Model selection by resampling penalization

Sylvain Arlot

Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Model selection by resampling penalization

Sylvain Arlot

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

Abstract: We define a new family of resampling-based penalization procedures for model selection in a very general framework. It generalizes several methods (including Efron's bootstrap penalties and the recently proposed leave-one-out penalties, Arlot (2008c)) to any exchangeable weighted bootstrap resampling scheme. In the heteroscedastic regression framework, assuming the models to have a particular structure, we prove that these penalties satisfy a non-asymptotic oracle inequality with a leading constant close to 1. In particular, they are asympotically optimal. We then use these resampling penalties to define an estimator which adapts simultaneously to the smoothness of the regression function and the heteroscedasticity of the noise. This is remarkable because these penalties are general purpose devices, which have not been built specifically to handle heteroscedastic data. We have thus proven that resampling penalties are naturally adaptive to heteroscedasticity. In addition, a simulation study shows that these penalties improve simultaneously V-fold cross-validation, in particular when the signal-to-noise ratio is not large.

AMS 2000 subject classifications: Primary 62G09 ; secondary 62G08, 62M20.
Keywords and phrases: non-parametric statistics, resampling, non-asymptotic, exchangeable weighted bootstrap, model selection, penalization, non-parametric regression, adaptivity, heteroscedastic data, histogram.

1. Introduction

Model selection has received much interest in the last decades. When its final goal is prediction, it can be seen more generally as the question of choosing between the outcomes of several prediction algorithms. With such a general formulation, a very natural (and classical) answer is the following. First, estimate the prediction error for each model (or algorithm). Then, select the model which minimizes this criterion. Model selection procedures mainly differ on the way of making this estimation.

It is natural to think of the empirical risk (also known as the apparent error or the resubstitution error) as an estimator of the prediction error. This can fail dramatically, because it uses the same data for building predictors and for comparing them, making this estimate strongly biased for models involving a number of parameters growing with the sample size.

In order to correct this drawback, *cross-validation* methods (Allen (4), Stone (67)) rely on a data-splitting idea for estimating the prediction error with much less bias. In particular, V-fold cross-validation (VFCV, Geisser (36)) is a popular procedure in practice because it is both general and computationally tractable. There is a huge number of papers about the properties of cross-validation methods, showing that they are efficient for a suitable choice of the way the data is split (or V for VFCV). Asymptotic optimality results of leave-one-out cross-validation (*i.e.* the $V = n$ case) in regression have been proven for instance by Li (49) and Shao (62). However, when V is fixed, VFCV can be asymptotically suboptimal, as shown by Arlot (9). We refer to the latter paper for more references on cross-validation methods, including the small amount of available non-asymptotic results.

Another way to correct the empirical risk for its bias is *penalization*. Basically, it states that a good choice can be made by minimizing the sum of the empirical risk (how do algorithms fit the data) and some complexity measure of the algorithms (called the penalty). This is the case of FPE (Akaike (2)), AIC (Akaike (3)) and Mallows' C_p or C_L (Mallows (51)), to name but a few.

In this article, we aim at defining *efficient* penalization procedures, *i.e.* such that their quadratic risk is asymptotically equivalent to the risk of the oracle. This property is often called *asymptotic optimality*. It does not mean that the procedure finds out a "true model" (which may even not exist), which would be the *consistency* problem. A procedure is efficient when it makes the best possible use of the data in terms of the quadratic risk of the final estimator. This property is desired when the final goal of model selection is *prediction* or *estimation*. According to the previous approach, the ideal penalty for prediction is the difference between the prediction error (the "true risk") and the empirical risk, and penalties should be data-dependent estimates of this quantity.

Many penalties (or complexity measures) have been proposed. Consider for instance regression and least-square estimators on finite-dimensional vector spaces (the models). When the design is fixed and the noise-level constant equal to σ, Mallows' C_p penalty (51) (equal to $2n^{-1}\sigma^2 D$ for a D-dimensional space, and it

can be modified according to the number of models (20; 59)) has some optimality properties (63; 49; 14; 21). However, such a penalty linear in the dimension may fail with heteroscedastic data (as shown by Arlot (6), Chap. 4).

In the binary supervised classification framework, VC-dimension-based penalties have also been proposed. They have the drawback of being independent of the underlying measure, so that they are adapted to the worst case. They have been improved with data-dependent complexity estimates, such as Rademacher complexities (45; 17) (generalized by Fromont with resampling ideas (33)), but these ones may still be too large because they are global complexity measures. The localization idea then led to local Rademacher complexities (18; 46) which are tight estimates of the ideal penalty, but involve huge (and sometimes unknown) constants and may be very difficult to compute in practice. Hence, there is still some need for easy-to-compute margin adaptive penalties.

It appears that all these penalties have serious drawbacks which make them less often used in practice than cross-validation methods: AIC and Mallows' C_p rely on strong assumptions (like homoscedasticity of the data, and linearity of the models) and some mainly asymptotic arguments; global complexities are far too pessimistic; local Rademacher complexities are hard to compute, and even much harder to calibrate. There is another approach for designing penalties in a general framework, which does not have these drawbacks *a priori*, which is *resampling*.

Efron's resampling heuristics (29) (first stated for the bootstrap, then generalized to exchangeable weighted bootstrap by Mason and Newton (54) and Præstgaard and Wellner (58)) basically states that the distribution of any function of the (unknown) law of the data and the sample can be estimated by drawing "resamples" from the initial sample. In particular, this heuristics can be used to estimate the variance of an estimator (Efron (29)), a prediction error (Wu (69), Efron and Tibshirani (32)) or the ideal penalty (Efron (30; 31) and Ishiguro, Sakamoto and Kitagawa (43) with the bootstrap; Shao (61) with the m out of n bootstrap[1]; Arlot (9) with a V-fold subsampling scheme). The asymptotic optimality of Efron's bootstrap penalty for some maximum likelihood estimators has been proven by Shibata (64). Notice also that the aforementioned global and local Rademacher complexities are using an i.i.d. Rademacher scheme for estimating different upper bounds on the ideal penalty, and Fromont's penalties (34) are generalizing the global ones to the exchangeable weighted bootstrap.

The first goal of this paper is to define and study *general-purpose penalties*. This means that they should be well-defined in (almost) every framework, and perform reasonably well in most of them, including regression and classification. The main interest of such penalties would be the ability to face difficult problems (*e.g.* heteroscedastic data, a non-smooth regression function, or the fact that the oracle model attains fast rates of estimation), *without knowing them in advance*. From the practical viewpoint, such a property is crucial.

[1] Notice that Shao's goal was not efficiency but consistency.

For this, we propose to use the aforementioned resampling heuristics, with the general exchangeable weighted bootstrap. This defines a wide family, called "Resampling Penalization" (RP), which includes Efron's and Shao's penalties, as well as the n-fold penalties defined by Arlot (9). To our knowledge, it has never been proposed with such general resampling schemes, so that it contains a wide range of completely new procedures. Notice that RP is well-defined in a very general framework, which includes regression and classification, but also many other possible fields of application. Even if we focus on least-square regression for our proofs, we obviously do not mean that RP should be restricted to this framework.

We investigate the model selection performance (in terms of efficiency) of RP, with a *unified approach* for all the "exchangeable" resampling schemes. For instance, our results make a comparison of the bootstrap and subsampling quite straightforward, which is not common in the resampling literature (except a few asymptotic results, *cf.* Barbe and Bertail (15)).

Our viewpoint is *non-asymptotic*, which has 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. As noticed in (9), with such data, VFCV can have serious drawbacks which can be naturally fixed when using penalization procedures, because they are more flexible. It is worth noticing that such a non-asymptotic approach is not common in the model selection literature, and more generally there are few non-asymptotic results concerning general resampling methods.

Another important point 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. As already mentioned, simple model selection procedures such as Mallows' C_p may not work, and it is natural to hope that resampling methods may be robust to heteroscedasticity. In this article, both theoretical and simulation results confirm this fact.

We mainly prove two kinds of results. First, making very mild assumptions on the distribution of the data, we prove a non-asymptotic oracle inequality for

RP with a leading constant close to 1 (Thm. 1). It is satisfied for several kinds of resampling schemes (including bootstrap, leave-one-out, half-subsampling and i.i.d. Rademacher weighted bootstrap) and implies the asymptotic optimality of RP, even when the data is highly heteroscedastic. For proving such a result, we have to assume that each model is the vector space of piecewise constant functions (histograms) on some partition of the feature space. This is quite a restriction, but we conjecture it to be mainly technical, and that RP stays efficient in a much more general framework. We provide some evidence for this in Sect. 8.3. Another reason for studying extensively the toy model of histograms is that we can use it to derive heuristics for the general case. Our main goal here is to help practical users, who would like to know how to use resampling for performing model selection.

Second, we use RP to build an estimator simultaneously adaptive to the smoothness of the regression function (assuming that it is α-hölderian for some unknown $\alpha \in (0,1]$) and to the unknown noise-level $\sigma(\cdot)$ (Thm. 2). This result may seem surprising since RP has never been designed specifically for such a purpose. We interpretate this as another evidence that RP is *naturally adaptive*, and should work well in several other difficult frameworks.

As already noticed, several similar results for other algorithms exist in the literature, for instance for Mallows' C_p (with homoscedastic data only), VFCV and leave-one-out cross-validation. In addition, there exists several minimax adaptive estimators for heteroscedastic data (assuming only that the noise-level is smooth enough), such as the ones of Efromovich and Pinsker (28) or Galtchouk and Pergamenschikov (35). The interest of RP is both its *generality* (contrary to Mallows' C_p and specific adaptive estimators) and its *flexibility* (contrary to VFCV, see (9)).

We conduct a simulation experiment (Sect. 6) with small sample sizes. RP is shown to be competitive with Mallows' C_p for "easy" problems, and much better for some harder ones (*e.g.* with a variable noise-level). On the other hand, a well-calibrated RP has almost always better performances than classical VFCV. Thus, RP may be of great interest in situations where no *a priori* information is known about the data. It is an efficient alternative to VFCV, which is able to deal with difficult problems, while being close to the best procedures that are fitted for easier problems.

This article is organized as follows. The general Resampling Penalization algorithm (RP) is defined in Sect. 2. We focus on the histogram regression case in Sect. 3, for which we state our main theorems in Sect. 4. The influence of the resampling weights on the performance of the procedure is theoretically investigated in Sect. 5. We then present a simulation experiment in Sect. 6. The practical implementation of RP is considered in Sect. 7. We discuss the strengths and weakness of RP in Sect. 8. Finally, Sect. 9 is devoted to the proofs. Notice also that some additional material (other simulation experiments and proofs) has been reported into a technical appendix (7).

2. A general model selection algorithm

2.1. Setting

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 i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ 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 (or prediction error) $P\gamma(t) := \mathbb{E}_{(X,Y)\sim P}[\gamma(t,(X,Y))]$. 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), it is equivalent to consider the excess loss $\ell(s,t) = P\gamma(t) - P\gamma(s) \geq 0$ 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*). A natural method for choosing some data-dependent $\widehat{m} \in \mathcal{M}_n$ is to minimize over $\mathcal{M}_n$ a criterion $\mathrm{crit}(m)$ which estimates the true prediction loss $P\gamma(\widehat{s}_m(P_n))$. The more natural criterion may then be the re-substitution error $P_n\gamma(\widehat{s}_m(P_n))$, but it underestimates the true prediction loss, in particular from the non-asymptotic viewpoint where the complexity of the algorithms $\widehat{s}_m$ is allowed to increase with n. This leads to the penalization idea, which is to correct this bias by adding some data-dependent complexity measure $\mathrm{pen}(m)$ to the re-substitution error. In other words, we define

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ P_n\gamma(\widehat{s}_m(P_n)) + \mathrm{pen}(m) \} \quad , \tag{1}$$

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

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma(\widehat{s}_m(P_n)) \quad , \tag{2}$$

and we would like $\mathrm{pen}(m)$ to be as close to $\mathrm{pen}_{\mathrm{id}}(m)$ as possible for every $m \in \mathcal{M}_n$.

In such a general setting, it is natural to think that resampling may be used to estimate the ideal penalty.

2.2. The resampling heuristics

Let us recall briefly the *resampling heuristics*, which has been introduced by Efron (29) in the context of variance estimation. Basically, it 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 $\mathbb{E}_W\left[F(P_n, P_n^W)\right]$. The expectation $\mathbb{E}_W[\cdot]$ means that we integrate w.r.t. the resampling randomness only. Let us emphasize that $\mathrm{pen}_{\mathrm{id}}(m)$ has this form.

Later on, this heuristics has been generalized to other resampling schemes, with the exchangeable weighted bootstrap (Mason and Newton (54), Præstgaard and Wellner (58)). 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), and such that $\forall i$, $\mathbb{E}[W_i] = 1$. In this article, we will also assume that $\forall i$, $W_i \geq 0$ a.s. and $\mathbb{E}[W_i^2] < \infty$.

We mainly consider the following weights, which include the more classical resampling schemes:

1. *Efron* (m), $m \in \mathbb{N}\backslash\{0\}$ (Efr): $((m/n)W_i)_{1\leq i\leq n}$ is a multinomial vector with parameters $(m; n^{-1}, \ldots, n^{-1})$. A classical choice is $m = n$.
2. *Rademacher* (p), $p \in (0;1)$ (Rad): (pW_i) are independent, with a Bernoulli (p) distribution. A classical choice is $p = 1/2$.
3. *Poisson* (μ), $\mu \in (0, \infty)$ (Poi): (μW_i) are independent, with a Poisson (μ) distribution. A classical choice is $\mu = 1$.
4. *Random hold-out* (q), $q \in \{1, \ldots, n\}$ (Rho): $W_i = (n/q)\mathbb{1}_{i\in I}$ with I uniform random subset (of cardinality q) of $\{1, \ldots, n\}$. A classical choice is $q = n/2$.
5. *Leave-one-out* (Loo) = Rho $(n-1)$.

In the following, when the parameter is not mentioned, it has its "classical" value.

Remark 1. The terminology above is made to give explicit links with some classical resampling schemes. See (54; 39; 68) for more details about classical resampling weights names, as well as other classical examples.

- The name "Efron" comes from the classical choice $m = n$ for which Efron weights actually are Efron's bootstrap weights. When $m < n$, this is the m out of n bootstrap, used for instance by Shao (61).
- The name "Rademacher" for the i.i.d. Bernoulli weights comes from the classical choice $p = 1/2$ for which $(W_i - 1)_i$ are i.i.d. Rademacher random variables. Using this resampling scheme to estimate some upper bound on $\mathrm{pen}_{\mathrm{id}}(m)$ can lead to either global or local Rademacher complexities.
- Poisson weights are often used as approximations to Efron weights, via the so-called "Poissonization" technique (*cf.* (68), Chap. 3.5, and (33)). They are known to be efficient for estimating several non-smooth functionals (Barbe and Bertail (15), Chap. 3; see also Mammen (52), Sect. 1.4).

- The Random hold-out (q) weights can also be called "delete-$(n-q)$ jackknife", as well as the Leave-one-out weights also refer to the jackknife (sometimes called cross-validation). They are both resampling schemes without replacement (see Ex. 3.6.14 in (68)), more often called *subsampling weights* (see *e.g.* Politis, Romano and Wolf (57) on subsampling). They are thus very close to the idea of splitting the data into a training set and a validation set (*e.g.* leave-one-out, hold-out and cross-validation). Indeed, if one defines the training set as

$$\{ (X_i, Y_i) \text{ s.t. } W_i \neq 0 \}$$

and the validation set as its complementary, there is a one-to-one correspondence between the two ideas.

Applying the resampling heuristics (with any exchangeable weight, for instance one of the above) to estimate the ideal penalty (defined by (2)), we get

$$\mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) - P_n^W \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) \right] \quad . \tag{3}$$

As explained in introduction, we would like this quantity to be an unbiased estimated of $\mathrm{pen}_{\mathrm{id}}(m)$ for each model m. However, the asymptotic theory of exchangeable bootstrap empirical processes (for instance, Theorem 3.6.13 of van der Vaart and Wellner (68)) shows that this does not hold for general W, although it does in the bootstrap case (*i.e.* $\mathrm{Efr}(n)$ weights). More precisely, it is asymptotically unbiased if and only if $\forall i$, $\mathbb{E}\left(W_i - 1\right)^2 \approx 1$, which holds for several classical weights: $\mathrm{Efr}(n)$, $\mathrm{Rad}(1/2)$, $\mathrm{Poi}(1)$ and $\mathrm{Rho}(n/2)$, but not in the general case. Intuitively, this condition means that the variability of the weights has to be of the right order, so that $P_n - P_n^W$ mimics $P - P_n$. Otherwise, one may have to multiply the resampling estimate (3) by a constant $C_W > 0$.

In the framework developped in Sect. 3, it will turn out that

$$C_W \sim_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(W_i - 1\right)^2 \right)^{-1}$$

does not depend on the model m, and we will give non-asymptotic expressions for it (see Tab. 1). In the general case, we suggest to use a data-driven method to estimate C (which is discussed in Sect. 7.3.1), whereas the resampling penalty only estimates the shape of the ideal one.

2.3. Resampling penalization

We are now in position to make explicit the resampling penalization algorithm, in the general framework.

Algorithm 1 (Resampling penalization).

1. Choose a resampling scheme, *i.e.* the law $\mathcal{L}(W)$ of a weight vector W.

2. Choose a constant $C \geq C_W$.
3. Compute the following resampling penalty for each $m \in \mathcal{M}_n$:
$$\mathrm{pen}(m) = C\mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) - P_n^W \gamma \left(\widehat{s}_m \left(P_n^W \right) \right) \right] \quad . \tag{4}$$
4. Choose $\widehat{m}$ according to (1).

Remark 2.
1. Algorithm 1 above is a generalization of several model selection procedures. With a bootstrap resampling scheme and $C = 1$, it is Efron's bootstrap penalty (30). In the log-likelihood framework, it has also been called the EIC procedure by Ishiguro, Sakamoto and Kitagawa (43). With a m out of n bootstrap resampling scheme and $C = 1$, it has been proposed and studied by Shao (61) in the context of model identification (notice that $C_W \neq 1$ for $\mathrm{Efr}(m)$ weights if $m \neq n$; this is a crucial point that we will discuss later). A (non-exchangeable) V-fold subsampling scheme has also been proposed recently in (9), as an alternative to V-fold cross-validation.
2. When W are the "leave-one-out" weights, RP is not identical to the classical cross-validation model selection algorithm. However, according to (9), when $C = n - 1$, it is identical to Burman's n-fold corrected cross-validation (23), hence very close to the uncorrected one.
3. We allow C to be larger than C_W because overpenalizing may be fruitful in a non-asymptotic viewpoint, *e.g.* when there is few noisy data. The simulation study of Sect. 6 provides experimental evidence for this fact.
4. Since $\widehat{m}$ is computed through a plug-in method, Algorithm 1 seems to be reasonable only if $\mathcal{M}_n$ is not too large. Otherwise, Birgé and Massart (21) have proved that penalization procedures have to be enlarged, taking into account the complexity of $\mathcal{M}_n$.

In order to study Algorithm 1 from the non-asymptotic viewpoint, it is crucial to compare the expectation of the resampling penalty (4) with the one of the ideal penalty (2). This is a quite hard problem in general, since asymptotic results can no longer be used. This is why, in the rest of the paper, we restrict ourselves to least-square regression on histogram models. Notice that we do not consider histograms as a final goal, but only a first theoretical step, from which we can derive heuristics concerning Algorithm 1 in general. We describe this framework in the next section.

3. The histogram regression case

3.1. Setting

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
$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{5}$$
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.

