlarging K_1 so that $K_1 n_0^{-2} \geq 1$. The final remark concerning ϵ_n holds true because we can replace the threshold dimensions $\ln(n)^6$ and $cn(\ln(n))^{-1}$ for "small" and "large" models by some powers of n, as soon as the exponents are not taken too far from 0 (resp. 1).

We now get the more classical oracle inequality (13) by noticing that $l(s, \widehat{s}_m) \leq A^2$ a.s., so that

$$\begin{aligned}
\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right] &\leq \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\Omega_n} \right] + \left\| l(s, \widehat{s}_{\widehat{m}}) \right\|_\infty \mathbb{P}\left(\Omega_n^c \right) \\
&\leq \left[2\eta - 1 + \ln(n)^{-1/5} \right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\} \right] + \frac{A^2 K_1}{n^2} \quad . \quad \square
\end{aligned}$$

B.4. Expectations.

B.4.1. *Proof of Prop. 1.*

Ideal criterion. We have to compute $\mathbb{E}\left[P\gamma\left(\widehat{s}_m \right) - P\gamma\left(s_m \right) \right] = \mathbb{E}\left[p_1(m) \right]$. Assume that $\widehat{s}_m$ is well-defined, *i.e.* $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda > 0$. Using that s_m minimizes $P\gamma(t)$ over $t \in S_m$, we have

$$p_1(m) = \sum_{\lambda \in \Lambda_m} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda \right)^2 = \sum_{\lambda \in \Lambda_m} \frac{1}{n^2 \widehat{p}_\lambda} \frac{p_\lambda}{\widehat{p}_\lambda} S_{\lambda,1}^2 \quad \text{so that} \quad \mathbb{E}^{\Lambda_m}\left[p_1(m) \right] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \frac{p_\lambda}{\widehat{p}_\lambda} \left(\sigma_\lambda \right)^2 \quad . \tag{37}$$

The result (3) follows, with $\delta_{n,p_\lambda} = e^0_{\mathcal{B}(n,p_\lambda)} - 1$ if $p_1 = \widetilde{p_1}^{(0)}$, or $\delta_{n,p_\lambda} = e^0_{\mathcal{B}(n,p_\lambda)} - 1 + np_\lambda(1-p_\lambda)^n$ if $p_1 = \widetilde{p_1}$. In each case, the proof of Lemma 3 gives non-asymptotic bounds on δ_{n,p_λ}.

V-fold criterion. By definition (1), on the event on which $\widehat{s}_m^{(-j)}$ is well-defined for every j,

$$\mathrm{crit}_{\mathrm{VFCV}}(m) = \frac{1}{V} \sum_{j=1}^{V} \left[p_1^{(-j)}(m) + \left(P_n^{(j)} - P \right) \gamma\left(\widehat{s}_m^{(-j)} \right) \right] \quad .$$

The second term is centered conditionally to $(X_i, Y_i)_{i \notin B_j}$, so that we only have to compute $\mathbb{E}\left[p_1^{(-j)} \right]$ for every j. Since $(X_i, Y_i)_{i \notin B_j}$ is an i.i.d. sample of size $n - \mathrm{Card}(B_j)$, we can apply

the above computation of $\mathbb{E}[p_1]$. Using a convention similar to $\widetilde{p_1}^{(0)}$ (which can be used on real data, since it does not depend on P), the result (4) holds with

$$\delta_{n,p_\lambda}^{(VF)} = \frac{1}{V}\sum_{j=1}^{V}\left[\frac{n-n/V}{n-\mathrm{Card}(B_j)}\left(e^0_{\mathcal{B}(n-\mathrm{Card}(B_j),p_\lambda)}-1\right)+\frac{\mathrm{Card}(B_j)}{n-\mathrm{Card}(B_j)}-\frac{1}{V-1}\right] .$$

From Lemma 3, we deduce that if $n^{-1}\max_j \mathrm{Card}(B_j) \le c_B < 1$, then

$$\frac{-1}{1-c_B}e^{-np_\lambda(1-c_B)} - L\epsilon_n^{reg} \le \delta_{n,p_\lambda}^{(VF)} \le \frac{L}{(1-c_B)^{5/4}}(np_\lambda)^{-1/4} + L\epsilon_n^{reg} .$$

Similarly to the computation of $p_1(m)$, when $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$, we have

$$p_2(m) = \sum_{\lambda\in\Lambda_m}\widehat{p}_\lambda\left(\beta_\lambda-\widehat{\beta}_\lambda\right)^2 = \sum_{\lambda\in\Lambda_m}\frac{S_{\lambda,1}^2\mathbb{1}_{n\widehat{p}_\lambda>0}}{n^2\widehat{p}_\lambda} \quad \text{so that} \quad \mathbb{E}^{\Lambda_m}[p_2(m)] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}(\sigma_\lambda)^2 . \tag{38}$$

Notice that $\mathbb{E}^{\Lambda_m}[p_2(m)] = \mathbb{E}[\widetilde{p}_2(m)]$ on this event. Using Lemma 3, this proves the following.

Lemma 7. *If* $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \ge B \ge 1$,

$$\left(1-e^{-B}\right)\mathbb{E}[\widetilde{p}_2(m)] \le \mathbb{E}\left[\widetilde{p_1}^{(0)}(m)\right] \le \mathbb{E}[\widetilde{p_1}(m)] \le \left(1+\sup_{np\ge B}\delta_{n,p}\right)\mathbb{E}[\widetilde{p}_2(m)] \tag{39}$$

where $\delta_{n,p}$ comes from Prop. 1. A similar result holds with p_2 instead of $\widetilde{p}_2$ inside the expectation.

B.4.2. *Proof of Prop. 2.* First of all, notice that all this proof is made *conditionally to* $(\mathbb{1}_{X_i\in I_\lambda})_{1\le i\le n,\,\lambda\in\Lambda_m}$. The outline of the proof is to prove that $\mathbb{E}^{\Lambda_m}[\mathrm{pen}_{\mathrm{VF}}]$ can be derived from the case where W satisfies an exchangeability condition, for which we can use Lemma 8 below. This is why we consider more generally the penalty $\mathrm{pen}_W\left(m,(X_i,Y_i)_{1\le i\le n}\right)$, defined by (11) for a general weight vector $W\in\mathbb{R}^n$, strengthening its dependence on the distribution of W and the data. When W is the subsampling weight vector of interest, pen_W coincides with the definition of $\mathrm{pen}_{\mathrm{VF}}$ in Algorithm 2.

Let σ be a random permutation of $\{1,\dots,n\}$, independent from W and the data, and uniform over the permutations that leave $(\mathbb{1}_{X_i\in I_\lambda})_{1\le i\le n,\,\lambda\in\Lambda_m}$ invariant. Defining $\widetilde{W} = \left(W_{\sigma(i)}\right)_{1\le i\le n}$,

$$\begin{aligned}\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\widetilde{W}}\left(m,(X_i,Y_i)_{1\le i\le n}\right)\right] &= \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_W\left(m,\left(X_{\sigma^{-1}(i)},Y_{\sigma^{-1}(i)}\right)_{1\le i\le n}\right)\right]\\ &= \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_W\left(m,(X_i,Y_i)_{1\le i\le n}\right)\right]\end{aligned}$$

since the penalty does not depend on the order of $(W_i,X_i,Y_i)_{X_i\in I_\lambda}$ (for the first equality), and $(X_i,Y_i)_{X_i\in I_\lambda}$ is exchangeable (for the second equality). Moreover, for every $\lambda\in\Lambda_m$, $(\widetilde{W}_i)_{X_i\in I_\lambda}$

is exchangeable and independent from $(X_i, Y_i)_{X_i \in I_\lambda}$. We can thus use Lemma 8 to compute $\mathrm{pen}_{\widetilde{W}}(m)$. Then,