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 coincides with the quadratic risk: $\ell(s,t) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2$. To each model S_m, we associate the *empirical risk minimizer*

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

(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:

$$s_m = \sum_{\lambda\in\Lambda_m}\beta_\lambda\mathbb{1}_{I_\lambda} \quad \text{and} \quad \widehat{s}_m = \sum_{\lambda\in\Lambda_m}\widehat{\beta}_\lambda\mathbb{1}_{I_\lambda} \ ,$$

$$\text{where} \quad \beta_\lambda := \mathbb{E}\left[Y \mid X\in I_\lambda\right] \quad \widehat{\beta}_\lambda := \frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda} Y_i \quad \widehat{p}_\lambda := P_n(X\in I_\lambda) \ .$$

Similarly, we can compute the resampling versions of these quantities

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

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

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.* $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$. For this reason, we remove models for which $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda = 0$ (for instance by adding $+\infty\mathbb{1}_{\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda=0}$ to the penalty).

3.2. A modified algorithm for histograms

Assuming that $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$ and defining $p_\lambda := P(X\in I_\lambda)$, we can compute the ideal penalty (see (46) and (47) in Sect. 9.7):

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

Hence, the resampling penalty defined by (4) can be written

$$\begin{aligned}\mathrm{pen}(m) &= \mathbb{E}_W\left[(P_n-P_n^W)\gamma(\widehat{s}_m^W)\right] \\ &= \sum_{\lambda\in\Lambda_m}\mathbb{E}_W\left[\left(\widehat{p}_\lambda+\widehat{p}_\lambda^W\right)\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[(P_n-P_n^W)\gamma(\widehat{s}_m)\right] \\ &= \sum_{\lambda\in\Lambda_m}\left(\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\right]\right) \qquad (6)\end{aligned}$$

$\mathcal{L}(W)$	Efr(m)	Rad(p)	Poi(μ)	Rho(q)	Loo
C_W	m/n	$p/(1-p)$	μ	$q/(n-q)$	$n-1$

TABLE 1
*C_W for several resampling schemes (*cf. *Sect. 4.4.1).*

since $\mathbb{E}[W_i] = 1$ for every i implies $\mathbb{E}_W\left[(P_n - P_n^W)\gamma(\widehat{s}_m)\right] = 0$. Unfortunately, this penalty (6) is well-defined if and only if $\widehat{s}_m^W$ is a.s. uniquely defined, *i.e.* $W_\lambda > 0$ for every $\lambda \in \Lambda_m$. This can not be assumed for general resampling schemes, considering that

$$\mathbb{P}\left(\forall i \geq 2,\, W_i = 0\right) > 0$$

for most of them, including the bootstrap. We then have to modify Algorithm 1 in the histogram regression case.

Algorithm 2 (Resampling penalization for histograms).

0. Replace $\mathcal{M}_n$ by
$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathcal{M}_n \text{ s.t. } \min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\} \geq 3 \right\} \quad .$$
1. Choose a resampling scheme $\mathcal{L}(W)$.
2. Choose a constant $C \geq C_W$ where C_W is defined in Tab. 1.
3. Define, for each $m \in \widehat{\mathcal{M}}_n$, the resampling penalty $\mathrm{pen}(m)$ as
$$C \sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W\left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0 \right] + \mathbb{E}_W\left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \right] \right) \quad . \tag{7}$$
4. Choose $\widehat{m} \in \arg\min_{m\in\widehat{\mathcal{M}}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m) \right\}$.

Remark 3.
1. At step 0, we remove more models than those for which $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda = 0$. When $n\widehat{p}_\lambda = 1$, it is clear that one can not estimate the quality of estimation of $\widehat{\beta}_\lambda$, which comes from only one observation, without making further assumptions on the noise-level σ. Hence, we can not hope that resampling will be able to estimate it directly. The reason why we choose to remove also models for which $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda = 2$ is that the oracle inequalities of Sect. 4 require it for several instances of weights (but not all of them). Notice also that such models have very poor prediction performance in general, because making predictions from only two observations is quite hazardous. This is why step 0 is very reasonable.
2. Several other conventions may have been chosen at step 3, in order to deal with the uniqueness issue. With (7), we have chosen to restrict to weight vectors such that $W_\lambda > 0$, separately for each $\lambda \in \Lambda_m$ and each $m \in \mathcal{M}_n$. Together with step 0, this avoids to underestimate the penalty pen when $\min_{\lambda\in\Lambda_m}\left\{ n\widehat{p}_\lambda \right\}$ is small. We refer to Sect. 8.1 of (6) and to (9) for further considerations about this issue.

4. Main results

In this section, we give some non-asymptotic properties of Resampling Penalization (Algorithm 2) for model selection. The first one is an oracle inequality, with leading constant close to 1. In particular, it implies the asymptotic optimality of RP. The second one is an adaptivity result, when the regression function is assumed to belong to some Hölderian ball. A quite interesting point is that both results stay valid with very mild assumptions on the distribution of the noise, which is allowed to be non-gaussian and highly heteroscedastic.

Throughout this section, 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 weight vector W is chosen among Efr, Rad, Poi, Rho and Loo (defined in Sect. 2.2).

4.1. Oracle inequality

Theorem 1. *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*
$$C_{\text{b}}^- D_m^{-\beta_1} \leq \ell\left(s, s_m\right) \leq C_{\text{b}}^+ D_m^{-\beta_2} \quad .$$
- ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{L}(X)$*:* $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\text{r},\ell}^X$.

Let $\widehat{m}$ be the model chosen by Algorithm 2 (under restrictions (**P1** − **3**)*, with $C = C_W$). Then, there exists a constant $K_1 > 0$ and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_1 n^{-2}$,*

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left(1 + \epsilon_n\right) \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s, \widehat{s}_m\right) \right\} \quad . \tag{8}$$

Moreover, we have

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right) \right] \leq \left(1 + \epsilon_n\right) \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s, \widehat{s}_m\right) \right\} \right] + \frac{A^2 K_1}{n^2} \quad . \tag{9}$$

The constant K_1 may depend on constants in (**Ab**)*,* (**An**)*,* (**Ap**)*,* ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *and* (**P1** − **3**)*, but not on n. The term ϵ_n is smaller than $\ln(n)^{-1/5}$; 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_1.*

The non-asymptotic oracle inequality (8) implies that Algorithm 2 is *a.s. asymptotically optimal* in this framework, when $C \sim_{n\to\infty} C_W$. This means that if $\mathcal{M}_n$ contains a model that takes well into account the smoothness of s and the shape of the noise $\sigma(X)$, the Resampling Penalization algorithm does as well as

this oracle model for estimation. Since this does not require any knowledge about the smoothness of s, the heteroscedasticity of σ or any other property satisfied by P, it is a *naturally adaptive algorithm*. We give in the next subsection a framework where Resampling Penalization can be used to build an adaptive estimator.

Notice that (8) is somehow stronger than an adaptivity result because of the leading constant close to one (whereas estimators are said "adaptive" when they attain the correct estimation rate up to a constant independent from n, but possibly much greater than one). Hence, once $\mathcal{M}_n$ is well chosen (and our assumptions on $\mathcal{M}_n$ do not forbid this), one can hope that RP leads to an adaptive estimator which should be close to the optimal ones.

We now make some comments concerning the assumptions of Thm. 1:

1. When $C \in [C_W; \eta C_W]$ for some $\eta > 1$, the same result holds with a leading constant $2\eta - 1 + \epsilon_n$ instead of $1 + \epsilon_n$ in (8) and (9).
2. With (**P3**), we restrict ourselves to the five "classical" exchangeable weights of Sect. 2.2. What we really need is that (1) W is exchangeable, (2) $R_{1,W}(n,p)+R_{2,W}(n,p) \approx 2C_W$ for np large enough (with a non-asymptotic control on the ratio between these two quantities, as in the proof of Prop. 2), and (3) $R_{1,W}(n,p) + R_{2,W}(n,p) > (1+\epsilon)C_W$ for some $\epsilon > 0$, as soon as $np \geq T \geq 2$ (as in Lemma 11). Then, replacing the threshold 3 by T at step 0 of Algorithm 2, the results of Thm. 1 still hold.
   The first two conditions hold for all the exchangeable weights considered in Prop. 2. The third one is satisfied for most of them, as soon as T is large enough (see Remark 11 below Lemma 11). They certainly also hold for several other resampling schemes, to which Thm. 1 would then be extended.
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, as explained in Sect. 4.3.
4. When X has a lower bounded density w.r.t. Leb, ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied for "almost piecewise regular" histograms, including all those considered in the simulation study of Sect. 6.
5. The upper bound in (**Ap**) holds when $(I_\lambda)_{\lambda \in \Lambda_m}$ is regular on $\mathcal{X} \subset \mathbb{R}^k$ and s α-hölderian with $\alpha > 0$, with $\beta = 2\alpha k^{-1}$.

To finish with, let us comment more extensively the lower bound in (**Ap**), which may seem unintuitive at first sight. Indeed, it means that s is not too well approximated by the models S_m. Notice that it is classical to assume that $\ell(s, s_m) > 0$ for every $m \in \mathcal{M}_n$, for proving the asymptotic optimality of Mallows' C_p (*cf.* Shibata (63), Li (49) and Birgé and Massart (21)). Moreover, the stronger assumption (**Ap**) has already been made by Stone (66) and Burman (24) in the density estimation framework, for the same technical reasons as ours.

Let us now explain why we use it in our proof. According to Remark 9 in Sect. 9.2, when the lower bound in (**Ap**) is no longer assumed, (8) holds with two modifications in its right-hand side: the inf is restricted to models of dimension

larger than $\ln(n)^{\gamma_1}$, and there is a remainder term $\ln(n)^{\gamma_2} n^{-1}$ (where γ_1 and γ_2 are numerical). This is essentially the same as (8), unless there is a model of small dimension with a very small bias, and the lower bound in (**Ap**) is sufficient to ensure that this do not happen. Notice that if there is such a very small model very close to s, it is hopeless to obtain an oracle inequality with a penalty which estimates $\mathrm{pen}_{\mathrm{id}}$, simply because deviations of $\mathrm{pen}_{\mathrm{id}}$ around its expectation would be much larger than the excess loss of the oracle. In such a situation, BIC-like methods are more appropriate.

Another argument in favour of (**Ap**) is that it is not too strong, because it is at least satisfied in the following case: $(I_\lambda)_{\lambda \in \Lambda_m}$ is "regular" (as defined in Algorithm 3 below), X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$, and s is non-constant and α-hölderian (w.r.t. $\|\cdot\|_\infty$), with

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

We refer to Sect. 8.10 in (6) for more details about this claim (including complete proofs).

We finally mention that this is not the only case where (**Ap**) holds. In particular, an important point is that Thm. 1 does not need the distribution of X to satisfy any assumption, once the models are wisely chosen according to this distribution. As a consequence, if one has some prior knowledge on the distribution of X (*e.g.* thanks to unlabeled data), it is always possible to choose a collection of models which satisfy (**P1**), (**P2**) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$), and has good approximation properties w.r.t. the distribution of X (*i.e.* (**Ap**) is satisfied, uniformly over functions s belonging to some appropriate approximation space). Hence, the general formulation of (**Ap**) is crucial to make Thm. 1 valid *whatever the distribution of* X.

4.2. Adaptivity to the hölderian regularity

As noticed in the previous subsection, a natural framework in which Thm. 1 can be applied is when $\mathcal{X}$ is a compact subset of $\mathbb{R}^k$, X has a lower bounded density w.r.t. the Lebesgue measure and s is α-hölderian with $\alpha \in (0,1]$. Indeed, the latter condition on α ensures that regular histograms have good approximation properties. In this subsection, we investigate how these facts can be combined with Thm. 1 to build an adaptive estimator through Resampling Penalization. We first define the estimator that is likely to be adaptive to α. For the sake of simplicity[2], let us assume that $\mathcal{X}$ is a closed ball of $(\mathbb{R}^k, \|\cdot\|_\infty)$, say $[0,1]^k$.

Algorithm 3 (Resampling penalization with regular histograms).
For every $T \in \mathbb{N} \setminus \{0\}$, denote by $S_{m(T)}$ is the model of regular[3] histograms

[2] If $\mathcal{X}$ has a smooth boundary, Algorithm. 3 can be modified so that our proof of Thm. 2 stays valid.

[3] When $\mathcal{X}$ has a general shape, it is any partition $(I_\lambda)_{\lambda \in \Lambda_m}$ such that $\mathrm{Leb}(I_\lambda) \propto T^{-k}\,\mathrm{Leb}(\mathcal{X})$ and $\mathrm{diam}(I_\lambda) \propto T^{-1}$ for every $\lambda \in \Lambda_m$.

with T^k bins:

$$(I_\lambda)_{\lambda\in\Lambda_m} := \left(\prod_{i=1}^{k}\left[T^{-1}j_i; T^{-1}(j_i+1)\right)\right)_{0\leq j_1,\ldots,j_k\leq T-1} .$$

Then, define $(S_m)_{m\in\mathcal{M}_n} := \left(S_{m(T)}\right)_{1\leq T\leq n^{1/k}}$.

0. Replace $\mathcal{M}_n$ by

$$\widehat{\mathcal{M}}_n = \left\{ m\in\mathcal{M}_n \text{ s.t. } \min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\}\geq 3 \right\} .$$

1. Choose a resampling scheme $\mathcal{L}(W)$ among Efr, Rad, Poi, Rho and Loo.
2. Take the constant $C = C_W$ as defined in Tab. 1.
3. For each $m\in\widehat{\mathcal{M}}_n$, compute $\mathrm{pen}(m)$ defined by (7).
4. Choose $\widehat{m}\in\arg\min_{m\in\widehat{\mathcal{M}}_n}\{P_n\gamma(\widehat{s}_m)+\mathrm{pen}(m)\}$.
5. Define $\widetilde{s} := \widehat{s}_{\widehat{m}}$.

Theorem 2. *Let $\mathcal{Y}\subset\mathbb{R}$ and $\mathcal{X}=[0,1]^k$. Assume that the (X_i,Y_i)'s satisfy*

(**Ab**) *Bounded data: $\|Y_i\|_\infty\leq A<\infty$.*
(**An**) *Noise-level bounded from below: $\sigma(X_i)\geq\sigma_{\min}>0$ a.s.*
(**Ad$_\ell$**) *Density bounded from below:*

$$\exists c_X^{\min}>0,\quad \forall I\subset\mathcal{X},\qquad P(X\in I)\geq c_X^{\min}\mathrm{Leb}(I) .$$

(**Ah**) *Hölderian regression function: there exists $\alpha\in(0;1]$ and $R>0$ s.t.*

$$s\in\mathcal{H}(\alpha,R)\quad \text{i.e.}\quad \forall x_1,x_2\in\mathcal{X},\ |s(x_1)-s(x_2)|\leq R\|x_1-x_2\|_\infty^\alpha .$$

Let $\widetilde{s}$ be the estimator obtained through Algorithm 3. Define $\sigma_{\max}=\sup_{\mathcal{X}}|\sigma|\leq 2A$. Then, there exists positive constants K_2 and K_3 such that,

$$\mathbb{E}\left[\ell(s,\widetilde{s})\right]\leq K_2 R^{\frac{2k}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+k}} + K_3 A^2 n^{-2} . \tag{10}$$

If moreover the noise-level is smooth:

(**Aσ**) *σ is piecewise K_σ-Lipschitz with at most J_σ jumps,*

then, (10) *holds with $\|\sigma\|_{L^2(\mathrm{Leb})}$ instead of $\sigma_{\max}$, and assumption* (**An**) *can be removed.*

For both results, K_2 depends only on α and k. The constant K_3 depends only on k, η, constants in (**Ab**), (**An**), (**Ad$_\ell$**)*and* (**Ah**) *(and* (**Aσ**) *for the last result).*

We now compare the upper bounds given by Thm. 2 with classical minimax lower bounds on the estimation of functions in $\mathcal{H}(\alpha,R)$ with $\alpha\in(0,1]$. In the homoscedastic case, lower bounds have been proven by Stone (65), and generalized by several authors, such as Korostelev and Tsybakov (47) and Yang

and Barron (72), to name but a few. It then appears that the best attainable rate is

$$R^{\frac{2k}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \sigma^{\frac{4\alpha}{2\alpha+k}} \quad ,$$

up to a positive multiplicative constant independent from n, R and σ. This shows that the upper bound (10) attains the right estimation rate in terms of n, R and σ, without using the knowledge of α, R or σ.

Moreover, (10) is still valid in a wide heteroscedastic framework, without using the knowledge of the shape of the noise-level σ. Then, up to a multiplicative constant independent from n and R (but possibly of the order of a power of the ratio between $\sigma_{\max}$ and $\sigma_{\min}$), the upper bound (10) is the best possible estimation rate. Minimax lower bounds have also been proven in the heteroscedastic case (see *e.g.* Efromovich and Pinsker (28), Galtchouk and Pergamenschikov (35) and references therein), showing in particular that when $k = \alpha = 1$ and the noise-level is smooth enough, the best attainable estimation rate depends on σ through the multiplicative factor $\|\sigma\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+k}}$. This shows that the upper bound given by Thm. 2 under the assumption $(\mathbf{A}\sigma)$ is tight, even through its dependence on the noise-level. Up to our best knowledge, such an upper bound have never been obtained when $\alpha \in (0,1)$ and $k > 1$, even with estimators using the knowledge of α, σ and R.

Theorem 2 thus shows that Algorithm 3 defines an *adaptive estimator*, uniformly over distributions such that s belongs to some hölderian ball $\mathcal{H}(\alpha, R)$ with $\alpha \in (0,1]$ and the noise-level σ is not too pathological. This is a quite strong result. Of course, similar properties have already been proven for other estimators (see *e.g.* Efromovich and Pinsker (28), Galtchouk and Pergamenschikov (35)). The main difference is that *Resampling Penalization has not been designed specifically to have such a property*, contrary to these "*ad hoc*" estimators (see Sect. 8.1.4 for further comments). This shows that exchangeable resampling penalties are *naturally adaptive* to these features of the data. As we discuss in Sect. 8.3, this make us conjecture that these penalties are adaptive in several other frameworks, including difficult problems (heteroscedasticity of the data being a major issue in practice).

Remark 4.

1. The proof of Thm. 2 actually gives a stronger result, which is that $\widehat{s}_{\widehat{m}}$ attains the minimax rate of estimation on a set of probability larger than $1 - K_3' n^{-2}$. In particular, with probability one,

$$\limsup_{n\to\infty} \left(\ell\left(s, \widetilde{s}\right) R^{\frac{-2k}{2\alpha+k}} n^{\frac{2\alpha}{2\alpha+k}} \|\sigma\|_{L^2(\mathrm{Leb})}^{\frac{-4\alpha}{2\alpha+k}} \right) \leq K_2(\alpha, k) \quad .$$

2. If s is only piecewise α-hölderian, with at most J_s jumps (of height bounded by $2A$), then the same results hold, with K_3 depending also on J_s.
3. As for Thm. 1, the boundedness of the data and the lower bound on the noise level can be replaced by other assumptions. See Sect. 4.3 below for more details.

4.3. Alternative assumption sets

Actually, Thm. 1 and 2 are corollaries of a more general result, called Lemma 7 in Sect. 9.2. Then, the assumptions we make, in particular (**Ab**) and (**An**) on the distribution of the noise $\sigma(X)\epsilon$, are only sufficient conditions so that the assumptions of Lemma 7 hold true. We give in this subsection some alternative sufficient conditions.

On the one hand, one can remove (**An**): $\sigma(X) \geq \sigma_{\min} > 0$, by adding instead that $\mathcal{X} \subset \mathbb{R}^k$ is bounded, equipped with $\|\cdot\|_\infty$, $\mathbb{E}\left[\sigma(X)^2\right] > 0$ and

($\mathbf{Ar}_\mathbf{u}^\mathbf{d}$) Upper regularity of the partitions: $\exists c_{\mathrm{r,u}}^d, \alpha_d > 0$ such that

$$\forall m \in \mathcal{M}_n, \quad \max_{\lambda \in \Lambda_m} \left\{ \operatorname{diam}(I_\lambda) \right\} \leq c_{\mathrm{r,u}}^d D_m^{-\alpha_d} \ .$$

($\mathbf{Ar}_\mathbf{u}$) Upper regularity of the partitions for Leb: $\exists c_{\mathrm{r,u}} > 0$ such that

$$\forall m \in \mathcal{M}_n, \quad \max_{\lambda \in \Lambda_m} \left\{ \operatorname{Leb}(I_\lambda) \right\} \leq c_{\mathrm{r,u}} D_m^{-1} \ .$$

($\mathbf{A}\sigma$) σ is piecewise K_σ-Lipschitz with at most J_σ jumps.