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{n} \sum_{\lambda \in \Lambda_m} \left(R_{1,\widetilde{W}}(n, \widehat{p}_\lambda) + R_{2,\widetilde{W}}(n, \widehat{p}_\lambda) \right) (\sigma_\lambda)^2 \quad .$$

It now remains to compute $R_{1,\widetilde{W}}$ and $R_{2,\widetilde{W}}$. If V divides $n\widehat{p}_\lambda$, then $W_\lambda = 1$ a.s. and $R_{1,\widetilde{W}} = R_{2,\widetilde{W}} = (V-1)^{-1}$. For the general case, see the proof of Prop. 5.2 in [Arl07] (Sect. 5.7.2). □

Lemma 8 (Lemma 5.7 of [Arl07]). *Let S_m be the model of histograms adapted to some partition $(I_\lambda)_{\lambda \in \Lambda_m}$, $W \in [0;\infty)^n$ be a random vector such that for every $\lambda \in \Lambda_m$, $(W_i)_{X_i \in I_\lambda}$ is exchangeable and independent from $(X_i, Y_i)_{X_i \in I_\lambda}$. Define the Resampling Penalty for histograms as* (11)*, and assume* $\min_{\lambda \in \Lambda_m}\{n\widehat{p}_\lambda\} \geq 2$. *Then,*

$$\mathrm{pen}(m) = \frac{C}{n} \sum_{\lambda \in \Lambda_m} \left(R_{1,W}(n, \widehat{p}_\lambda) + R_{2,W}(n, \widehat{p}_\lambda) \right) \frac{n\widehat{p}_\lambda S_{\lambda,2} - S_{\lambda,1}^2}{n\widehat{p}_\lambda - 1} \quad , \qquad \textit{where} \tag{40}$$

$$R_{1,W}(n, \widehat{p}_\lambda) = \mathbb{E}^{\Lambda_m}\left[\frac{(W_{i_\lambda} - W_\lambda)^2}{W_\lambda^2} \,\middle|\, W_\lambda > 0 \right] \quad R_{2,W}(n, \widehat{p}_\lambda) = \mathbb{E}^{\Lambda_m}\left[\frac{(W_{i_\lambda} - W_\lambda)^2}{W_\lambda} \right] \quad . \tag{41}$$

and i_λ is any index such that $X_{i_\lambda} \in I_\lambda$.

B.5. Concentration results. In order to prove Thm. 1 and 2, we need to combine Prop. 1 and 2 with concentration inequalities, which are the purpose of the present section. Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda \in \Lambda_m}$, and assume that both (**Ab**) and (**An**) are satisfied (see the statement of Thm. 2).

Our first result has to deal with p_1 and p_2, which are the main components of the ideal penalty. Whereas concentration for p_2 can be obtained in a general framework (see [Arl07], Chap. 7), lower bounds on p_1 are completely new, up to our best knowledge.

Proposition 9. *Let $\gamma > 0$ and assume that $\min_{\lambda \in \Lambda_m}\{np_\lambda\} \geq B_n$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{A,\sigma_{\min},\gamma}\left[\ln(n)^2 D_m^{-1/2} + e^{-LB_n}\right] \mathbb{E}\left[p_2(m)\right] \tag{42}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{A,\sigma_{\min},\gamma}\left[\ln(n)^2 D_m^{-1/2} + \sqrt{D_m} e^{-LB_n}\right] \mathbb{E}\left[p_2(m)\right] \tag{43}$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{A,\sigma_{\min},\gamma} D_m^{-1/2} \ln(n) \mathbb{E}\left[p_2(m)\right] \quad . \tag{44}$$

In addition, if $B_n > 0$, there is an event of probability at least $1 - Ln^{-\gamma}$ on which

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n^{-1}\ln(n)} - L_{A,\sigma_{\min},\gamma} \ln(n)^2 D_m^{-1/2} \right) \mathbb{E}\left[\widetilde{p}_2(m)\right] \quad . \tag{45}$$

PROOF OF PROP. 9. According to the explicit expressions (37) and (38), $\widetilde{p_1}(m)$ and $p_2(m)$ are both U-statistics of order 2 conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda)}$. Then, we use Lemma 5, with $\xi_{i,\lambda}=Y_i-\beta_\lambda$, $a_\lambda=0$, $b_\lambda=p_\lambda(n\widehat{p}_\lambda)^{-2}$ for $\widetilde{p_1}$ and $b_\lambda=(n^2\widehat{p}_\lambda)^{-1}$ for p_2. This proves, for all $q\geq 2$,

$$\left\|\widetilde{p_1}(m)-\mathbb{E}^{\Lambda_m}[\widetilde{p_1}(m)]\right\|_q^{(\Lambda_m)} \leq \max_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\mathbb{1}_{\widehat{p}_\lambda>0}\right\} L_{A,\sigma_{\min}} D_m^{-1/2} q\mathbb{E}\left[p_2(m)\right] \tag{46}$$

$$\|p_2(m)-\mathbb{E}[p_2(m)]\|_q^{(\Lambda_m)} \leq L_{A,\sigma_{\min}} D_m^{-1/2} q\mathbb{E}\left[p_2(m)\right] \quad . \tag{47}$$

We deduce conditional concentration inequalities from those moment inequalities (for instance by Lemma 8.9 of [Arl07]), with a deterministic probability bound $1-Le^{-x}=1-n^{-\gamma}$. Hence, we deduce unconditional concentration inequalities, and the result follows for p_2. To control the remainder term for $\widetilde{p_1}$, we use 54 in Lemma 12.

We now have to control the distance between $\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}\right]$ and $\mathbb{E}\left[\widetilde{p_1}\right]$. First, if $B_n\geq 1$, we can use Lemma 4: taking $X_\lambda=n\widehat{p}_\lambda$ and $a_\lambda=p_\lambda\left(\sigma_\lambda\right)^2$, according to (37), we have $\widetilde{p_1}(m)=Z_{m,1}$ and the concentration inequality for $\widetilde{p_1}$ follows. On the other hand, if we only know that $B_n>0$, instead of using Lemma 4, we remark that

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}(m)\right] \geq \min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\right\}\mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \quad ,$$

and the result follows thanks to (55) in Lemma 12. □

We mention here a much classical result, which is a consequence of Bernstein's inequality, since it deals with sums of independent variables. We refer to [AM08] for a detailed proof.

LEMMA 10 (Prop. 3, [AM08]). *Let t be any deterministic predictor. For every $x\geq 0$, there is an event of probability at least $1-2e^{-x}$ on which*

$$\forall\eta>0,\quad \left|\left(P-P_n\right)\left(\gamma\left(t\right)-\gamma\left(s\right)\right)\right| \leq \eta l(s,t)+\left(\frac{4}{\eta}+\frac{8}{3}\right)\frac{A^2x}{n} \quad . \tag{48}$$

Finally, we consider the V-fold penalties defined by Algorithm 2.

PROPOSITION 11. *Let* $\mathrm{pen}(m)$ *be defined by* (10) *with the weights W defined in Algorithm 2 and $\gamma>0$. There is an event of probability at least $1-n^{-\gamma}$ on which, if* $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda>0$,

$$\left|\mathrm{pen}(m)-\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \leq C\left(\frac{1}{\min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\}}\vee\frac{1}{V}\right)L_{A,\sigma_{\min},\gamma}D_m^{-1/2}\ln(n)\mathbb{E}\left[p_2(m)\right] \quad . \tag{49}$$

PROOF OF PROP. 11. By definition (10), $\mathrm{pen}(m)=\mathbb{E}_W\left[Z\right]$ with

$$Z=\sum_{\lambda\in\Lambda_m}\left(\widehat{p}_\lambda+\widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda-\widehat{\beta}_\lambda^W\right)^2 \quad =\sum_{\lambda\in\Lambda_m}\frac{1+W_\lambda}{n^2\widehat{p}_\lambda W_\lambda^2}\left[\sum_{X_i\in I_\lambda}\left(W_\lambda-W_i\right)\left(Y_i-\beta_\lambda\right)\right]^2 \quad . \tag{50}$$