On the other hand, the boundedness assumption (**Ab**) can be removed, at the price of adding the following: $\mathcal{X} \subset \mathbb{R}$ is bounded measurable and

($\mathbf{A}_{\mathrm{gauss}}$) The noise is sub-gaussian: there exists $c_{\mathrm{gauss}} > 0$ such that

$$\forall q \geq 2, \forall x \in \mathcal{X}, \quad \mathbb{E}\left[|\epsilon|^q \mid X = x \right]^{1/q} \leq c_{\mathrm{gauss}} \sqrt{q} \ .$$

($\mathbf{A}\sigma_{\max}$) Noise-level bounded from above: $\sigma^2(X) \leq \sigma_{\max}^2 < +\infty$ a.s.

($\mathbf{As}_{\max}$) Bound on the target function: $\|s\|_\infty \leq A$.

(**Al**) s is B-Lipschitz, piecewise C^1 and non-constant: $\pm s' \geq B_0 > 0$ on some interval $J \subset \mathcal{X}$ with $\operatorname{Leb}(J) \geq c_J > 0$.

($\mathbf{Ar}_{\ell,\mathbf{u}}$) Regularity of the partitions for Leb: $\exists c_{\mathrm{r},\ell}, c_{\mathrm{r,u}} > 0$ such that

$$\forall m \in \mathcal{M}_n, \forall \lambda \in \Lambda_m, \quad c_{\mathrm{r},\ell} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \leq \operatorname{Leb}(I_\lambda) \leq c_{\mathrm{r,u}} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \ .$$

($\mathbf{Ad}_\ell$) Density bounded from below: $\exists c_X^{\min} > 0, \forall I \subset \mathcal{X}, \mathbb{P}(X \in I) \geq c_X^{\min} \operatorname{Leb}(I) \operatorname{Leb}(\mathcal{X})^{-1}$.

Notice that we here keep the uniform lower bound on the noise (**An**).

Finally, it is possible to remove simultaneously (**An**) and (**Ab**). See (7) for more details.

The above results mean that Thm. 1 holds for most "reasonably" difficult problems. Actually, the proof of Thm. 1 (Prop. 3 and Remark 8) shows that the resampling penalties are much closer to $\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m)\right]$ than $\operatorname{pen}_{\mathrm{id}}(m)$ itself, provided that our concentration inequalities are tight. As a consequence, the (ideal) penalization algorithm which uses $\operatorname{pen}(m) = \mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m)\right]$ as a penalty does not outperform much the resampling penalization algorithm. Up to differences within ϵ_n (we assumed here that $C = C_W$), they perform equally well on a set of probability $1 - K_1 n^{-2}$.
Hence, for every assumption set such that our proof gives an oracle inequality for the penalty $\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m)\right]$, the same proof gives a very similar oracle inequality for the resampling penalties.

4.4. Probabilistic tools

The main results stated in this section rely on several probabilistic tools, which may be of self-interest: accurate computation of the expectations and concentration inequalities for the resampling penalties, and bounds on expectations of the inverses of several classical random variables. Their originality comes from their non-asymptotic nature: we provide explicit bounds on the deviations or the remainder terms for finite sample sizes.

4.4.1. Expectations of resampling penalties

Using only the exchangeability of the weights, we are able to compute the resampling penalty explicitly (Lemma 8 in Sect. 9.7). This can be used to compare its expectation to the one of the ideal penalty. We start with a result valid for *general exchangeable weights*.

Proposition 1. *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$, and $W \in [0,\infty)^n$ be an exchangeable random vector independent from the data. Denote by $\mathbb{E}^{\Lambda_m}[\cdot]$ expectations conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$. Then, defining $\mathrm{pen}_{\mathrm{id}}(m)$ by (2) and $\mathrm{pen}(m)$ by (7), if $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$,*

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(1+\frac{p_\lambda}{\widehat{p}_\lambda}\right)\sigma_\lambda^2 \tag{11}$$

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

$$\text{with}\quad \sigma_\lambda^2 := \mathbb{E}\left[\left(Y-s(X)\right)^2 \,\middle|\, X\in I_\lambda\right] \tag{13}$$

$$\text{and}\quad R_{1,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1-W_\lambda)^2}{W_\lambda^2}\,\middle|\, X_1\in I_\lambda, W_\lambda>0\right] \tag{14}$$

$$R_{2,W}(n,\widehat{p}_\lambda) = \mathbb{E}\left[\frac{(W_1-W_\lambda)^2}{W_\lambda}\,\middle|\, X_1\in I_\lambda\right]\ . \tag{15}$$

In particular,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(2+\delta_{n,p_\lambda}\right)\sigma_\lambda^2 \tag{16}$$

where $\delta_{n,p}$ only depends on (n,p) and is small when the product np is large: there is a numerical constant L_1 such that $|\delta_{n,p}| \leq L_1(np)^{-1/4}$.

Remark 5.
- In order to make the expectation in (16) well-defined, we have to take a convention for $\mathrm{pen}_{\mathrm{id}}(m)$ when $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda = 0$. See Sect. 9.1 and 9.7 for more details.
- Combining Prop. 1 with Lemma 8.4 of (6), a similar result holds for non-exchangeable weights (with only a modification of the definitions of $R_{1,W}$ and $R_{2,W}$.

$\mathcal{L}(W)$	Efr(m)	Rad(p)	Poi(μ)	Rho(q)	Loo
$R_{2,W}(n,\widehat{p}_\lambda)$	$\frac{n}{m}\left(1-\frac{1}{\widehat{np}_\lambda}\right)$	$\frac{1}{p}-1$	$\frac{1}{\mu}\left(1-\frac{1}{\widehat{np}_\lambda}\right)$	$\frac{n}{q}-1$	$\frac{1}{n-1}$
C_W	m/n	$p/(1-p)$	μ	$q/(n-q)$	$n-1$

TABLE 2
C_W for several resampling schemes.

In the general heteroscedastic framework of (5), Prop. 1 shows that resampling penalties take into account the fact that σ_λ^2 actually depends on $\lambda \in \Lambda_m$. This is a major difference with the classical Mallows' C_p penalty

$$\mathrm{pen}_{\mathrm{Mallows}}(m) := \frac{2\mathbb{E}\left[\sigma(X)^2\right] D_m}{n}$$

which does not take into account the variability of the noise level over $\mathcal{X}$. A more detailed comparison with Mallows' C_p is made in Sect. 8.1.1.

Moreover, assuming that $R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda)$ does not depend too much on $\widehat{p}_\lambda$ (at least when $n\widehat{p}_\lambda$ is large), it is possible to choose

$$C = C(W) = C_W \approx \frac{2}{R_{1,W}(n,1) + R_{2,W}(n,1)} \ ,$$

which only depends on n and $\mathcal{L}(W)$. Then, comparing (12) and (16), we would have a penalty close to the ideal one in expectation. Together with concentration inequalities, this result is at the core of our proof of the oracle inequalities of Sect. 4.1. As shown in the next subsection, $R_{1,W} + R_{2,W}$ have this property several interesting exchangeable weights W.

We focus now on the example of resampling weights given in Sect. 2.2. For each of them, we can compute explicitly $R_{2,W}(n,\widehat{p}_\lambda)$ (see Tab. 2) and show that $R_{1,W}(n,\widehat{p}_\lambda) \approx R_{2,W}(n,\widehat{p}_\lambda)$ when $n\widehat{p}_\lambda$ is large. As a consequence, we can define $C_W \approx R_{2,W}^{-1}$ as in Tab. 2. Then, the following proposition, combined with (16) in Prop. 1, shows that resampling penalties with $C = C_W$ are approximately unbiased estimates for the ideal penalty, assuming that $\min_{\lambda\in\Lambda_m}\{np_\lambda\}$ is large enough.

Proposition 2. *Let W be an exchangeable resampling weight vector among Efr(m_n), Rad(p), Poi(μ), Rho($\lfloor n/2 \rfloor$) and Loo, and define C_W as in Tab. 2. Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ and* pen(m) *defined by* (7)*. Then, there exists real numbers $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ (depending only on n, $\widehat{p}_\lambda$ and the resampling scheme $\mathcal{L}(W)$) such that*

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{C_W n}\sum_{\lambda\in\Lambda_m}\left(2+\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}\right)\sigma_\lambda^2 \ . \tag{17}$$

If $m_n n^{-1} \geq B > 0$ (Efr), $p \in (0;1)$ (Rad) or $\mu > 0$ (Poi), then,

$$\forall n \in \mathbb{N}\backslash\{0\}, \forall \widehat{p}_\lambda \in (0,1], \quad \left|\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}\right| \leq L_2\left(n\widehat{p}_\lambda\right)^{-1/4} \ ,$$

where $L_2 > 0$ is a numerical constant (for Rho($\lfloor n/2 \rfloor$) and Loo) or depends respectively on B (Efr), p (Rad) or μ (Poi). More accurate bounds for each kind of weights are given by (62) *to* (66).

Remark 6. Prop. 2 can also be generalized to Rho(q_n) weights with $0 < B_- \leq q_n n^{-1} \leq B_+ < 1$, but the bound on $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ only holds for $n\widehat{p}_\lambda \geq L(B_-, B_+)$ (and L_2 depends on B_-, B_+). See the proof of Prop. 2 for details.

Remark 7. With the explicit computation of $R_{1,W}$ and $R_{2,W}$ for several resampling weights, we can enlighten several known results.

- In the maximum log-likelihood framework, Shibata (64) showed the asymptotical equivalence of two bootstrap penalization methods. The first penalty, denoted by B_1, is Efron's bootstrap penalty (30). It is defined by (4) with $C = 1$ and Efron (n) weights. The second penalty, denoted B_2, was proposed by Cavanaugh and Shumway (26). It is the equivalent of

$$2\widehat{p}_1(m) := 2\mathbb{E}_W \left[P_n \left(\gamma(\widehat{s}^W_m) - \gamma(\widehat{s}_m) \right) \right]$$

  in the log-likelihood contrast. In the least-square regression framework (with histogram models), the proofs of Prop. 1 and 2 show that

$$\mathbb{E}^{\Lambda_m} \left[2\widehat{p}_1(m) \right] = \frac{2}{n} \sum_{\lambda \in \Lambda_m} R_{1,W}(n, \widehat{p}_\lambda) \sigma^2_\lambda \approx \mathbb{E}^{\Lambda_m} \left[\mathrm{pen}(m) \right]$$

  for several resampling schemes, including Efron's bootstrap (for which $C_W = 1$). The concentration results of Sect. 9.9 show that this remains true without expectations. Our result is thus a non-asymptotic version of Shibata's (64), for general resampling weights, in the least-square regression framework.
- With Efron (m_n) weights (and a bootstrap selection procedure close to RP, but with $C = 1$), Shao (61) showed that $m_n = n$ leads to an inconsistent model selection procedure for identification. On the contrary, when $m_n \to \infty$ and $m_n \ll n$, the bootstrap selection procedure becomes consistent. Notice that the constant $C = 1$ is then much larger than $C_W = m_n/n$. Considering that identification needs overpenalization within a factor that goes to infinity with n, (12) gives a simple explanation to this phenomenon.

4.4.2. Concentration inequalities for resampling penalties

According to (46) and (47), the ideal penalty is a U-statistics of order 2, conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{(i,\lambda \in \Lambda_m)}$. From the asymptotic viewpoint, this is sufficient to show that resampling gives a consistent estimate of it (Arcones and Giné (5) considered the bootstrap case; Hušková and Janssen (42) extended it to the exchangeable weighted bootstrap). In our non-asymptotic framework, we need the following result.

Proposition 3. *Let W be an exchangeable weight vector and* $\mathrm{pen}(m)$ *the corresponding Resampling Penalty defined by* (7)*. Let $\gamma > 0$ and $A_n \geq 2$. Assume that for every $q \geq 2$,*

$$\frac{\sqrt{D_m \sum_{\lambda \in \Lambda_m} m_{q,\lambda}^4}}{\sum_{\lambda \in \Lambda_m} m_{2,\lambda}^2} \leq a_\ell q^{\xi_\ell} \quad \text{where} \quad m_{q,\lambda} := \left(\mathbb{E}\left[|Y - s_m(X)|^q \mid X \in I_\lambda \right] \right)^{1/q} .$$

Then, there are constants $K_4, K_5 > 0$ and an event of probability at least $1 - K_4 n^{-\gamma}$ on which

$$\begin{aligned} \left| \mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) \right] \right| \mathbb{1}_{\min_{\lambda \in \Lambda_m} \left\{ n\widehat{p}_\lambda \right\} \geq A_n} &\leq C K_5 \\ \times \sup_{np \geq A_n} \left\{ R_{1,W}(n,p) + R_{2,W}(n,p) \right\} \frac{\ln(n)^{\xi_\ell + 1}}{\sqrt{A_n D_m}} &\mathbb{E}\left[p_2(m) \right] \end{aligned} \tag{18}$$

where $R_{1,W}$ and $R_{2,W}$ are defined by (14) *and* (15)*. The constant K_4 is numerical, while K_5 only depends on a_ℓ, ξ_ℓ, γ.*

If moreover W satisfies the assumptions of the second part of Prop. 2 and C_W is defined as in Tab. 2, then there is a constant $K_W > 0$ such that

$$\left| \mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)] \right| \mathbb{1}_{\min_{\lambda \in \Lambda_m} \left\{ n\widehat{p}_\lambda \right\} \geq A_n} \leq \frac{C K_5 K_W \ln(n)^{\xi_\ell + 1}}{C_W \sqrt{A_n D_m}} \mathbb{E}\left[p_2(m) \right] . \tag{19}$$

For the Rad(p) weights, K_W is smaller than $(1-p)^{-1}$ times a numerical constant. For the other weights, K_W is numerical.

Notice that the moment condition holds under the assumptions of Thm. 1 as well as the alternative assumptions of Sect. 4.3. It is here stated in its most general form.

Remark 8. With the $A_n^{-1/2}$ factor, we obtain better bounds for resampling penalties than for ideal penalties (see Prop. 12 below).

Although we do not know how tight are our bounds, such a phenomenon is classical with bootstrap and may be understood in the asymptotic viewpoint through Edgeworth expansions (Hall (38)). In a non-asymptotic gaussian framework, (10) (see Sect. 2.3) show the same property for resampling estimators, which concentrates at the rate n^{-1} instead of $n^{-1/2}$ (n being the amount of data). Since A_n plays the role of n in our case, it is reasonable to believe that the gain $A_n^{-1/2}$ may not be improved without some more assumptions.

This stresses the fact that resampling penalties do not estimate the ideal penalties themselves but their expectations. Thus, our procedure (with $C = C_W$) cannot take into account the fact that $\mathrm{pen}_{\mathrm{id}}(m)$ may be far from its expectation.

4.4.3. Expectations of inverses

For any non-negative random variable Z, we define

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

This quantity appears in the explicit formulas for $R_{1,W}$ when W is among the examples of resampling weights of Sect. 2.2. We then need non-asymptotic bounds on this quantity when Z has a binomial, hypergeometric or Poisson distribution, in our proof of Prop. 2. Former results concerning e_Z^+ can be found in papers by Lew (48) (for general Z), Jones and Zhigljavsky (44) (for the Poisson case) or Žnidarič (73) (for the binomial and Poisson case), but they are either asymptotic or not accurate enough. The following results solve this issue.

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.

Binomial case

Lemma 4. *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), $\kappa_1 = 5.1$ and $\kappa_2 = 3.2$. Then, if $np \geq 1$,*

$$\kappa_2 \wedge \left(1 + \kappa_1 (np)^{-1/4}\right) \geq e^+_{\mathcal{B}(n,p)} \geq 1 - e^{-np} \tag{20}$$

$$\text{and} \qquad 2 + 3 \times 10^{-4} \geq e^+_{\mathcal{B}\left(n,\frac{1}{2}\right)} \geq \mathbb{1}_{n \geq 3} \ . \tag{21}$$

In particular, $e^+_{\mathcal{B}(n,p)} \to 1$ when $np \to \infty$, which can be derived from (73). Notice that (20) was stated and proven in (9), where it is called Lemma 3. The second inequality is useful for studying Rademacher weights.

Hypergeometric case Recall that an hypergeometric random variable $X \sim \mathcal{H}(n,r,q)$ is defined by

$$\forall k \in \{0, \ldots, q \wedge r\}, \quad \mathbb{P}(X = k) = \frac{\binom{r}{k}\binom{n-r}{q-k}}{\binom{n}{q}} \ .$$

Lemma 5. *Let $n, r, q \in \mathbb{N}$ such that $n \geq r \geq 1$ and $n \geq q \geq 1$.*

1. *General lower-bound:*

$$e^+_{\mathcal{H}(n,r,q)} \geq 1 - \mathbb{1}_{r \leq n-q} \exp\left(-\frac{qr}{n}\right) \ . \tag{22}$$

2. *General upper-bound: Let $\epsilon \in (0;1)$ and $\kappa_3(\epsilon) = 0.9 + 1.4 \times \epsilon^{-2}$.*

$$\text{If} \qquad r \geq 2 \quad \text{and} \quad \frac{n}{q} \leq (1-\epsilon)\frac{2r}{2 + \sqrt{3(r+1)\ln(r)}}$$

$$\text{Then,} \qquad 1 + \kappa_3(\epsilon)\frac{n}{q}\sqrt{\frac{\ln(r)}{r}} \geq e^+_{\mathcal{H}(n,r,q)} \ . \tag{23}$$

3. *"Rho" case: if $n \geq 2$,*

$$14.3 \geq \sup_{r \geq 1}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2} \rfloor)}\right\} \qquad \text{and} \qquad 3 \geq \sup_{r \geq 26}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2} \rfloor)}\right\} \ . \tag{24}$$

4. "Loo" case:

$$1+\frac{\mathbb{1}_{r\geq 2}}{n(r-1)} \geq e^+_{\mathcal{H}(n,r,n-1)} = 1+\frac{1}{n}\left(\frac{(n-1)r}{n(r-1)}\mathbb{1}_{r\geq 2}-1\right) \geq 1-\frac{\mathbb{1}_{r=1}}{n} \quad . \tag{25}$$

5. "Lpo" case: if $n \geq r \geq n-q+1 \geq 2$,

$$\frac{r}{r-n+q} \times \frac{n^{n-q}}{n(n-1)\cdots(q+1)} \geq e^+_{\mathcal{H}(n,r,q)} \geq 1 \quad . \tag{26}$$

In particular, if $\sup_k \left\{ n_k q_k^{-1} \wedge (n_k - q_k) \right\} < +\infty$ and $n_k \geq r_k \to +\infty$, then $e^+_{\mathcal{H}(n_k,r_k,q_k)} \to 1$ when $k \to \infty$.

Poisson case

Lemma 6. *For every* $\mu > 0$, $\mathcal{P}(\mu)$ *denotes the Poisson law with parameter* μ. *Then,*

$$\left(2-2e^{-2\mu}\right) \wedge \left(1+\frac{2(1+e^{-3})}{(\mu-2)_+}\right) \geq e^+_{\mathcal{P}(\mu)} \geq \mathbb{1}_{\mu\geq 1.61} \vee \left(1-e^{-\mu}\right) \quad . \tag{27}$$

In particular, $e^+_{\mathcal{P}(\mu)} \to 1$ when $\mu \to \infty$, which can be derived from (44; 73).

5. Comparison of the weights

According to Thm. 1, any resampling scheme among Efr, Rad, Poi, Rho and Loo leads to an asymptotically optimal procedure. Even from the non-asymptotic viewpoint, it is not quite clear to distinguish between these weights with our main result. This comes from the equality of these resampling penalties in expectation at first order (Prop. 2), while deviations are negligible in front of expectations (Prop. 3).