For every $q \geq 1$, using Jensen inequality and the independence between W and the data (conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{i,\lambda}$),

$$\begin{aligned} \left\| \mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] \right\|_q^{(\Lambda_m)} &\leq \left\| Z - \mathbb{E}^{\Lambda_m}\left[Z \mid W \right] \right\|_q^{(\Lambda_m)} \\ &\leq \sup_{W_0 \in \mathrm{supp}(W)} \left\{ \left\| Z - \mathbb{E}^{\Lambda_m}\left[Z \mid W = W_0 \right] \right\|_q^{(W_0,\Lambda_m)} \right\} \end{aligned} \tag{51}$$

where $\mathrm{supp}(W)$ is the support of the resampling weight vector W distribution (conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{i,\lambda}$) and $\|\cdot\|_q^{(W_0,\Lambda_m)}$ denotes the q-th moment conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{(i,\lambda)}$ and $W = W_0$. In other words, the deviations of pen are smaller than those of the worse case with a deterministic weight vector $W_0 \in \mathrm{supp}(W)$.

From now on, we work conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{(i,\lambda)}$ and assume that $W \in \mathbb{R}^n$ is deterministic, among those authorized by Algorithm 2. Denote by $X_{(1,\lambda)}, \ldots, X_{(n\widehat{p}_\lambda,\lambda)}$ the data such that $X_i \in I_\lambda$. According to (50), Lemma 5 with $r_\lambda = n\widehat{p}_\lambda$, $a_\lambda = 0$, $b_\lambda = (1 + W_\lambda)(n^2 \widehat{p}_\lambda W_\lambda^2)^{-1}$ and $\xi_{i,\lambda} = \left(W_{(i,\lambda)} - W_\lambda \right) \left(Y_{(i,\lambda)} - \beta_\lambda \right)$ shows that

$$\left\| Z - \mathbb{E}^{\Lambda_m}\left[Z \right] \right\|_q^{(W,\Lambda_m)} \leq \frac{LA^2 q}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \left(\frac{1 + W_\lambda}{n\widehat{p}_\lambda W_\lambda^2} \sum_{i=1}^{n\widehat{p}_\lambda} \left(W_{(i,\lambda)} - W_\lambda \right)^2 \right)^2} \ .$$

We now fix some $\lambda \in \Lambda_m$ and write $n\widehat{p}_\lambda = aV + b \geq 1$ with $a, b \in \mathbb{N}$ and $0 \leq b \leq V - 1$. Since W is in the support of the V-fold weights distribution of Algorithm 1, there is an $\epsilon \in \{0, 1\}$ such that

$$\{ W_i \text{ s.t. } X_i \in I_\lambda \} = \left\{ 0 \text{ repeated } a + \epsilon \text{ times}, \frac{V}{V-1} \text{ repeated } r_\lambda - a - \epsilon \text{ times} \right\} \ .$$

Hence,

$$W_\lambda = 1 + \frac{b - V\epsilon}{(V-1)(aV+b)} \quad \text{and} \quad \sum_{i=1}^{r_\lambda} \left(W_{(i,\lambda)} - W_\lambda \right)^2 \leq L \times \left[1 \vee \frac{n\widehat{p}_\lambda}{V} \right] \ ,$$

so that for every $q \geq 2$,

$$\left\| \mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] \right\|_q^{(\Lambda_m)} \leq LA^2 q \left[\frac{1}{\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \}} \vee \frac{1}{V} \right] \mathbb{E}^{\Lambda_m}\left[p_2(m) \right] \ .$$