As a consequence, differences between these weights can only come from second-order terms, either in the expectations or in the sizes of the deviations of resampling penalties. As a first step, we compare in this section second-order terms in the expectations of the penalties (*i.e.* differences between second-order terms in (16) and (17)), for a fixed sample size. For asymptotic considerations, we refer to Barbe and Bertail (15), Chap. 2, where Edgeworth expansions are used to compare the accuracy of estimation with many exchangeable weights. The asymptotic results mentioned in Sect. 4.4.3 may also be useful.

In Sect. 4.4.1, we show that $\mathrm{pen}_{\mathrm{id}}(m)$ and $\mathrm{pen}(m)$ have the same expectation, up to small terms δ_{n,p_λ} and $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$. We deduce that

$$\mathbb{E}\left[\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\overline{\delta}^{(\mathrm{penW})}_{n,p_\lambda} - \delta_{n,p_\lambda}\right)(\sigma_\lambda)^2 \tag{28}$$

$$\text{with} \qquad \overline{\delta}^{(\mathrm{penW})}_{n,p_\lambda} := \mathbb{E}\left[\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} \,\middle|\, \widehat{p}_\lambda > 0\right] \quad .$$

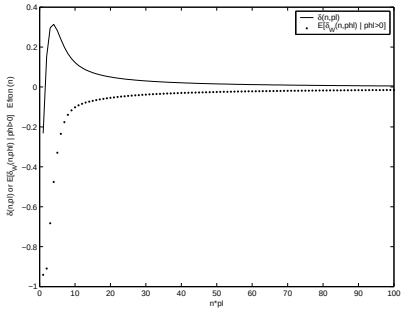

FIG 1. $\delta_{n,p} > 0 > \overline{\delta}_{n,p}^{(\mathrm{penEfr(n)})}$.

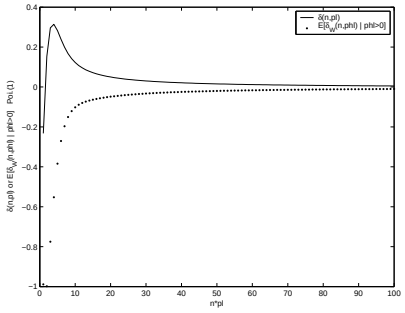

FIG 2. $\delta_{n,p} > 0 > \overline{\delta}_{n,p}^{(\mathrm{penPoi(1)})}$.

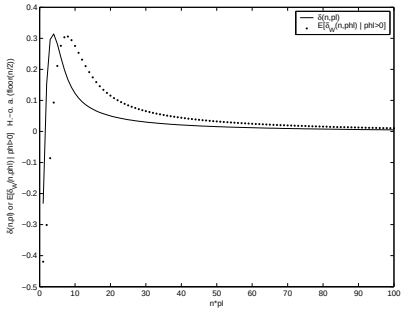

FIG 3. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRho}(n/2))}$ *for* $np \geq 6$.

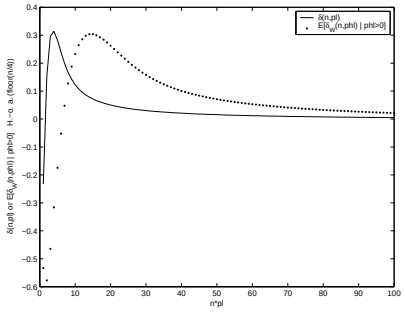

FIG 4. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRho}(n/4))}$ *for* $np \geq 9$.

Since we have explicit expressions for $\delta_{n,p}$ and $\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penW})}$, we have been able to compute numerically $\delta_{n,p}$ and $\overline{\delta}_{n,p}^{(\mathrm{penW})}$ as a function of np for several resampling schemes, with $n = 200$. The results are given on Fig. 1 to 6 (with straight lines for $\delta_{n,p}$ and dots for $\overline{\delta}_{n,p}^{(\mathrm{penW})}$).

It follows that Loo is the most accurate, even when np is small. On the contrary, Rho $(n/2)$ and Rad give overestimations of $\delta_{n,p}$ (except when np is small, where they are underpenalizing). It also seems that the bias of Rho (q) is a decreasing function of q, as illustrated by Fig. 4. Finally, Efr and Poi are strongly underestimating the ideal penalty, mostly because of the $1 - (n\widehat{p}_\lambda)^{-1}$ term in $R_{1,W}$ and $R_{2,W}$.

This can be summed up as follows:

$$\mathrm{penRad} \approx \mathrm{penRho} > \mathrm{penLoo} >> \mathrm{penEfr} \approx \mathrm{penPoi} \ , \tag{29}$$

where ">>" means a comparatively large gap, but still negligible at first order. As a consequence, we can expect that the leave-one-out penalty is the most efficient, closely followed by Rad and Rho. However, from the non-asymptotic viewpoint, it turns out that it is generally better to overpenalize slightly (and sometimes strongly, see the simulations of Sect. 6 and the discussion of Sect. 7.3.2).

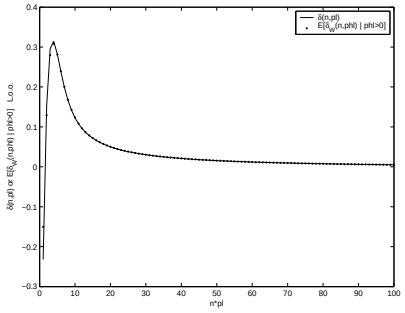

FIG 5. $\delta_{n,p} \approx \overline{\delta}_{n,p}^{(\mathrm{penLoo})}$.

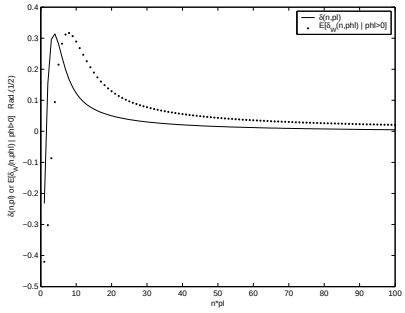

FIG 6. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRad}(1/2))}$ *for* $np \geq 6$.

Then, the ordering of (29) may also be the one of the prediction performance of RP according to the resampling scheme. This will be confirmed by the simulation study of Sect. 6.

Another interesting point is the following: we observe that $\overline{\delta}_{n,p}^{(\mathrm{penRho})} \propto \delta_{n,p}$ when np is large enough. Then, provided that we do not consider histograms with too small bins (w.r.t. $\mathcal{L}(X)$), penLoo and penRho are almost equivalent, up to the choice of the factor C. If a wise tuning of C is possible, we just have to choose between Loo and Rho according to computation issues (see the discussion of Sect. 7.2).

6. Simulation study

As an illustration of the results of Sect. 4, we compare the performances of Algorithm 2 (with several resampling schemes), Mallows' C_p and V-fold cross-validation on some simulated data.

6.1. 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. 7; smooth with jumps for HS, see Fig. 8), 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. 9 to 12. Their last difference lies in the families of models. Defining

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

$$(I_\lambda)_{\lambda\in\Lambda_{(k_1,k_2)}} = \left(\left[\frac{j}{2k_1}; \frac{j+1}{2k_1}\right)\right)_{0\leq j\leq k_1-1} \cup \left(\left[\frac{1}{2}+\frac{j}{2k_2}; \frac{1}{2}+\frac{j+1}{2k_2}\right)\right)_{0\leq j\leq k_2-1},$$

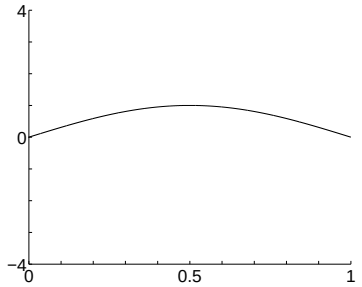

FIG 7. $s(x) = \sin(\pi x)$

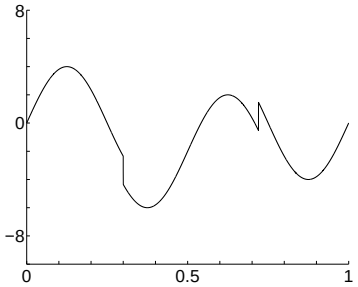

FIG 8. $s(x) = \text{HeaviSine}(x)$ *(see (27))*

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

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

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

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

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

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 quite efficient, hence the model selection problem more challenging.

We compare the following algorithms:

Mal Mallows' C_p penalty: $\text{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$ where $\widehat{\sigma}^2$ is the variance estimator (31). The non-asymptotic validity of this procedure for model selection in homoscedastic regression has been assessed by Baraud (13).

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

VFCV Classical V-fold cross-validation, with $V \in \{2, 5, 10, 20\}$ (defined as in (9)).

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

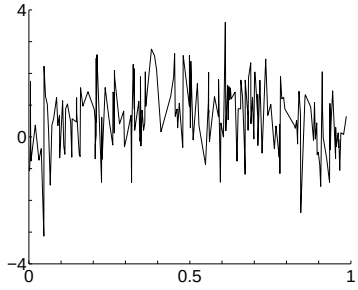

FIG 9. *S1:* $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$

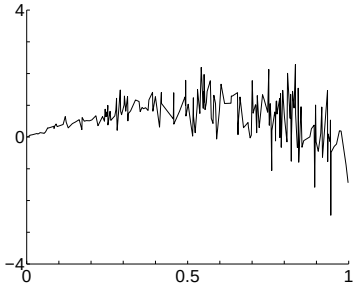

FIG 10. *S2:* $s(x) = \sin(\pi x)$, $\sigma(x) = x$, $n = 200$

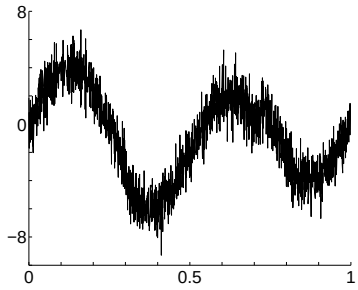

FIG 11. *HSd1: HeaviSine,* $\sigma \equiv 1$, $n = 2048$

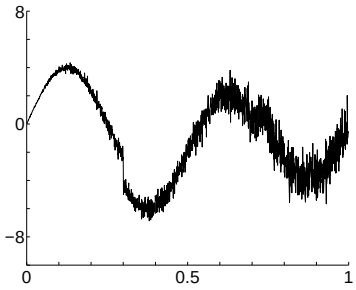

FIG 12. *HSd2: HeaviSine,* $\sigma(x) = x$, $n = 2048$

penEfr Efron (n) penalty, $C = C_W = 1$.
penRad Rademacher penalty, $C = C_W = 1$.
penRho Random hold-out ($n/2$) penalty, $C = C_W = 1$.
penLoo Leave-one-out penalty, $C = C_W = n - 1$.

For each of these, we also consider the same penalties multiplied by 5/4 (and we denote them by a + symbol added after the shortened names). This intends to test for overpenalization (the choice of the factor 5/4 being arbitrary and certainly not optimal, see Sect. 7.3.2).

In each experiment, for each simulated data set, we first remove the models with less than 2 data points in one piece of their associated partition. 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 $\ell\left(s, \widehat{s}_{\widehat{m}}\right)$ (and the excess loss $\ell\left(s, \widehat{s}_m\right)$ 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:

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

Basically, C_{or} is the constant that should appear in an oracle inequality like (9), and $C_{\mathrm{path-or}}$ corresponds to a pathwise oracle inequality like (8). As C_{or} and $C_{\mathrm{path-or}}$ approximatively give the same rankings between algorithms, we only report C_{or} in Tab. 3.

6.2. Results and comments

First of all, our experiments show the interest of both Resampling Penalization (RP) and VFCV in several difficult frameworks, with relatively small sample sizes. Although they can not compete with simple procedures such as Mallows' C_p from the computational viewpoint, they are much more efficient when the noise is heteroscedastic (S2 and HSd2). In these hard frameworks, the performances of RP and VFCV are comparable to those of the "ideal deterministic penalty" $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$. Notice that in the case of HSd2, penRad and penRho do better than any linear penalty (possibly with a slope that depends on both the data and the unknown distribution P; see Sect. 8.1.2). 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 longer the case when σ is smaller, and less obvious when n is larger (see respectively experiments S0.1 and S1000 in (7)). We would like to insist on the importance of this phenomenon, which is seldom mentioned in theoretical papers because it vanishes in the asymptotic framework, and it is quite hard to find from theoretical results.

TABLE 3

*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 $\sqrt{N}$; $N = 1000$). In each column, the more accurate algorithms (taking the uncertainty into account) 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
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$	1.919 ± 0.03	2.296 ± 0.05	1.028 ± 0.004	$\mathbf{1.102 \pm 0.004}$
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]+$	$\mathbf{1.792 \pm 0.03}$	$\mathbf{2.028 \pm 0.04}$	$\mathbf{1.003 \pm 0.003}$	$\mathbf{1.089 \pm 0.004}$
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 ± 0.04	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
penRad	1.973 ± 0.04	2.485 ± 0.06	1.018 ± 0.003	$\mathbf{1.102 \pm 0.004}$
penRho	1.982 ± 0.04	2.502 ± 0.06	1.018 ± 0.003	$\mathbf{1.103 \pm 0.004}$
penLoo	2.080 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penEfr	2.597 ± 0.07	3.152 ± 0.07	1.067 ± 0.005	1.114 ± 0.005
penRad+	$\mathbf{1.799 \pm 0.03}$	$\mathbf{2.137 \pm 0.05}$	$\mathbf{1.002 \pm 0.003}$	$\mathbf{1.095 \pm 0.004}$
penRho+	$\mathbf{1.798 \pm 0.03}$	$\mathbf{2.142 \pm 0.05}$	$\mathbf{1.002 \pm 0.003}$	$\mathbf{1.095 \pm 0.004}$
penLoo+	$\mathbf{1.844 \pm 0.03}$	$\mathbf{2.215 \pm 0.05}$	$\mathbf{1.004 \pm 0.003}$	$\mathbf{1.096 \pm 0.004}$
penEfr+	2.016 ± 0.05	2.605 ± 0.06	1.011 ± 0.003	$\mathbf{1.097 \pm 0.004}$

We now compare RP and VFCV. According to the four experiments of Tab. 3, it is quite clear that RP with Rad or Rho resampling schemes outperforms VFCV for any V, even without overpenalizing. The only exception to this is HSd1 where 2-fold cross-validation has particularly good performance. We refer to (9) for a non asymptotic study of the performance of V-fold cross-validation, both from the theoretical viewpoint and on these simulated data. In a nutshell, it appears that VFCV overpenalizes within a factor $1 + 1/(2(V-1))$, while the V-fold criterion has a variance which decreases with V. Then, when overpenalization is necessary (*e.g.* in S1, S2 or HSd1), small values of V can outperform the leave-one-out ($V = n$). But the increased variability of the criterion allow RP to do much better as soon as we overpenalize at the right level. The reason why penRad and penRho also perform slightly better without overpenalization is that they naturally overpenalize a little when $C = C_W = 1$ (*cf.* Sect. 5).

Let us now consider the performance of RP with several exchangeable resampling schemes. The two best ones are Rad and Rho, in the four experiments, with or without overpenalization. Then, Loo does slightly worse (but not always significantly), and Efr much worse. Looking carefully at the values of the penalties, it appears that Rad and Rho are slightly overpenalizing, Loo is exactly at the right level, and Efr underpenalizes (as well as Poi, which has performances quite similar to the ones of Efr, see (7)). Notice that this comparison can also be derived from theoretical computations (*cf.* Sect. 5). Since overpenalization is benefic in the four experiments of Tab. 3, this explains why penRad and penRho slightly outperform penLoo. In the case of Efron's boostrap penalty, underpenalizing implies overfitting, which can explain the comparatively bad performances we observe in Tab. 3.

We conclude this section with 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, resampling penalties remain significantly better than Mallows' C_p. Hence, the performances of Mallows' C_p in Tab. 3 are not only due to a bad estimation of the mean noise-level (see also Sect. 8.1).
- Eight additional experiments are reported in (7), showing similar results with various n, σ and s (although the assumptions of Thm. 1 are not always satisfied).
- Resampling penalties with a V-fold subsampling scheme have also been studied in Sect. 4 of (9), on the same simulated data. It is interesting to notice that exchangeable resampling schemes always give better performance that non-exchangeable ones (in a significant way when V is small),

except for the schemes which tend to underestimate the ideal penalty, like Efr and Poi.

7. Practical implementation

7.1. Computational cost

An exact computation of resampling penalties with exchangeable weights (without using our formula (48) for histograms) would be either impossible or very greedy. We suggest two possible ways to fix this problem.

First, one can use a classical Monte-Carlo approximation, *i.e.* draw a small number B of weight vectors instead of considering each element of the support of $\mathcal{L}(W)$. Practical methods for this are addressed *e.g.* by Hall (38), appendix II, in the bootstrap case. In addition, a non-asymptotic estimation of the accuracy of Monte-Carlo approximation can be obtained via McDiarmid's inequality (*cf.* Prop. 2.11 by Arlot, Blanchard and Roquain (10) in another framework). We would thus have a practical way of quantifying what we loose by making such an approximation, and choose B consequently (at least for Rad, Rho and Loo weights).

Second, it is possible to use non-exchangeable weight vectors W, such that the cardinality of the support of $\mathcal{L}(W)$ is much smaller than n. The case-example of such weights is V-fold subsampling ones: given a partition $(B_j)_{1\leq j\leq V}$ of $\{1,\dots,n\}$ and J a uniform random variable over $\{1,\dots,V\}$, independent from the data, we define

$$\forall i \in \{1,\dots,n\}\,, \quad W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}\ .$$

This leads to the so-called V-fold penalties, which have been introduced and studied in (9). They are quite similar to VFCV (*e.g.* from the computational viewpoint), while being more flexible (the overpenalization factor is decoupled from the choice of V). Similarly to resampling penalties, this results in an improvement over VFCV, with a reasonable computational cost.

Both approaches have been tested on the simulated data of Sect. 6. The detailed results are given in (7).

7.2. Choice of the weights

In both Algorithms 2 and 3, there are two tuning parameters: the distribution of the weight vector W, and the constant C. In this subsection, we address the question of choosing the first one.

We have already investigated the influence of the weights from the theoretical viewpoint in Sect. 5, focusing on second-order terms in expectation. However, it is likely that the deviations of $\mathrm{pen}(m)$ around its expectation also differ according to the weight vector W (the upper bound in (19) having no reason to

be tight). With the simulation study of Sect. 6, we can make comparisons that take into account both phenomenon.

In terms of model selection efficiency, Tab. 3 shows that the best weights (for accuracy of prediction and for the variability[4] of this accuracy) are Rho and Rad, while Loo perform slightly worse. On the other hand, from both accuracy and variability viewpoints, Efron's bootstrap weights appear to perform worse than Rho, Rad and Loo, mainly because they lead to underpenalization.

Notice however that this comparison is strongly dependent from our choices[5] for the constant C_W, which make all penalties unbiased at first order, but possibly slightly under or over-penalizing. Then, this could result in different performances on data which do not require overpenalization. The computations of Sect. 5 show that Efron's bootstrap weights have a real drawback, which can not be fixed only by changing C_W.

From the computational viewpoint (when computing the penalties exactly), Loo weights are the only reasonable ones, while being almost as accurate as Rho and Rad. This is why we suggest their use, enlarging the constant C when needed (see Sect. 7.3.2 on overpenalization).

However, computing n empirical risk minimizers (or more complicated algorithms) for each model may not always be possible. Then, one should avoid using the Leave-one-out with a Monte-Carlo approximation, because this would give a large importance to a very small number of data points. Rho or Rad weights seem much more safe in this situation. Alternatively, one may consider the use of V-fold penalties (defined in (9)) as a good alternative when the computational power is limited.