The classical link between moment and concentration inequalities (*e.g.* Lemma 8.9 in [Arl07]) gives (49) conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{i,\lambda}$. We can remove this conditioning since the probability bound $1 - n^{-\gamma}$ is deterministic. $\square$

B.6. Expectation of inverses of binomials (proof of Lemma 3). Let $Z \sim \mathcal{B}(n,p)$. By Jensen inequality,

$$e_Z^+ \geq \mathbb{P}(Z>0) = 1-(1-p)^n \geq 1-e^{-np} \ .$$

For the upper bound, define

$$e^0_{\mathcal{L}(Z)} := \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z>0}\right] = e_Z^+\mathbb{P}(Z>0) \ , \tag{52}$$

so that we can focus on $e^0_{\mathcal{B}(n,p)}$.

The bound by κ_4 follows from Lemma 4.1 of [GKKW02], according to which

$$\forall n \in \mathbb{N}, \forall p \in [0,1], \quad e^0_{\mathcal{B}(n,p)} \leq \frac{2np}{(n+1)p} \leq 2 \ . \tag{53}$$

We can now assume that $np \geq A \geq 29.17$ since otherwise, $1+\kappa_3(np)^{-1/4} \geq \kappa_4$. Using that $\mathbb{P}(1>Z>0)=0$, we have for every $\alpha>0$,

$$e^0_{\mathcal{B}(n,p)} = np\mathbb{E}\left[Z^{-1}\mathbb{1}_{\alpha\mathbb{E}[Z]>Z>0}\right] + np\mathbb{E}\left[Z^{-1}\mathbb{1}_{Z\geq\alpha\mathbb{E}[Z]}\right] \leq np\mathbb{P}\left(\alpha np > Z > 0\right) + \alpha^{-1} \ .$$

We now bound the probability on the right-hand side thanks to Bernstein's inequality (*e.g.* Prop. 2.9 of [Mas07]):

$$\forall \theta>0, \quad \mathbb{P}\left(Z \leq \left(1-\sqrt{2\theta}-\frac{\theta}{3}\right)np\right) \leq e^{-\theta np} \ ,$$

and $\theta = A^{-1/2}$. Straightforward computations shows that

$$\sup_{np\geq A}\{e^+_{\mathcal{B}(n,p)}\} \leq \left[\frac{1}{1-\sqrt{2}A^{-1/4}-\frac{1}{3}A^{-1/2}} + Ae^{-\sqrt{A}}\right]\frac{1}{1-e^{-A}} \ ,$$

from which the result follows.

B.7. A technical lemma. Because of the randomness of the design, we have to ensure that the empirical frequencies $n\widehat{p}_\lambda$ are not too far from the expected ones np_λ.

LEMMA 12. *Let $(p_\lambda)_{\lambda\in\Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ a multinomial vector of parameters $(n;(p_\lambda)_{\lambda\in\Lambda_m})$, $\gamma>0$. Assume that* $\mathrm{Card}(\Lambda_m) \leq n$ *and* $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n > 0$. *There is an event of probability at least $1-Ln^{-\gamma}$ on which the following three inequalities hold.*

$$\max_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\mathbb{1}_{\widehat{p}_\lambda>0}\right\} \leq L\times(\gamma+1)\ln(n) \tag{54}$$

$$\min_{\lambda\in\Lambda_m}\left\{\frac{p_\lambda}{\widehat{p}_\lambda}\right\} \geq \frac{1}{2+(\gamma+1)B_n^{-1}\ln(n)} \tag{55}$$

$$\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq \frac{\min_{\lambda\in\Lambda_m}\{np_\lambda\}}{2} - 2(\gamma+1)\ln(n) \tag{56}$$

PROOF OF LEMMA 12. Those three results come from Bernstein's inequality (*e.g.* Prop. 2.9 of [Mas07]) applied to $n\widehat{p}_\lambda$: for every $\lambda \in \Lambda_m$, there is a set of probability $1 - 2n^{-(\gamma+1)}$ on which

$$np_\lambda - \sqrt{2np_\lambda(\gamma+1)\ln(n)} - \frac{(\gamma+1)\ln(n)}{3} \leq n\widehat{p}_\lambda \leq np_\lambda + \sqrt{2np_\lambda(\gamma+1)\ln(n)} + \frac{(\gamma+1)\ln(n)}{3} \quad .$$