Let us emphasize that this analysis and the subsequent advices should be considered with caution. First, we believe that the deviations of the penalties around their expectations should be understood much better, because they can be comparable or even larger than the second-order terms in expectations. Second, in a more general framework, the comparison may be different. For instance, the variability of the leave-one-out procedure is known to be quite large in classification (Hastie, Tibshirani and Friedman (40)), while this phenomenon disappears when the empirical minimization algorithm $(X_i, Y_i)_{1 \leq i \leq n} \mapsto \widehat{s}_m$ is stable (Molinaro, Simon and Pfeiffer (56)). This results in a quite different picture for choosing V for V-fold cross-validation according to the framework we consider, as noticed in (9). We expect that such differences may arise for choosing between exchangeable resampling weights. As a consequence, this question deserves further investigations, either theoretical or empirical, in any framework

[4] which is more an indicator of the *stability* of the performance of RP than of the variance of the resampling penalty. However, it remains an interesting measure, since one may prefer an algorithm which performs always equally well compared to another one with better mean efficiency but very poor performances on a small probability event.

[5] However, it is quite unclear how to change C_W in order to optimize each penalty in the general case. This is why we have chosen in Tab. 2 to take C_W as "simple" as possible.

to which Resampling Penalization can be applied, such as binary classification, regression with a different contrast function, or density estimation.

Finally, we remark that the bias of the bootstrap penalty has already been noticed by Efron (30; 31), who proposed several ways to correct it, including a double bootstrap procedure and the .632 bootstrap. The novelty of our approach is to propose to use (for instance) a leave-one-out resampling scheme instead of the boostrap, so that it is much less necessary to correct for this bias. This shows the main interest of considering the general family of Resampling Penalties, instead of only one particular resampling scheme.

7.3. Choice of the constant C

7.3.1. Optimal constant for bias

From the asymptotical viewpoint, the optimal $C = C^\star$ for prediction is generally the one for which pen is an unbiased estimator of the ideal penalty $\mathrm{pen}_{\mathrm{id}}$ (at least for "reasonable" models). This is how we have defined C_W in the histogram case, and Prop. 1 and 2 provide a non-asymptotic control on the bias. This allowed us to derive non-asymptotic oracle inequalities, which imply the asymptotic optimality of RP. Hence[6], $C^\star \sim_{n\to\infty} C_W$.

However, even if the asymptotic theory tells us that resampling penalties of the general Algorithm 1 are asymptotically unbiased for every fixed model when $C = C_W$ (see Sect. 2.2), we have no guarantee that the same result holds when the dimension of m is allowed to grow with n. It is likely that the constant C which makes pen unbiased sometimes differ from C_W from the non-asymptotic viewpoint, out of the histogram framework. Then, choosing C would either require additional theoretical studies (which would be interesting, but may be quite hard in general), or a data-driven calibration method.

For the latter point, we suggest to use the so-called "slope heuristics" to choose C. This heuristics (and the subsequent algorithm) were first proposed by Birgé and Massart (21) in a Gaussian homoscedastic framework. Their claim is that the optimal penalty is twice the minimal penalty, *i.e.* the one under which the selected model is obviously too large. This result has been extended to some non-Gaussian heteroscedastic data (and so more general shapes for the ideal penalty) by Arlot and Massart (11). One can then conjecture that this algorithm works in a quite general framework. Then, it would be possible to estimate the shape of $\mathrm{pen}_{\mathrm{id}}$ by resampling, and the constant C_W with the slope heuristics.

7.3.2. Overpenalization

A careful look at the proof of Thm. 1 shows that a similar oracle inequality holds for any constant $C > 4C_W/5$, the leading constant staying close to one

[6] See the proof of Thm. 1 in (9) to prove that asymptotic optimality requires $C^\star \sim_{n\to\infty} C_W$ as soon as there are enough models close to the oracle.

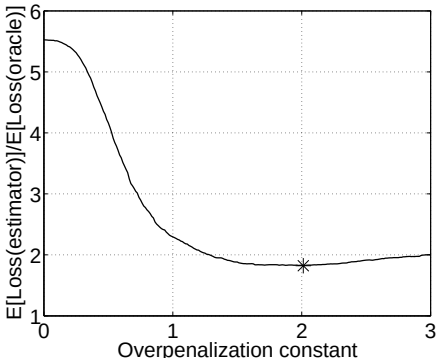

FIG 13. *The non-asymptotic need for overpenalization: the prediction performance* C_{or} *(defined in Sect. 6.1) of the model selection procedure* (1) *with* $\mathrm{pen}(m) = C_{\mathrm{ov}}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ *is represented as a function of* C_{ov}*. Data and models are the ones of experiment (S2):* $n = 200$*,* $\sigma(x) = x$*,* $s(x) = \sin(\pi x)$*. See Sect. 6 for more details.*

when $C \sim C_W$ *asymptotically*. In other words, when the sample size n is small, we have no guarantee that the optimal constant $C^\star$ is exactly equal to C_W. The simulations of Sect. 6 also supports this fact: *overpenalization*, *i.e.* taking $C = C_{\mathrm{ov}}C_W$ with $C_{\mathrm{ov}} > 1$, can improve the prediction performance of $\widehat{s}_{\widehat{m}}$ when n is small and σ large, or when s is non-smooth.

This problem would appear even if we knew the "optimal" constant $C^\star$ such that pen is non-asymptotically unbiased. On Fig. 13, we have estimated the model selection performance of the penalty $C_{\mathrm{ov}}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ as a function of C_{ov}, for experiment (S2) of Sect. 6. It appears that the optimal overpenalization constant $C_{\mathrm{ov}}^\star$ seems to be between 1.5 and 2.35 for this particular problem. More generally, the drawback of $C^\star$ is that it does not take into account the deviations of $\mathrm{pen}_{\mathrm{id}}(m)$ around its expectation. To avoid the possible overfit induced by these fluctuations, the constant C must be slightly enlarged. A major issue remains: how to estimate $C_{\mathrm{ov}}^\star$ from the data only, since it strongly depends on n, σ, the smoothness of s and the number of models in $\mathcal{M}_n$?

Several proposals can be made. If computational cost does not matter, one can think of choosing C_{ov} by V-fold cross-validation, but this would lead to a more than greedy algorithm. We here suggest a way of using the power of the resampling idea, for choosing the overpenalization factor wisely, with almost the same computational complexity as the initial algorithm.

Our idea relies on the fact that overpenalization is mainly due to the deviations of $\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$ around its expectation. If we had a simultaneous *confidence region* at level α for $\left(\mathrm{pen}_{\mathrm{id}}(m)\right)_{m\in\mathcal{M}_n}$ — instead of point estimates $\left(\mathrm{pen}(m)\right)_{m\in\mathcal{M}_n}$ — then we would be able to derive a confidence set for the oracle model $m^\star$ (instead of a single point estimate $\widehat{m}$). Overpenalizing could then be obtained by picking up the more parcimonious model in that confidence set (other ways of choosing a single model at final are possible). The point is that such a confidence region can be obtained by resampling, since Efron's heuristics

provides a way of estimating the whole distribution of any function of both P and P_n. In this particular case, this amounts to replace the expectation w.r.t. the resampling randomness in (4) by a quantile:

$$\mathrm{pen}_\alpha(m) := C_W \inf \left\{ t \in \mathbb{R} \text{ s.t. } P_n^W \left[P_n \gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W \gamma\left(\widehat{s}_m\left(P_n^W\right)\right) > t \right] \leq \alpha \right\} . \tag{30}$$

We have no theoretical guarantee that

$$\left(\left[\mathrm{pen}_{1-\alpha/2}(m), \mathrm{pen}_{1-\alpha/2}(m) \right] \right)_{m \in \mathcal{M}_n}$$

is a simultaneous confidence region for $\left(\mathrm{pen}_{\mathrm{id}}(m)\right)_{m\in\mathcal{M}_n}$ with level α (or even $\alpha\,\mathrm{Card}(\mathcal{M}_n)$, thanks to the union bound), but the resulting model selection procedure is overpenalizing, in the sense that it underfits more and more when α goes down to zero. Then, it is tempting to replace the choice of an overpenalization constant C_{ov} by the one of a "level" α, much easier to interpetate, and whose calibration may not depend too much on the distribution P of the data. It can thus be left to the final user, which has to decide up to which "level" he wants to be sure not to overfit.

A theoretical study of this procedure deserves further investigation and is beyond the scope of this paper. Let us just mention two references that may be helpful for defining and studying it properly. First, a non-asymptotic control of the level of some high-dimensional resampling-based confidence regions have been obtained by Arlot, Blanchard and Roquain (10), with general exchangeable resampling schemes. Second, the use of relative bounds (*i.e.* a simultaneous confidence region for $\left(\mathrm{pen}_{\mathrm{id}}(m) - \mathrm{pen}_{\mathrm{id}}(m')\right)_{m,m'\in\mathcal{M}_n}$) has been proposed by Audibert (12) to define model selection procedures in the statistical learning framework (see also Catoni (25)). Our proposal can be seen as mixing up these two ideas together.

Notice that such a modification can be made with the classical V-fold cross-validation (the level α being a multiple of V^{-1}), but not with Mallows' C_p penalty. This is not the only drawback of Mallows' C_p, as shown in Sect. 8.1.1. Further comments on overpenalization are given in (6; 9) for instance.

8. Discussion

8.1. Comparison with other procedures

In this article, we propose a family of model selection procedure, "Resampling Penalties" (RP), which can be used in almost any "reasonable" framework. Moreover, we provide both theoretical and empirical evidence that it indeed performs well for least-square histogram regression, with very few assumptions on the distribution of the data. In particular, this includes many kinds of heteroscedastic data, which are generally very hard to deal with. This robustness is a key property of RP, which may not fail dramatically in the most difficult

situations, while performing reasonably well for the simpler ones. However, computing the resampling penalties may be quite long, even with the suggestions made in Sect. 7.1, when minimizing $P_n^W \gamma(t,\cdot)$ over $t \in S_m$ is hard. In such cases, we need some clues for choosing between simple procedures (*e.g.* Mallows' C_p) and RP. In particular, for "easy" problems, RP can behave worse than Mallows', simply because it is more general. We would like to know what are those "easy" problems, for which we can avoid unnecessary long computations.

8.1.1. Mallows' Cp

Mallows' C_p penalty is equal to $2\sigma^2 D_m n^{-1}$ for a model m of dimension D_m, when the noise-level σ is constant. Non-asymptotic results about Mallows'-like penalties can be found in (16; 13; 14; 21). They imply that Mallows' C_p is asymptotically optimal in the homoscedastic framework, when $\mathcal{M}_n$ is not too large.

When the (constant) noise-level σ is unknown, one has to estimate it. Introducing artificially a model $S_{\lfloor n/2 \rfloor}$ of dimension $\lfloor n/2 \rfloor$, a classical estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{d^2\left(Y_{1\dots n}, S_{\lfloor n/2 \rfloor}\right)}{n - \lfloor n/2 \rfloor} \quad , \tag{31}$$

where $Y_{1\dots n} = (Y_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ and d the Euclidean distance on $\mathbb{R}^n$. Baraud (13; 14) then showed that the resulting data-driven model selection procedure satisfies some non-asymptotic oracle inequalities.

Assume for the sake of simplicity that n is even, and choose $S_{n/2}$ such that each piece of the associated partition contains exactly two data points. Reordering the (X_i, Y_i) according to X_i, we then have

$$\mathrm{pen}_{\mathrm{Mallows}}(m) = \frac{2D_m}{n^2} \sum_{i=1}^{n/2} \left(Y_{2i} - Y_{2i-1}\right)^2 \quad ,$$

so that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{Mallows}}(m)\right] = \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(D_m \widehat{p}_\lambda\right) \left(\sigma_\lambda^r\right)^2 + \frac{2D_m}{n^2} \sum_{i=1}^{n/2} \left(s(X_{2i}) - s(X_{2i-1})\right)^2 \tag{32}$$

$$\text{where} \qquad \left(\sigma_\lambda^r\right)^2 := \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X \in I_\lambda\right] \quad .$$

This should be compared with

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^d\right)^2\right) \tag{33}$$

$$\text{where} \qquad \left(\sigma_\lambda^d\right)^2 := \mathbb{E}\left[\left(s(X) - s_m(X)\right)^2 \,\middle|\, X \in I_\lambda\right] \quad .$$

Hence, both Mallows' C_p and the ideal penalty are in expectation the sum of a "variance" term (involving the $\left(\sigma_\lambda^r\right)^2$) and a "bias" term (involving the $\left(\sigma_\lambda^d\right)^2$). However, when s is smooth and $\min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\}$ is large, the "bias" term in (32) is negligible in front of the one of (33), which means that Mallows' C_p may be underpenalizing a little when the "bias" component of $\mathrm{pen}_{\mathrm{id}}$ is large.

On the other hand, the "variance" component of $\mathrm{pen}_{\mathrm{id}}$, which is the main one in general, is deformed in Mallows' C_p: the part of the penalty corresponding to I_λ is multiplied by $D_m\widehat{p}_\lambda$, which may not be close to 1 when the model m is not regular w.r.t. $\mathcal{L}(X)$. This happens for instance in the experiments S2 and HSd2 of Sect. 6.

These two main differences between Mallows' C_p and Resampling Penalization enlighten several possibly "hard" problems:

- heteroscedastic noise, with irregular histograms and X uniform (*e.g.* S2, HSd2 or Svar2 of Sect. 6),
- heteroscedastic noise, with regular histograms and X highly non-uniform on $\mathcal{X}$,
- regression function s with jumps (*e.g.* HeaviSine[7]) or large non-smooth areas (*e.g.* Doppler).

In either of those cases, one should avoid the use of Mallows'-like penalties, and we suggest RP as an efficient alternative. As explained below, the first class of problems can make any linear penalty suboptimal.

8.1.2. Linear penalties

The simplicity of Mallows' C_p comes from the fact that its shape is fixed *a priori* as linear in the dimension D_m of the models:

$$\mathrm{pen}(m) = \widehat{K}D_m \tag{34}$$

and there is only one constant $\widehat{K}$ to determine. In the case of Mallows' C_p, we take

$$\widehat{K}_{\mathrm{Mallows}} = 2\sigma^2 n^{-1} \quad \text{or} \quad 2\widehat{\sigma}^2 n^{-1}$$

if the mean variance level σ is unknown. Following the slope heuristics of Birgé and Massart (21), one can also define a data-dependent constant $\widehat{K}_{\mathrm{slope}}$ and define $\mathrm{pen}_{\mathrm{shape}}(m) = \widehat{K}_{\mathrm{slope}}D_m$, which is proved to be efficient in a homoscedastic framework (21; 11).

However, in view of (11), the ideal penalty is not linear in general, even in expectation. There are even frameworks in which any penalty of the form (34) is suboptimal, meaning that it can not satisfy any oracle inequality with leading constant smaller than some $\kappa > 1$ (see (6), Chap. 4). This is a quite strong result, since it even applies to procedures using the knowledge of s and σ. In particular,

[7]However, in experiment HSd1, Mallows' C_p still behaves quite well compared to RP. We do not know whether the non-smoothness of s can actually make Mallows' C_p fail.

even the following *optimal linear penalization algorithm* $\mathrm{pen}_{\mathrm{opt,lin}}(m) := K^\star D_m$ is suboptimal, where

$$K^\star \in \arg\min_{K>0} \left\{ P\gamma\left(\widehat{s}_{\widehat{m}(K)}\right)\right\} \quad \text{and} \quad \forall K>0,\, \widehat{m}(K) \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + KD_m \right\} \ .$$

Resampling penalties do not have the same drawback.

We compared resampling penalization with the ideal linear penalization in the four experiments of Sect. 6. The latter has a better performance for S1, S2 and HSd1. This is not surprising in "easy" situations, where Mallows' C_p is almost optimal, since $\mathrm{pen}_{\mathrm{opt,lin}}$ is always better than Mallows'. It is less intuitive for S2, which is more difficult, because of heteroscedasticity. Considering that $\mathrm{pen}_{\mathrm{opt,lin}}$ uses the knowledge of the true distribution P, one can understand that it is sufficient to keep a good performance for "intermediate" problems. However, in experiment HSd2, the ideal linear penalization has a constant $C_{\mathrm{or}} = 1.18 \pm 0.01$. This is worse than resampling penalization, for which $C_{\mathrm{or}} \leq 1.11$. Thus, the most difficult problem of Sect. 6 (with a complex family of models, heteroscedasticity and bias) gives another example where linear penalties are definitely not adapted.

8.1.3. Refined versions of Mallows'

In least-square regression and other frameworks, several penalties have been defined as refinements of Mallows' C_p, in Gaussian frameworks (Barron, Birgé and Massart (16), Birgé and Massart (21)) as well as in non-Gaussian ones (Baraud (14), Sauvé (59)). Basically, when $\mathrm{Card}(\mathcal{M}_n)$ is polynomial in n, these penalties are linear in D_m. So, they have at least the same drawbacks as the optimal linear penalty above.

When $\mathrm{Card}(\mathcal{M}_n)$ is larger (*e.g.* exponential in n), one has to take a larger penalty of the form

$$\mathrm{pen}(m) = KD_m\left(1 + c\ln\left(\frac{n}{D_m}\right)\right) \ .$$

With such a family of models, one can not use Resampling Penalization without modifications. Indeed, uniform deviations for $\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$ derived from the union bound may be too large, so that the model selection procedure can fail.

In order to solve this issue, we propose to apply Algorithm 1 to $(\widetilde{S}_C)_{1\leq C\leq n}$ instead of $(S_m)_{m\in\mathcal{M}_n}$, where

$$\widetilde{S}_C := \bigcup_{C_m = C} S_m$$

and $C_m \in \{1,\dots,n\}$ is any complexity measure (for instance the dimension of S_m as a vector space). This new model selection problem satisfies the polynomial complexity assumption. By grouping models according to C_m, we allow the

resampling procedure to detect the complexity of $\mathcal{M}_n$ through the complexity of each aggregated model $\widetilde{S}_C$. However, it is not straightforward to extend our results to this case when each S_m is an histogram model, because $\widetilde{S}_C$ is not. Results in this framework would be very interesting, since they could be related to the CART algorithm (defined by Breiman *et al.* (22); see (60) for an application of CART to variable selection).

8.1.4. Ad hoc *procedures*

One of the main points of Thm. 1 and 2 is that Resampling Penalization works in an heteroscedastic framework, contrary to Mallows' C_p. However, it is possible to adapt Mallows' penalty to heteroscedasticity, for instance by splitting $\mathcal{X}$ into disjoint subsets $(\mathcal{X}_k)_{1\leq k\leq K_n}$. Then, replacing $\sigma^2 D_m$ by $\sum_{k=1}^{K_n} \sigma_k^2 D_{m,k}$ (where σ_k and $D_{m,k}$ are local indexes for the noise and the complexity of S_m), and choose K_n such that both K_n and nK_n^{-1} go to infinity with n at an appropriate rate, we obtain a procedure that should be (asymptotically) optimal in the histogram case if σ is Lipschitz with a finite number of jumps. In the least-square regression framework, Efromovich and Pinsker (28) and Galtchouk and Pergamenschikov (35) defined other procedures that are more generally minimax adaptive in the heteroscedastic case.

These procedures may perform a little better than Resampling Penalization. We call them "*ad hoc*" because they are specially designed for the heteroscedastic case and a particular family of estimators. On the contrary, RP is a general-purpose device. It was neither built to be adaptive to heteroscedastic noises, nor to take advantage of a specific model (regression, histograms).

When no information is available on the data, or when no known algorithm can make use of such informations, we suggest the use of RP. Moreover, it may happen that available informations are partial or wrong. Then, using an *ad hoc* procedure may be catastrophic, whereas a general device like RP would still work. In a nutshell, choose RP if you have no useful information or if you do not trust them.

8.1.5. Other model selection procedures by resampling

The most well-known resampling-based model selection procedure is cross-validation. For practical reasons, it is often used in its V-fold version, which may have some tricky behaviors, in particular for choosing V (see Yang (71)). This can also be shown in our simulation experiments (Sect. 6, Tab. 3): in HSd1, $V = 2$ is better than $V \in \{5, 10, 20\}$. In (9), we explained this phenomenon by some bias of the V-fold criterion, that strongly depends on V. We also use Resampling Penalization for defining an alternative to V-Fold Cross-Validation, which does not have this drawback. The resulting "V-fold penalties" are thus an improvement on VFCV. In this paper, we have proven that RP with several exchangeable resampling schemes — generalizing the $V = n$ case — perform even better (or at least as well).

We have already mentioned the existence of several bootstrap model selection procedures (30; 61; 64). As noticed in Remark 7, the ones studied by Shibata (64) are quite close to RP, although stated in a less general form. In particular, they are restricted to Efron(n) weights, which are the worst ones in our framework, according to the simulation study of Sect. 6. Moreover, they do not consider useful to suggest that the penalty may be multiplied by a factor $C \neq 1$. With our formulation of RP, we have disconnected the question of choosing the weights from the overpenalization problem. This is crucial because it allows to obtain asymptotic optimality and overpenalization (and probably also consistency, see Sect. 8.2) with many resampling schemes. Providing such a unified approach for these resampling methods, our result enlighten for instance that the results of Shao (61) with Efron(m) weights are probably more related to his choice $C = 1$ than to some intrinsic property of the m out of n bootstrap (see Sect. 8.2).

8.2. Consistency

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

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

There is a huge amount of papers about model selection for identification; we refer to Shao (62) for general asymptotic results with linear models, and the introduction of papers by Yang (70; 71) 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 (1)). Thus, our explicit computations of the resampling penalties enlighten the results of Shao (61) on bootstrap model selection. Basically, Shao proved that resampling penalties with $C = 1$ and Efron (n) weights are inconsistent, while $C = 1$ and Efron(m) weights are consistent when $m \ll n$. In the histogram framework, we have proven that the ratio between Efron (m) resampling penalties and the ideal penalty is close to Cnm^{-1}. Then, Shao's results may be mostly due to the choice $C = 1$. Our conjecture is that choosing any $C \gg C_W$ would be sufficient to obtain a similar consistency result for resampling penalties with many exchangeable weights. This may be a crucial improvement because m out of n bootstrap weights are probably not the best ones in terms of variability (in particular when the computational cost matters, it may be hard to take into account all the data with these weights when n/m is large). In addition, as shown in Sect. 5, Loo, Rho and Rad weights may be more accurate than Efron(m) ones. These preliminary evidence show that the consistency properties of RP (with a suitable choice of the constant C) deserve further investigations.

8.3. Prediction in a general framework, including classification

Our results on Resampling Penalization are restricted to the histogram case so that we can wonder whether they stay valid in a general framework. We conjecture that they do, for the following reasons. First, Algorithm 1 can be applied (maybe up to some little modifications, as for histograms) to any model selection problem. It relies on the resampling idea, which is known to be quite robust in a wide variety of situations. With our Thm 1 and 2, we have confirmed this point for heteroscedastic regression, while RP has not been designed specifically for this problem.

Second, we mainly make the histogram assumption in order to control the expectations of resampling and ideal penalties from a non-asymptotic viewpoint, so that it is mainly technical. We have already mentioned that the same comparison is valid asymptotically in a much more general framework. In addition, several of the key concentration inequalities we use have been extended in Chap. 7 of (6) to a general framework, including bounded regression and binary classification.

In the classification setting, we even believe that RP can improve the performances of the classical resampling-based penalties: global and local Rademacher complexities. The former ones, introduced by Koltchinskii (45) and Bartlett, Boucheron and Lugosi (17), are using an i.i.d. Rademacher scheme to estimate

$$\mathrm{pen}_{\mathrm{id,g}}(m) := \sup_{t \in S_m} \{ (P - P_n)\gamma(t) \} \quad \text{instead of} \quad \mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) \ .$$

They have also been generalized by Fromont (34) to general exchangeable resampling schemes. Their main drawback is that they are much too large compared to the ideal penalty, so that they can not attain fast rates of estimation when the margin condition (introduced by Mammen and Tsybakov (53)) holds.

This is why localized penalties (*e.g.* local Rademacher complexities) have then been introduced to take into account the closeness of $\widehat{s}_m$ and s (50; 18; 19; 46). It is shown in these papers that localized penalties are close enough to $\mathrm{pen}_{\mathrm{id}}(m)$, so that the resulting algorithms can benefit of the margin condition. With RP, we precisely try to improve Fromont's bootstrap penalties (33) by estimating $\mathrm{pen}_{\mathrm{id}}(m)$ instead of $\mathrm{pen}_{\mathrm{id,g}}(m)$. Then, RP can be considered as local penalties.

But contrary to local Rademacher complexities, RP are estimating the ideal penalty itself, not a complicated upper bound defined as a fixed point of a local modulus of continuity. These are two crucial points in favour of RP, which may be closer to the ideal penalty and much easier to compute.

Moreover, local Rademacher complexities generally depend on several constants which are huge according to theory, and not only a multiplicative factor. They may thus be hard to calibrate from the data, in order to obtain at least an asymptotically optimal procedure. On the contrary, RP depends only on a multiplicative factor, which can for instance be chosen thanks to the slope heuristics (*cf.* Sect. 7). As a consequence, we can conjecture that RP (*e.g.* combined with the slope heuristics, as in (11)) is adaptive to the margin condition,

and simultaneously asymptotically optimal. A rigorous proof of this fact would of course be of much interest. We draw in Chap. 7 of (6) some possible paths towards such a proof.

8.4. Conclusion

This article intends to help the practical user to answer the following question: when shall Resampling Penalization be used? To sum up, we list below the advantages and drawbacks of RP *vs.* the classical methods.

Advantages of RP

- generality: well-defined in almost any framework.
- robustness and versatility: designed for the cautious user.
- adaptivity: to several properties, *e.g.* heteroscedasticity and smoothness of the target.
- flexibility: possibility of overpenalization, either for non-asymptotic prediction or for identification.

Drawbacks of RP

- computation time: one may prefer V-fold algorithms, such as V-fold cross-validation, or the V-fold subsampling version of resampling penalties (see (9)).
- possibly outperformed by Mallows' C_p (in easy cases) or *ad hoc* procedures (in some particular frameworks, when one has some information on the data).

9. Proofs

9.1. Notations

Before starting the proofs, we introduce some notations or conventions:

- The letter L designs "some positive numerical constant, possibly different from some place to another". In the same way, a positive constant which depends on $c_1, \ldots, c_k$ is denoted by $L_{c_1,\ldots,c_k}$, and if $(\mathbf{A})$ denotes a set of assumptions, $L_{(\mathbf{A})}$ is any positive constant that depends on the parameters appearing in $(\mathbf{A})$.
- By convention, $\infty \mathbb{1}_E$ and $\mathbb{1}_E/0$ are both equal to zero when E does not hold.
- For any $x \in \mathbb{R}$, $x_+ := x \vee 0$ and $x_- := (-x) \vee 0$.
- For any non-negative random variable Z, $e^0_{\mathcal{L}(Z)} := \mathbb{E}\left[Z\right] \mathbb{E}\left[Z^{-1} \mathbb{1}_{Z>0}\right]$.

- For every model $m \in \mathcal{M}_n$,

$$p_1(m) := P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) \qquad p_2(m) := P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$$
$$\overline{\delta}(m) := (P_n - P)\left(\gamma(s_m) - \gamma(s)\right) \quad .$$

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

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

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

$$\widetilde{p_1}(m) = \widetilde{p_1}^{(0)}(m) + \sum_{\lambda \in \Lambda_m} p_\lambda (\sigma_\lambda)^2 \mathbb{1}_{\widehat{p}_\lambda = 0} \qquad \text{with} \qquad \widetilde{p_1}^{(0)}(m) = \sum_{\lambda \in \Lambda_m} \frac{p_\lambda \mathbb{1}_{\widehat{p}_\lambda > 0}}{(n\widehat{p}_\lambda)^2} S_{\lambda,1}^2 \quad . \tag{35}$$

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

Notice that whatever the convention we choose (and even if we keep their original definition), $p_1(m)$ and $p_2(m)$ 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. $\widetilde{p}_2$).

9.2. General framework

9.2.1. Bounded assumption set ***(Bg)***

There is some noise: $\|\sigma(X)\|_2 > 0$.

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}; c_{\mathrm{rich}}\sqrt{n}]$.

(**P3**) The weights are exchangeable, among Efr, Rad, Poi, Rho and Loo.

(**P4**) The constant C is well chosen: $\eta C_W \geq C \geq C_W$.

(**Ab**) Bounded data: $\|Y_i\|_\infty \leq A < \infty$.

($\mathbf{A_{m,\ell}}$) Local moment assumption: there exists $a_\ell, \xi_\ell \geq 0$ such that for every $q \geq 2$, for every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$,

$$P_m^\ell(q) := \frac{\sqrt{D_m \sum_{\lambda \in \Lambda_m} m_{q,\lambda}^4}}{\sum_{\lambda \in \Lambda_m} m_{2,\lambda}^2} \leq a_\ell q^{\xi_\ell} \quad .$$

(**Ap**) Polynomial decreasing of the bias: there exists $\beta_1 \geq \beta_2 > 0$ and $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ such that, for every $m \in \mathcal{M}_n$,

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq \ell\left(s, s_m\right) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{A_Q}$) For every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$,

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[p_2(m)\right]}{D_m} = \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \sigma_\lambda^2 \geq c_Q^- > 0$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) Lower regularity of the partitions for $\mathcal{L}(X)$: there exists $c_{\mathrm{r},\ell}^X > 0$ such that for every $m \in \mathcal{M}_n$, $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\mathrm{r},\ell}^X$.

9.2.2. Unbounded assumption set ***(Ug)***

We remove (**Ab**) from (**Bg**), and add

($\mathbf{A}\sigma_{\max}$) Noise-level bounded from above: $\sigma^2(X) \leq \sigma_{\max}^2 < \infty$ a.s.

($\mathbf{As}_{\max}$) Bound on the target function: $\|s\|_\infty \leq A < \infty$.

($\mathbf{A}_{\mathbf{g},\epsilon}$) Global moment assumption for the noise: there exists $a_{g\epsilon}, \xi_{g\epsilon} \geq 0$ such that for every $q \geq 2$,

$$P^{g\epsilon}(q) := \|\epsilon\|_q \leq a_{g\epsilon} q^{\xi_{g\epsilon}}$$

($\mathbf{A}\delta$) Global moment assumption for the bias: there is a constant $c_{\Delta,m}^g > 0$ such that, for every $m \in \mathcal{M}_n$ of dimension $D_m \geq D_0$,

$$\|s - s_m\|_\infty \leq c_{\Delta,m}^g \|s(X) - s_m(X)\|_2$$

9.2.3. General result

Lemma 7. *Let $n \in \mathbb{N} \setminus \{0\}$, $\gamma_0 > 0$ and $\widehat{m}$ given by Algorithm 2. Assume that either* (**Bg**) *or* (**Ug**) *holds with constants independent from n.*

Then, there exists a constant K_1 (that depends on γ_0 and all the constants in (**Bg**) *(resp.* (**Ug**)*), but not on n) such that*

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left[2\eta - 1 + \ln(n)^{-1/5}\right] \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s, \widehat{s}_m\right)\right\} \tag{36}$$

with probability at least $1 - K_1 n^{-\gamma_0}$.

The proof of this lemma is made in Sect. 9.6.

Remark 9. If we remove the lower bound in (**Ap**) from the assumption set, then, the proof of Lemma 7 shows that there are constants $\gamma_1, \gamma_2 > 0$ (depending only on ξ_ℓ, resp. ξ_ℓ and $\xi_{g\epsilon}$) and an event of probability at least $1 - K_1 n^{-\gamma_0}$ on which

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left[2\eta - 1 + \ln(n)^{-1/5}\right] \inf_{\substack{m \in \mathcal{M}_n \\ D_m \geq \ln(n)^{\gamma_1}}} \left\{\ell\left(s, \widehat{s}_m\right)\right\} + \frac{\ln(n)^{\gamma_2}}{n} \quad . \tag{37}$$

Remark 10. In the infimum in (36), there may be some $m \in \mathcal{M}_n$ such that $\widehat{s}_m$ is not well defined. We take by convention $\ell\left(s, \widehat{s}_m\right) = \infty$ in those cases.

From the proof, there is a constant $c > 0$ (that depends on $\alpha_{\mathcal{M}}$, γ_0 and $c_{\mathrm{r},\ell}^X$) such that every model of dimension smaller than $cn\left(\ln(n)\right)^{-1}$ belongs to $\widehat{\mathcal{M}}_n$ on the event where (36) holds. For each of these models,

$$\ell\left(s, \widehat{s}_m\right) = \ell\left(s, s_m\right) + \widetilde{p_1}^{(0)}(m) = \ell\left(s, s_m\right) + \widetilde{p_1}(m)$$

so that we can restrict the infimum to models of dimension lower than $cn\left(\ln(n)\right)^{-1}$ with any of these conventions for $\ell\left(s, \widehat{s}_m\right)$.

9.3. Proof of Thm. 1

We apply Lemma 7 with $\gamma_0 = 2$. In order to deduce (8), it remains to show that $(\mathbf{A_{m},}_{\ell})$ and $(\mathbf{A_Q})$ are satisfied. This is true with $D_0 = 1$ since for every $m \in \mathcal{M}_n$,

$$P_m^{\ell}(q) = \frac{\sqrt{\sum_{\lambda \in \Lambda_m} m_{q,\lambda}^4}}{\sqrt{D_m} Q_m^{(p)}} \leq \frac{\left\| Y - s_m(X) \right\|_{\infty}^2}{Q_m^{(p)}} \leq \frac{4A^2}{Q_m^{(p)}}$$

$$Q_m^{(p)} := \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \left[\left(\sigma_{\lambda}^r\right)^2 + \left(\sigma_{\lambda}^d\right)^2 \right] \geq \sigma_{\min}^2 \ .$$

Let Ω_n be the event on which (8) holds true. Then,

$$\begin{aligned}
\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] &= \mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right) \mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right) \mathbb{1}_{\Omega_n^c}\right] \\
&\leq \left[2\eta - 1 + \epsilon_n\right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{\ell\left(s, \widehat{s}_m\right)\right\}\right] + A^2 K_1 \mathbb{P}\left(\Omega_n^c\right)
\end{aligned}$$

which proves (9). Following Remark 10, (9) also holds with $\mathcal{M}_n$ replaced by

$$\left\{ m \in \mathcal{M}_n \ \text{ s.t. } \ D_m \leq c(\alpha_{\mathcal{M}}, c_{\mathrm{r},\ell}^X) n \ln(n)^{-1} \right\}$$

and the convention $p_1(m) = \widetilde{p_1}^{(0)}(m)$. □

9.4. Proof of Thm. 1: alternative assumptions

We prove in this section the statements of Sect. 4.3.

9.4.1. No uniform lower bound on the noise-level

When $\sigma(X)$ is allowed to go to zero, we only need another proof for $(\mathbf{A_Q})$:

$$Q_m^{(p)} \geq \frac{\left\|\sigma\right\|_{L^2(\mathrm{Leb})}^2}{2c_{\mathrm{r,u}}} - \frac{K_{\sigma}^2 \left(c_{\mathrm{r,u}}^d\right)^2 \mathrm{diam}(\mathcal{X})^2}{D_m^{2\alpha_d}} - \frac{J_{\sigma} \left\|\sigma(X)\right\|_{\infty}^2}{2D_m} \quad \text{(by Lemma 16).}$$

Hence, $(\mathbf{A_{m},}_{\ell})$ and $(\mathbf{A_Q})$ hold true uniformly on models $m \in \mathcal{M}_n$ such that $D_m \geq D_0 = L_{(\mathbf{Bg})}$. □

9.4.2. Unbounded data

We still use Lemma 7, but the proof is a little longer.

Pathwise oracle inequality We prove it for a general γ_0 (since we need it for the classical oracle below). We have to prove $(\mathbf{A_{m},}_{\ell})$, $(\mathbf{A_Q})$, $(\mathbf{A_{g,}}_{\epsilon})$ and $(\mathbf{A}\delta)$. The first three ones are almost straightforward: for every $m \in \mathcal{M}_n$,

$$\begin{aligned} P_m^{\ell}(q) &= \frac{\sqrt{\sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4}}{\sqrt{D_m} Q_m^{(p)}} \leq \frac{\left(2A + c_{\text{gauss}}\sqrt{q}\sigma_{\max}\right)^2}{Q_m^{(p)}} \leq \frac{L_{c_{\text{gauss}},\sigma_{\max},A} q}{Q_m^{(p)}} \\ Q_m^{(p)} &\geq \sigma_{\min}^2 \\ P^{g\epsilon}(q) &\leq \sigma_{\max} c_{\text{gauss}} \sqrt{q} \ . \end{aligned}$$

For the last one, we use Lemma 17 (with $(\mathbf{Al})$, $(\mathbf{Ar}_{\ell,\mathbf{u}})$ and $(\mathbf{Ad}_{\ell})$) which shows that

$$c_{\Delta,m}^g \leq L_{(\mathbf{Ug})} \qquad \text{if} \quad D_m \geq D_0 = L_{(\mathbf{Ug})} \ .$$

Classical oracle inequality Let Ω_n be the event on which (8) holds true with $\gamma_0 = 6 + \alpha_{\mathcal{M}}$. As in the bounded case, we only have to upper bound

$$\begin{aligned} \mathbb{E}^{\Lambda_m}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\Omega_n^c}\right] &\leq \sqrt{\mathbb{P}(\Omega^c)}\sqrt{\mathbb{E}^{\Lambda_m}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)^2\right]} \qquad \text{by Cauchy-Schwartz} \\ &\leq \sqrt{K_1} n^{-\gamma_0/2} \sqrt{\mathbb{E}^{\Lambda_m}\left[2\left\|s\right\|_{\infty}^2 + 2p_1(\widehat{m})^2\right]} \\ &\leq L_{(\mathbf{Ug})} n^{-\gamma_0/2}\left[1 + \sqrt{\mathbb{E}^{\Lambda_m}\left[\sum_{m\in\mathcal{M}_n} p_1(m)^2 \mathbb{1}_{m\in\widehat{\mathcal{M}}_n}\right]}\right] \ . \end{aligned}$$

For every $m \in \widehat{\mathcal{M}}_n$, we have to compute $\mathbb{E}^{\Lambda_m}[(p_1(m))^2]$ (and derive a bound on it, even very poor). Starting from (46), we have

$$\begin{aligned} \mathbb{E}^{\Lambda_m}\left[p_1(m)^2\right] &= \frac{1}{n^2}\sum_{\lambda\in\Lambda_m}\left(\frac{p_\lambda}{\widehat{p}_\lambda}\right)^2 \mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] + \frac{1}{n^2}\sum_{\lambda\neq\lambda'}\left[\frac{p_\lambda}{\widehat{p}_\lambda}\frac{p_{\lambda'}}{\widehat{p}_{\lambda'}} m_{2,\lambda}^2 m_{2,\lambda'}^2\right] \\ &\leq \sum_{\lambda\in\Lambda_m}\mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] + \sum_{\lambda\neq\lambda'}\left(\sigma_{\max}^2 + (2A)^2\right)^2 \leq D_m^2 L_{(\mathbf{Ug})} \leq n^2 L_{(\mathbf{Ug})} \end{aligned}$$

since

$$\begin{aligned} \mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] &= \mathbb{E}^{\Lambda_m}\left[\frac{\left(\sum_{X_i\in I_\lambda}(Y_i - \beta_\lambda)\right)^4}{(n\widehat{p}_\lambda)^2}\right] = \frac{m_{4,\lambda}^4}{n\widehat{p}_\lambda} + \frac{6(n\widehat{p}_\lambda - 1)m_{2,\lambda}^4}{n\widehat{p}_\lambda} \\ \text{and} \quad D_m \sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4 &\leq \left(a_\ell q^{\xi_\ell}\right)^2\left(\sigma_{\max}^2 + (2A)^2\right)^2 \ . \end{aligned}$$

Using that $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$, we obtain

$$\mathbb{E}^{\Lambda_m}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right) \mathbb{1}_{\Omega_n^c}\right] \leq L_{(\mathbf{Ug})} n^{1+(\alpha_{\mathcal{M}}-\gamma_0)/2}$$

which proves (9). □

9.5. Proof of Thm. 2

In this proof, we denote by (**H**) the set of assumptions made in Thm. 2. All the assumptions of Thm. 1 are satisfied, except maybe the lower bound in (**Ap**) (for $(\mathbf{Ar}_\ell^{\mathbf{X}})$, we use $(\mathbf{Ad}_\ell)$ and the fact that all the models are "regular"). We start from (37) in Remark 9 below Lemma 7. The constants γ_i are then numerical, because the data is bounded.