For (54), if $np_\lambda \geq 8(\gamma+1)\ln(n)$, the lower bound gives the result. Otherwise, remark only that $(p_\lambda/\widehat{p}_\lambda)\mathbb{1}_{\widehat{p}_\lambda>0} \leq np_\lambda \leq 8(\gamma+1\ln(n)$. For (55), use the upper bound and remark that $np_\lambda(\gamma+1)\ln(n)B_n^{-1} \geq (\gamma+1)\ln(n)$. For (56), use the lower bound and remark that $\sqrt{2np_\lambda(\gamma+1)\ln(n)} \leq (np_\lambda)/2 + (\gamma+1)\ln(n)$. Finally, the union bound gives the result since $\mathrm{Card}(\Lambda_m) \leq n$. □

ACKNOWLEDGMENTS

The author would like to thank gratefully Pascal Massart for several fruitful discussions.

REFERENCES

[ACH99] Marc Aerts, Gerda Claeskens, and Jeffrey D. Hart. Testing the fit of a parametric function. *J. Amer. Statist. Assoc.*, 94(447):869–879, 1999.

[Ada05] Radoslaw Adamczak. Moment inequalities for u-statistics, 2005.

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

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

[Alp99] Ethem Alpaydin. Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. *Neur. Comp.*, 11(8):1885–1892, 1999.

[AM08] Sylvain Arlot and Pascal Massart. Slope heuristics for heteroscedastic regression on a random design. In preparation, 2008.

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

[Bar00] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[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.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[BG04] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105 (electronic), 2004.

[BM06] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields*, 134(3), 2006.

[Bre96] Leo Breiman. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6):2350–2383, 1996.

[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.

[Bur90] Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–345, 1990.

[Bur02] Prabir Burman. Estimation of equifrequency histograms. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

[CR08] Alain Celisse and Stéphane Robin. Non-parametric density estimation by exact leave-p-out cross-validation. *C.S.D.A.*, 2008. To appear.

[CW79] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79.

[Die98] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neur. Comp.*, 10(7):1895–1924, 1998.

[DJ95] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

[DR98] Devdatt Dubhashi and Desh Ranjan. Balls and bins: a study in negative dependence. *Random Structures Algorithms*, 13(2):99–124, 1998.

[Efr79] Bradley Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

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

[EP96] Sam Efromovich and Mark Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996.

[Fro07] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3):165–207, 2007.

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

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GLZ00] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[GP05] Leonid Galtchouk and Sergey Pergamenshchikov. Efficient adaptive nonparametric estimation in heteroscedastic models. Université Louis Pasteur, IRMA, Preprint, 2005.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables, with applications. *Ann. Statist.*, 11(1):286–295, 1983.

[Lew76] Robert A. Lew. Bounds on negative moments. *SIAM J. Appl. Math.*, 30(4):728–731, 1976.

[Li87] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.

[Mal73] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[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.

[MN92] David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3):1611–1624, 1992.

[MSP05] Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[PW93] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[Sha93] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[Sha97] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264, 1997. With comments and a rejoinder by the author.

[Shi81] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[Sto85] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[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.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[Yan06] Yuhong Yang. Comparing learning methods for classification. *Statist. Sinica*, 16(2):635–657, 2006.

[Yan07] Yuhong Yang. Consistency of cross validation for comparing regression procedures. Accepted by Annals of Statistics, 2007.

[Zha93] Ping Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313, 1993.

[Žni05] Marko Žnidarič. Asymptotic expansions for inverse moments of binomial and poisson distributions. arXiv:math.ST/0511226, November 2005.

Sylvain Arlot
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathématiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select
E-mail: sylvain.arlot@math.u-psud.fr