Let $m(T_0) \in \mathcal{M}_n$ be the model of dimension T_0^k closest to $R^{\frac{2k}{2\alpha+k}} n^{\frac{k}{2\alpha+k}} \sigma_{\max}^{\frac{-2k}{2\alpha+k}}$. We must have

$$2^{-1} R^{\frac{2}{2\alpha+k}} n^{\frac{1}{2\alpha+k}} \sigma_{\max}^{\frac{-2}{2\alpha+k}} \leq T_0 \leq 2R^{\frac{2}{2\alpha+k}} n^{\frac{1}{2\alpha+k}} \sigma_{\max}^{\frac{-2}{2\alpha+k}} \quad .$$

This dimension is larger than $\ln(n)^{\gamma_1}$ and smaller than $cn(\ln(n))^{-1}$ if $n \geq L_{(\mathbf{H}),c}$. Hence, from the proof of Lemma 7, this model belongs to $\widehat{\mathcal{M}}_n$ and has a finite excess loss on the large probability event of Lemma 7. Moreover, this excess loss is smaller than

$$L\left(\ell\left(s, s_{m(T_0)}\right) + \mathbb{E}\left[\widetilde{p_1}^{(0)}(m(T_0))\right]\right)$$

when $n \geq L_{(\mathbf{H})}$. Since

$$\begin{aligned}
\ell\left(s, s_{m(T_0)}\right) &\leq R^2 T_0^{-2\alpha} \\
\mathbb{E}\left[\widetilde{p_1}^{(0)}(m(T_0))\right] &\leq \sup_{np \geq 0} e^0_{\mathcal{B}(n,p)} \times \frac{1}{n} \sum_{\lambda \in \Lambda_{m(T_0)}} \left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right) \\
&\leq \frac{2}{n}\left(R^2 T_0^{1-2\alpha} + \sum_{\lambda \in \Lambda_{m(T_0)}} (\sigma_\lambda^r)^2\right) \leq \frac{2R^2 T_0^{1-2\alpha}}{n} + \frac{2\sigma_{\max}^2 D_{m(T_0)}}{n} \quad ,
\end{aligned}$$

(the bound on $e^0_{\mathcal{B}(n,p)}$ coming from (37), Lemma 4.1), there is an event of probability at least $1 - K_1' n^{-2}$ on which

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq K_2 R^{\frac{2k}{2\alpha+k}} n^{\frac{-2\alpha}{2\alpha+k}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+k}} + \frac{\ln(n)^{\gamma_2}}{n} \quad ,$$

where K_2 only depends on k and α. Notice that the constant K_1 has been replaced by a larger one, K_1', so that the probability bound is nonpositive when n is too small. Enlarging K_1' once more, one can also drop off the $\ln(n)^{\gamma_2} n^{-1}$ term by adding 1 to the constant K_2. We then take expectations as in the proof of Thm. 1, and deduce (10).

When $(\mathbf{A}\sigma)$ holds, we replace $\sigma_{\max}$ by $\|\sigma\|_{L^2(\mathrm{Leb})}$ in the definition of $m(T_0)$. Then, for every $\lambda \in \Lambda_{m(T_0)}$ such that there is no jump of σ on I_λ,

$$\begin{aligned}(\sigma_\lambda^r)^2 \leq \max_{I_\lambda} \sigma^2 &\leq \left(\frac{K_\sigma}{T_0} + \sqrt{\int_{\mathcal{X}} \sigma^2(t)\,\mathrm{Leb}(dt)} \right)^2 \\ &\leq \left(1+\theta^{-1}\right) \frac{K_\sigma^2}{T_0^2} + (1+\theta) \int_{\mathcal{X}} \sigma^2(t)\,\mathrm{Leb}(dt)\end{aligned}$$

for every $\theta > 0$ (since $\mathrm{Leb}(\mathcal{X}) = 1$). If σ jumps on I_λ (and there are at most J_σ such λ), we simply bound $\max_{I_\lambda} \sigma^2$ by $\sigma_{\max}^2$. As a consequence, taking $\theta = T_0^{-1}$, we get

$$\begin{aligned}\mathbb{E}\left[\widetilde{p_1}^{(0)}(m(T_0)) \right] &\leq \frac{2}{n} \left(R^2 T_0^{1-2\alpha} + \sum_{\lambda \in \Lambda_{m(T_0)}} (\sigma_\lambda^r)^2 \right) \\ &\leq \frac{2R^2 T_0^{1-2\alpha}}{n} + \frac{2D_{m(T_0)} \|\sigma\|_{L^2(\mathrm{Leb})}^2}{n} + \frac{L_{(\mathbf{H})}}{n}\end{aligned}$$

and the end of the proof does not change. In this second case, we can also remove $(\mathbf{An})$ because all the assumptions stated in the first part of Sect. 4.3 are satisfied. □

9.6. Proof of Lemma 7

We first give the complete proof in the bounded case. Then, we will explain how it can be extended to the unbounded case.

9.6.1. Bounded case

For every $m \in \mathcal{M}_n$, define $\mathrm{pen}'_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}_{\mathrm{id}}(m) + (P - P_n)\gamma(s)$. By definition of $\mathrm{pen}_{\mathrm{id}}$ and $\widehat{m}$, we have for every $m \in \widehat{\mathcal{M}}_n$,

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m})\right) \leq \ell\left(s, \widehat{s}_m\right) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m)\right) \quad . \quad (38)$$

The idea of the proof is to show that $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ is negligible in front of $\ell\left(s, \widehat{s}_m\right)$ for "reasonable" models (*i.e.*, those which are likely to be either selected by Algorithm 2, 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. 9.9.

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,γ_0} on which the concentration inequalities of Prop. 3 and 12 and Lemma 13 and 15, hold with $\gamma = \alpha_{\mathcal{M}} + \gamma_0$ (or similarly $x = (\alpha_{\mathcal{M}} + \gamma_0)\ln(n)$), for every $m \in \mathcal{M}_n$. Using assumption $(\mathbf{P1})$, the union bound gives $\mathbb{P}\left(\Omega_{n,\gamma_0}\right) \geq 1 - L_{c_{\mathcal{M}}} n^{-\gamma_0}$.

First, let $c, \gamma_1 > 0$ be two constants to be chosen later, and consider $\widetilde{\mathcal{M}}_n$, the set of models $m \in \mathcal{M}_n$ such that $\ln(n)^{\gamma_1} \leq D_m \leq cn(\ln(n))^{-1}$. According

to $(\mathbf{Ar}^{\mathbf{X}}_{\ell})$, for every such m, $B_n(m) \geq c^X_{\mathrm{r},\ell} c^{-1} \ln(n)$, so that (77) ensures that $A_n(m) \geq \ln(n)$ on Ω_{n,γ_0}, if $c \leq L_{c^X_{\mathrm{r},\ell},\alpha_{\mathcal{M}},\gamma_0}$. In particular, $m \in \widehat{\mathcal{M}}_n$ on Ω_{n,γ_0}. We also assume that $n \geq \exp(D_0)$, so that $D_m \geq D_0$ for every $m \in \widetilde{\mathcal{M}}_n$ if $\gamma_1 \geq 1$. Now, using both bounds on D_m, by construction of Ω_{n,γ_0},

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

is smaller than $L_{(\mathbf{Bg})} \ln(n)^{-1} \left(\ell\left(s, s_m\right) + \mathbb{E}\left[p_2(m)\right] \right)$ on this event, provided that $c \leq L_{c^X_{\mathrm{r},\ell},\gamma}$ (to ensure that $B_n(m)$ is large enough) and $\gamma_1 \geq 2\xi_\ell + 6$. We now fix $c = L_{c^X_{\mathrm{r},\ell},\gamma} > 0$ and $\gamma_1 = L_{\xi_\ell}$ that satisfy these conditions. Using Prop. 2, Lemma 9 and the lower bound on $B_n(m)$, we have for every $m \in \widetilde{\mathcal{M}}_n$

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

as soon as $n \geq L_{(\mathbf{Bg})}$ (this restriction is necessary because the bounds are in terms of excess loss of $\widehat{s}_m$ instead of $\ell\left(s, s_m\right) + \mathbb{E}\left[p_2\right]$). Combined with (38), this gives: if $n \geq L_{(\mathbf{Bg})}$,

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \mathbb{1}_{\widehat{m} \in \widetilde{\mathcal{M}}_n} \leq \left[2\eta - 1 + \frac{L_{(\mathbf{Bg})}}{\ln(n)^{1/4}} \right] \times \inf_{m \in \widetilde{\mathcal{M}}_n} \left\{ \ell\left(s, \widehat{s}_m\right) \right\} \ . \tag{39}$$

Second, we prove that any minimizer $\widehat{m}$ of the penalized empirical criterion $\mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$ belongs to $\widetilde{\mathcal{M}}_n$ on the event Ω_{n,γ_0}. Define, for every $m \in \mathcal{M}_n$, $\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma\left(s\right)$, which has the same minimizers over $\widehat{\mathcal{M}}_n$ as crit. According to $(\mathbf{P2})$, there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If $n \geq L_{(\mathbf{Bg})}$, $m_0 \in \widetilde{\mathcal{M}}_n$, from which we deduce (using $(\mathbf{Ap})$)

$$\mathrm{crit}'(m_0) \leq \ell\left(s, s_{m_0}\right) + \left|\overline{\delta}(m_0)\right| + \mathrm{pen}(m_0) \leq L_{(\mathbf{Bg})} \left(n^{-\beta_2/2} + n^{-1/2} \right) \ . \tag{40}$$

On the other hand, if $D_m < \ln(n)^{\gamma_1}$, we have

$$\begin{aligned} \mathrm{crit}'(m) &\geq \ell\left(s, s_m\right) - \left|\overline{\delta}(m)\right| - p_2(m) \\ &\geq C^-_{\mathrm{b}} \left(\ln(n)\right)^{-\gamma_1\beta_1} - L_{A,\gamma_0} \sqrt{\frac{\ln(n)}{n}} - L_{(\mathbf{Bg})} \frac{\ln(n)^{1+\xi_\ell+\gamma_1}}{n} \end{aligned} \tag{41}$$

on Ω_{n,γ_0}. In addition, if $D_m > cn(\ln(n))^{-1}$ and $m \in \widehat{\mathcal{M}}_n$, by Lemma 11, $\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) - p_2(m)\right] \geq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right]/4$. As a consequence, by construction of Ω_{n,γ_0}, we have $\mathrm{pen}(m) - p_2(m) \geq (1 - L_{(\mathbf{Bg})} n^{-1/4}) \mathbb{E}\left[p_2(m)\right]$ on it, so that

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) - p_2(m) - \left|\overline{\delta}(m)\right| \geq L_{(\mathbf{Bg})} \ln(n)^{-1} \tag{42}$$

when $n \geq L_{(\mathbf{Bg})}$. Comparing (40), (41) and (42), it follows that $\widehat{m} \in \widetilde{\mathcal{M}}_n$ on Ω_{n,γ_0}, provided that $n \geq L_{(\mathbf{Bg})}$.

Finally, we show that the infimum can be extended to $\mathcal{M}_n$ in the right-hand side of (39), with the convention $\ell(s, \widehat{s}_m) = +\infty$ if $A_n(m) = 0$. Using similar arguments as above (as well as the definition of Ω_{n,γ_0}, in particular (71) for large models), we have $\ell(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{Bg})} \left(n^{-\beta_2/2} + n^{-1/2} \right)$ on Ω_{n,γ_0}. On the other hand, for every $m \in \mathcal{M}_n$, if $D_m < \ln(n)^{\gamma_1}$, $\ell(s, \widehat{s}_m) \geq \ell(s, s_m) \geq L_{(\mathbf{Bg})} \ln(n)^{-\gamma_1 \beta_1}$ while if $D_m > cn(\ln(n))^{-1}$, $\ell(s, \widehat{s}_m) \geq L_{(\mathbf{Bg})} \ln(n)^{-2}$ on Ω_{n,γ_0} as soon as $n \geq L_{(\mathbf{Bg})}$. Hence, if $n \geq L_{(\mathbf{Bg})}$, no model $m \notin \widetilde{\mathcal{M}}_n$ can contribute to the infimum in the right-hand side of (39). This concludes the proof of (36) in the bounded case.

9.6.2. Unbounded case

The proof of the bounded case has to be slightly modified. In the definition of Ω_{n,γ_0}, we replace the concentration inequalities of Lemma 13 by the ones of Lemma 14. We then have another condition for chosing γ_1, which is $\gamma_1 \geq 2\xi_{g\epsilon}+3$. The rest of the proof of (39) is unchanged.

In order to prove that $\widehat{m} \in \widetilde{\mathcal{M}}_n$, (41) has to be slightly changed because of the use of (75) instead of (72) to bound $\overline{\delta}(m)$. The final part of the proof is then modified similarly. □

9.6.3. Proof of Remark 9

We now prove the assertion made in Remark 9 below Lemma 7. Starting from (39), we can prove in the same way that $D_{\widehat{m}} \leq cn \ln(n)^{-1}$, but not the corresponding lower bound. Let $m \in \widehat{\mathcal{M}}_n$ such that $D_m < \ln(n)^{\gamma_1}$. Assume first that

$$\ell(s, s_m) \geq \frac{1}{1 - \ln(n)^{-1}} \left[(2\eta - 1 + \epsilon_n) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{\ln(n)^{\xi_\ell + \gamma_1 + 2}}{n} \right] , \tag{43}$$

where $\epsilon_n \leq L_{(\mathbf{Bg})} \ln(n)^{-1/4}$ comes from (8). Then, on Ω_{n,γ_0}, using (72) (with $\eta = \ln(n)^{-1}$), and (43),

$$\begin{aligned}
\mathrm{crit}'(m) &\geq \ell(s, s_m) - \left| \overline{\delta}(m) \right| - p_2(m) \\
&\geq (2\eta - 1 + \epsilon_n) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{\ln(n)^{\xi_\ell + \gamma_1 + 2}}{n} - L_A \frac{\ln(n)}{n} - L_{(\mathbf{Bg})} \frac{\ln(n)^{\xi_\ell + \gamma_1 + 1}}{n} \\
&\geq (2\eta - 1 + \epsilon_n) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{\ln(n)^{\xi_\ell + \gamma_1 + 2}}{2n} , \qquad (44)
\end{aligned}$$

provided that $n \geq L_{(\mathbf{Bg})}$. On the other hand, let $m_0 \in \arg\min_{m' \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_{m'}) \}$. Since $m_0 \in \widetilde{\mathcal{M}}_n$, on Ω_{n,γ_0},

$$\mathrm{crit}'(m_0) = \ell(s, \widehat{s}_{m_0}) + \mathrm{pen}(m_0) - \mathrm{pen}_{\mathrm{id}}(m_0) \leq (2\eta - 1 + \epsilon_n) \ell(s, \widehat{s}_{m_0}) ,$$

and this upper bound is smaller than the lower bound in (44).

Hence, on Ω_{n,γ_0}, if $D_{\widehat{m}} < \ln(n)^{\gamma_1}$, it can not satisfy (43). Using that

$$\widetilde{p_1}(m) \leq L_{(\mathbf{Bg})} \ln(n)^{\xi_\ell+2} \frac{D_m}{n}$$

for every $m \in \mathcal{M}_n$ such that $D_m \leq cn\ln(n)^{-1}$ on Ω_{n,γ_0}, we then have

$$\begin{aligned}
\ell\left(s, \widehat{s}_{\widehat{m}}\right) &= \ell\left(s, s_{\widehat{m}}\right) + \widetilde{p_1}(\widehat{m}) \\
&\leq \frac{2\eta - 1 + L_{(\mathbf{Bg})}\ln(n)^{-1/4}}{1 - \ln(n)^{-1}} \inf_{m \in \widetilde{\mathcal{M}}_n} \left\{ \ell\left(s, \widehat{s}_m\right) \right\} + L_{(\mathbf{Bg})} \frac{\ln(n)^{\xi_\ell+\gamma_1+2}}{n} \\
&\leq \left(2\eta - 1 + \ln(n)^{-1/5}\right) \inf_{m \in \widetilde{\mathcal{M}}_n} \left\{ \ell\left(s, \widehat{s}_m\right) \right\} + \frac{\ln(n)^{\xi_\ell+\gamma_1+3}}{n} \ , \qquad (45)
\end{aligned}$$

assuming that $n \geq L_{(\mathbf{Bg})}$. As a consequence, $\widehat{m}$ satisfies either (36) or (45) on Ω_{n,γ_0}. Finally, with the same arguments as in the proof of Lemma 7, we can extend the infimum in the right-hand side of (36) and (45) to the set of $m \in \mathcal{M}_n$ such that $D_m \geq \ln(n)^{\gamma_1}$, with the convention $\ell\left(s, \widehat{s}_m\right) = +\infty$ if $A_n(m) = 0$. Enlarging the constant K_1 to remove the conditions $n \geq L_{(\mathbf{Bg})}$, we have proven (37), with $\gamma_2 = \gamma_1 + \xi_\ell + 3$. The proof is quite similar in the unbounded case. □

9.7. Expectations

proof of Prop. 1. First, (11) and (16) directly come from Prop. 1 in (9) and its proof. The former result hold whatever the convention we take for p_1 and p_2 in Sect. 9.1. Since we use them in Sect. 3.2 and for proving concentration results, we recall below explicit expressions for p_1 and p_2 in the histogram framework:

$$p_1(m) = \sum_{\lambda \in \Lambda_m} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2 = \sum_{\lambda \in \Lambda_m} \frac{1}{n^2 \widehat{p}_\lambda} \frac{p_\lambda}{\widehat{p}_\lambda} S_{\lambda,1}^2 \qquad (46)$$

$$p_2(m) = \sum_{\lambda \in \Lambda_m} \widehat{p}_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2 = \sum_{\lambda \in \Lambda_m} \frac{S_{\lambda,1}^2 \mathbb{1}_{n\widehat{p}_\lambda > 0}}{n^2 \widehat{p}_\lambda} \ . \qquad (47)$$

Finally, (12) directly follows from the slightly more general Lemma 9 below (where W is allowed to depend on $\left(\mathbb{1}_{X_i \in I_\lambda}\right)_{(i,\lambda)}$). □

Lemma 8. *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* (7)*, and assume $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} \geq 1$. Then,*

$$\mathrm{pen}(m) = \frac{C}{n} \sum_{\lambda \in \Lambda_m} \left(R_{1,W}(n, \widehat{p}_\lambda) + R_{2,W}(n, \widehat{p}_\lambda) \right) \frac{n\widehat{p}_\lambda S_{\lambda,2} - S_{\lambda,1}^2}{n\widehat{p}_\lambda - 1} \mathbb{1}_{n\widehat{p}_\lambda \geq 2} \ , \qquad (48)$$

where $R_{1,W}$ and $R_{2,W}$ are defined by (14) *and* (15)*.*

proof of Lemma 8. First, as we have split $\mathrm{pen}_{\mathrm{id}}$ into p_1 and p_2 (plus a centered term), we split the penalty (without the constant C) into these two terms:

$$\widehat{p}_1(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] \tag{49}$$

$$\widehat{p}_2(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2\right] \ . \tag{50}$$

A key quantity to compute is the following: for every $\lambda\in\Lambda_m$ and $W_\lambda > 0$,

$$\begin{aligned}
\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda\right] &= \mathbb{E}_W\left[\widehat{p}_\lambda\left(\frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\left(1-\frac{W_i}{W_\lambda}\right)\right)^2 \,\middle|\, W_\lambda\right] \\
&= \frac{1}{n^2\widehat{p}_\lambda}\left[\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2\,\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)^2 \,\middle|\, W_\lambda\right]\right. \qquad (51) \\
&\left. + \frac{1}{n^2\widehat{p}_\lambda}\sum_{i\neq j, X_i\in I_\lambda, X_j\in I_\lambda}(Y_i-\beta_\lambda)(Y_j-\beta_\lambda)\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)\left(1-\frac{W_j}{W_\lambda}\right) \,\middle|\, W_\lambda\right]\right] .
\end{aligned}$$

Since the weights are exchangeable, $(W_i)_{X_i\in I_\lambda}$ is also exchangeable conditionally to W_λ and $(X_i)_{1\leq i\leq n}$. Thus, the "variance" term

$$R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) := \mathbb{E}_W\left[\left(W_i - W_\lambda\right)^2 \,\middle|\, W_\lambda\right]$$

does not depend from i (provided that $X_i\in I_\lambda$), and the "covariance" term

$$R_C(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) := \mathbb{E}_W\left[\left(W_i - W_\lambda\right)\left(W_j - W_\lambda\right) \mid W_\lambda\right]$$

does not depend from (i,j) (provided that $i\neq j$ and $X_i, X_j\in I_\lambda$). Moreover,

$$\begin{aligned}
0 &= \mathbb{E}_W\left[\left(\sum_{X_i\in I_\lambda}\left(W_i - W_\lambda\right)\right)^2 \,\middle|\, W_\lambda\right] \\
&= n\widehat{p}_\lambda R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) + n\widehat{p}_\lambda\left(n\widehat{p}_\lambda - 1\right) R_C(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W))
\end{aligned}$$

so that, if $n\widehat{p}_\lambda \geq 2$,

$$R_C(n, n\widehat{p}_\lambda, W_\lambda, W) = \frac{-1}{n\widehat{p}_\lambda - 1} R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W)) \ , \tag{52}$$

and $R_V(n, 1, W_\lambda, \mathcal{L}(W)) = 0$.

Combining (51) and (52), we obtain

$$\begin{aligned}
\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda\right] &= \frac{R_V(n, n\widehat{p}_\lambda, W_\lambda, \mathcal{L}(W))}{W_\lambda n^2\widehat{p}_\lambda}\mathbb{1}_{n\widehat{p}_\lambda\geq 2} \qquad (53) \\
&\quad \times\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - 1}S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda - 1}S_{\lambda,1}^2\right]
\end{aligned}$$

Combining (53) and (49) (resp. (53) and (50)), we deduce (48). □

Finally, let us recall the following lemma which compares p_1 and p_2 in expectation.

Lemma 9 (Lemma 7 of (9)). *If* $\min_{\lambda\in\Lambda_m}\{np_\lambda\}\geq B\geq 1$,

$$\left(1-e^{-B}\right)\mathbb{E}\left[\widetilde{p}_2(m)\right]\leq\mathbb{E}\left[\widetilde{p}_1^{(0)}(m)\right]\leq\mathbb{E}\left[\widetilde{p}_1(m)\right]\leq\left(1+\sup_{np\geq B}\delta_{n,p}\right)\mathbb{E}\left[\widetilde{p}_2(m)\right] \tag{54}$$

where $\delta_{n,p}$ *is the same as in* (16)*. A similar result holds with* p_2 *instead of* $\widetilde{p}_2$ *inside the expectation.*

9.8. Resampling constants

In this section, we prove some results relative to the exchangeable weights introduced in Sect. 2.2, in particular Prop. 2. We start with a lemma proving the formulas given in Tab. 2 for $R_{2,W}$.

Lemma 10. *Let* $n\in\mathbb{N}$ *and* $\widehat{p}_\lambda\in(0,1]$ *such that* $n\widehat{p}_\lambda\in\{1,\dots,n\}$*. Then, for every* $m\in\mathbb{N}\backslash\{0\}$, $p\in(0;1]$, $\mu>0$ *and* $q\in\{1,\dots,n\}$,

$$R_{1,\mathrm{Efr}(m)}=\frac{n}{m}e^+_{\mathcal{B}(m,\widehat{p}_\lambda)}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)\qquad R_{2,\mathrm{Efr}(m)}=\frac{n}{m}\left(1-\frac{1}{n\widehat{p}_\lambda}\right) \tag{55}$$

$$R_{1,\mathrm{Rad}(p)}=\frac{1}{p}e^+_{\mathcal{B}(n\widehat{p}_\lambda,p)}-1\qquad R_{2,\mathrm{Rad}(p)}=\frac{1}{p}-1 \tag{56}$$

$$R_{1,\mathrm{Poi}(\mu)}=\frac{1}{\mu}e^+_{\mathcal{P}(n\widehat{p}_\lambda\mu)}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)\qquad R_{2,\mathrm{Poi}(\mu)}=\frac{1}{\mu}\left(1-\frac{1}{n\widehat{p}_\lambda}\right) \tag{57}$$

$$R_{1,\mathrm{Rho}(q)}=\frac{n}{q}e^+_{\mathcal{H}(n,n\widehat{p}_\lambda,q)}-1\qquad R_{2,\mathrm{Rho}(q)}=\frac{n}{q}-1 \tag{58}$$

$$R_{1,\mathrm{Loo}}=\frac{n\widehat{p}_\lambda}{n(n\widehat{p}_\lambda-1)}\mathbb{1}_{n\widehat{p}_\lambda\geq 2}\qquad R_{2,\mathrm{Loo}}=\frac{1}{n-1} \tag{59}$$

where $\mathcal{B}$, $\mathcal{P}$ *and* $\mathcal{H}$ *are respectively the Binomial, Poisson and Hypergeometric distributions, and* $e^+_\mu=\mathbb{E}[Z]\,\mathbb{E}\left[Z^{-1}\,\middle|\,Z>0\right]$ *with* $Z\sim\mu$.

proof of Lemma 10. Since W is independent from the data, we can assume that the observations with $X_i\in I_\lambda$ are the $n\widehat{p}_\lambda$ first ones: $(X_1,Y_1),\dots,(X_{n\widehat{p}_\lambda},Y_{n\widehat{p}_\lambda})$. The random vector $(W_i)_{1\leq i\leq n\widehat{p}_\lambda}$ is then exchangeable (since W is). By definition of $W_\lambda=(n\widehat{p}_\lambda)^{-1}\sum_{i=1}^{n\widehat{p}_\lambda}W_i$, we deduce

$$\forall i\in\{1,\dots,n\widehat{p}_\lambda\},\quad\mathbb{E}_W\left[W_i\,|\,W_\lambda\right]=W_\lambda\ . \tag{60}$$

Then, the quantity

$$R_V(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W))=R_V(W_\lambda)=\mathbb{E}\left[\left(W_i-W_\lambda\right)^2\,\middle|\,W_\lambda\right]$$

appearing both in $R_{1,W}$ and $R_{2,W}$ is the variance of the weight W_i conditionally to W_λ.

Exchangeable subsampling weights We call *subsampling weight* any resampling weight W such that $W_i \in \{0, \kappa\}$ a.s. for every i. Such weights can be written $W_i = \kappa \mathbb{1}_{i \in I}$ for some random $I \subset \{1, \dots, n\}$. Rad and Rho are the two main examples of such weights, and they are both exchangeable. In their example 3.6.14, van der Vaart and Wellner (68) call this kind of weights "bootstrap without replacement". Using (60), we derive that

$$W_\lambda = \mathbb{E}_W \left[W_i \,|\, W_\lambda \right] = \kappa \mathbb{P} \left(W_i = \kappa \,|\, W_\lambda \right)$$

and thus

$$\mathcal{L} \left(W_i \,|\, W_\lambda \right) = \kappa \mathcal{B}(\kappa^{-1} W_\lambda) \qquad \text{and} \qquad R_V(W_\lambda) = W_\lambda(\kappa - W_\lambda) \ .$$

We then apply this result to Rad, for which $\kappa = p^{-1}$ and $\mathcal{L}(W_\lambda) = (n\widehat{p}_\lambda p)^{-1} \times \mathcal{B}(n\widehat{p}_\lambda, p)$ and deduce (56). In the Rho case, we have $\kappa = (n/q)$ and $\mathcal{L}(W_\lambda) = (q\widehat{p}_\lambda)^{-1} \mathcal{H}\left(n, n\widehat{p}_\lambda, q\right)$, so that (58) follows. The Loo is a particular case of Rho (with $q = n-1$), so that we only have to compute $e^+_{\mathcal{H}(n, n\widehat{p}_\lambda, n-1)}$. This is done with (25) in Lemma 5.

Efron Efron weights can also be written

$$W_i = \frac{n}{m} \operatorname{Card} \left\{ 1 \leq j \leq m \ \text{ s.t. } \ U_j = i \right\} \tag{61}$$

with $(U_j)_{1 \leq j \leq m}$ a sequence of i.i.d. random variables, uniform over $\{1, \dots, n\}$. From this, we deduce

$$\mathcal{L}(W_\lambda) = (m\widehat{p}_\lambda)^{-1} \mathcal{B}(m, \widehat{p}_\lambda) \qquad \text{and} \qquad \mathcal{L} \left(W_i \,|\, W_\lambda \right) = \frac{n}{m} \mathcal{B} \left(m\widehat{p}_\lambda W_\lambda, \frac{1}{n\widehat{p}_\lambda} \right) \ .$$

Thus,

$$R_V(W_\lambda) = \frac{n}{m} W_\lambda \left(1 - \frac{1}{n\widehat{p}_\lambda} \right)$$

and (55) follows.

Poisson One can check that the weights defined by (61), with $m = N_n \sim \mathcal{P}(\mu n)$ independent from the $(U_j)_{j \geq 1}$, are actually Poisson (μ) weights. This is the classical poissonization trick. Moreover, conditionally to W_λ and $N_n = m$, the same reasoning as for Efron(m) (with a multiplicative constant μ^{-1} instead of n/m) leads to (57). □

proof of Prop. 2. From (48), we obtain (17) with

$$\delta^{(\mathrm{penW})}_{n, \widehat{p}_\lambda} = C_W \left(R_{1,W}(n, \widehat{p}_\lambda) + R_{2,W}(n, \widehat{p}_\lambda) \right) - 2 \ .$$

Combining Lemma 10 with Lemma 4 (for Efr and Rad), Lemma 5 (for Rho and Loo) and Lemma 6 (for Poi), we obtain the following non-asymptotic bounds:

1. Efron (m_n): let $\kappa_1 = 5.1$ and $\kappa_2 = 3.2$,

$$(\kappa_2 - 1) \wedge \left(\frac{\kappa_1}{\left(Bn\widehat{p}_\lambda \right)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penEfr}(m_n))} \geq \frac{-2}{n\widehat{p}_\lambda} - e^{-Bn\widehat{p}_\lambda} \ . \tag{62}$$

2. Rademacher (p):

$$\frac{2}{1-p} \left[(\kappa_2 - 1) \wedge \left(\frac{\kappa_1}{\left(np\widehat{p}_\lambda \right)^{1/4}} \right) \right] \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRad}(p))} \geq \frac{-2e^{-pn\widehat{p}_\lambda}}{1-p} \tag{63}$$

$$\left(1 + 3 \times 10^{-4}\right) \wedge \left(\frac{\kappa_1 \times 2^{1/4}}{\left(n\widehat{p}_\lambda \right)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRad}(1/2))} \geq -\mathbb{1}_{n\widehat{p}_\lambda \leq 2} \ . \tag{64}$$

3. Poisson (μ):

$$1 \wedge \frac{2\left(1 + e^{-3}\right)}{\left(\mu n\widehat{p}_\lambda - 2 \right)_+} \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penPoi}(\mu))} \geq \frac{-2}{n\widehat{p}_\lambda} - \left(e^{-\mu n\widehat{p}_\lambda} \wedge \mathbb{1}_{\mu n\widehat{p}_\lambda < 1.61} \right) \ . \tag{65}$$

4. Random hold-out (q_n): on the one hand,

$$\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRho}(q_n))} = \frac{n}{n-q} \left(e^+_{\mathcal{H}(n, n\widehat{p}_\lambda, q_n)} - 1 \right) \geq \frac{e^{-B_- n\widehat{p}_\lambda}}{1 - B_+} \ ,$$

where the lower bounds assumes that $0 < B_- \leq q_n n^{-1} \leq B_+ < \infty$. On the other hand, under the same condition,

$$\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRho}(q_n))} \leq \frac{L}{B_-(1 - B_+)} \sqrt{\frac{\ln(n\widehat{p}_\lambda)}{n\widehat{p}_\lambda}}$$

provided that $n\widehat{p}_\lambda \geq L_{B_-,B_+}$. When $q_n = \lfloor n/2 \rfloor$, we combine this upper bound with (24).

5. Leave-one-out:

$$\frac{\mathbb{1}_{n\widehat{p}_\lambda \geq 2}}{n\widehat{p}_\lambda - 1} \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penLoo})} \geq -\mathbb{1}_{n\widehat{p}_\lambda = 1} \ . \tag{66}$$

□

A byproduct of the proof of Prop. 2 (combined with Lemma 9), is the following:

Lemma 11. *Assume that W is a weight vector among Efr, Rad, Poi, Rho and Loo. Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$, $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$ and $\mathrm{pen}(m)$ be defined by (7) with $C = C_W$ (cf. Tab. 2). Then, if $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} \geq 3$,*

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] \geq \frac{5}{4} \mathbb{E}^{\Lambda_m}\left[p_2(m)\right] \ . \tag{67}$$

Remark 11. Assuming that $\min_{\lambda\in\Lambda_m}\{\, n\widehat{p}_\lambda \,\} \geq T$ for some positive T, (67) still holds for:

- $\mathrm{Efr}(m_n)$ when $m_n n^{-1} \geq -T^{-1}\ln(3/4 - 2/T)$
- $\mathrm{Rad}(p)$ when $T \geq p^{-1}\ln[8/(3(1-p))]$
- $\mathrm{Poi}(\mu)$ when $T \geq 3$ and $\mu T \geq 1.61$
- $\mathrm{Rho}(q_n)$ when $T \geq n q_n^{-1}\ln[(4n)/(3(n-q_n))]$.

9.9. Concentration inequalities

In this subsection, we give concentration inequalities for both the ideal and resampling penalties.

9.9.1. Resampling penalties

proof of Prop. 3. According to (48), $\mathrm{pen}(m)$ is a U-statistics of order 2 conditionally to $(\mathbb{1}_{X_i\in I_\lambda})_{(i,\lambda)}$. Then, using Lemma 5 of (9), with

$$a_\lambda = \frac{R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda)}{n(n\widehat{p}_\lambda - 1)} \qquad b_\lambda = \frac{-\left(\, R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda)\,\right)}{n^2\widehat{p}_\lambda(n\widehat{p}_\lambda - 1)} \;,$$

we have, for all $q \geq 2$,

$$\begin{aligned}\left\|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right\|_q^{(\Lambda_m)} &\leq L_{a_\ell,\xi_\ell} D_m^{-1/2} A_n^{-1/2} \\ \times \sup_{np\geq A_n} \{\, R_{1,W}(n,p) + R_{2,W}(n,p)\,\}\, q^{\xi_\ell+1}\mathbb{E}\left[\, p_2(m)\,\right] \;.\end{aligned}$$

We deduce conditional concentration inequalities through the classical link between moments and concentration, with a probability bound $1 - n^{-\gamma}$. Since it is deterministic, this implies unconditional concentration inequalities.

The second statement follows from the proof of Prop. 2, where we can find non-asymptotic upper bounds on $2+\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} = C_W \times (\, R_{1,W}(n,\widehat{p}_\lambda) + R_{2,W}(n,\widehat{p}_\lambda)\,)$. □

9.9.2. Ideal penalty

Concentration properties for the ideal penalty are proven in (9), App. B.5, under the assumptions of Thm. 1. We here state them with assumption $(\mathbf{A_{m},}_\ell)$ instead of the boundedness assumption (**Ab**), so that they can be applied in the general framework of Sect. 9.2. Our first result has to deal with p_1 and p_2, which are the main components of the ideal penalty. Remark that concentration for p_2 can be obtained in a general framework (see (6), Chap. 7). On the contrary, we do not know any other non-asymptotic bound on the deviation of p_1.

Proposition 12. *Let $\gamma > 0$. Assume that $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n$ and $(\mathbf{A_m},\ell)$ $\forall q \geq 2$, $P_m^\ell(q) \leq a_\ell q^{\xi_\ell}$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{a_\ell,\xi_\ell,\gamma}\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{68}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{a_\ell,\xi_\ell,\gamma}\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{69}$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{a_\ell,\xi_\ell,\gamma}\frac{\ln(n)^{\xi_\ell+1}}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right] \ . \tag{70}$$

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

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n^{-1}\ln(n)} - L_{a_\ell,\xi_\ell,\gamma}\left[\frac{\ln(n)^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\right)\mathbb{E}\left[\widetilde{p_2}(m)\right] \ . \tag{71}$$

proof of Prop. 12. See the proof of Prop. 9 in (9), which only differ in the way of upper bounding $m_{q,\lambda}$, since the data is no longer assumed to be bounded. □

We now come to the centered term $\overline{\delta}(m)$. When the data is bounded, we have the following:

Lemma 13 (Prop. 3 of (11)). *Assume that $\|Y\|_\infty \leq A < \infty$. Then for all $x \geq 0$, there is an event of probability at least $1 - 2e^{-x}$ on which, for every $\eta > 0$,*

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta\ell\left(s, s_m\right) + \left(\frac{4}{\eta} + \frac{8}{3}\right)\frac{A^2 x}{n} \ . \tag{72}$$

In particular,

$$\left|\overline{\delta}(m)\right| \leq \frac{\ell\left(s, s_m\right)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q_m^{(p)}}\frac{\mathbb{E}^{\Lambda_m}[p_2(m)]}{\sqrt{D_m}}x \ . \tag{73}$$

proof of Lemma 13. We refer to (11) for a detailed proof of (72), which essentially relies on Bernstein's inequality. Then, we deduce (73) by choosing $\eta = D_m^{-1/2}$ and using the definition of $Q_m^{(p)}$. □

In the unbounded case, we need another concentration inequality for $\overline{\delta}(m)$. There are many strategies for this, since it is a sum of i.i.d. centered random variables. In our framework, the following result is sufficient.

Lemma 14. *Assume that $(\mathbf{A_g},_\epsilon)$ $\forall q \geq 2$, $P^{g\epsilon}(q) \leq a_{g\epsilon}q^{\xi_{g\epsilon}}$, $(\mathbf{A}\sigma_{\max})$ $\|\sigma(X)\|_\infty \leq \sigma_{\max}$ and $(\mathbf{A}\delta)$ $\|s - s_m\|_\infty \leq c^g_{\Delta,m}\|s(X) - s_m(X)\|_2$. Then, for every $x \geq 0$, there exists an event of probability at least $1 - e^{-x}$ on which*

$$\left|\overline{\delta}(m)\right| \leq \frac{L_{a_{g\epsilon},\xi_{g\epsilon},c^g_{\Delta,m}}x^{\xi_{g\epsilon}+1/2}}{\sqrt{D_m}}\left[\ell\left(s, s_m\right) + \frac{\sigma^2_{\max}}{Q_m^{(p)}}\mathbb{E}\left[p_2(m)\right]\right] \ . \tag{74}$$

On the other hand, if $(\mathbf{A}_{\mathbf{g},\epsilon})$ *and* $(\mathbf{A}\sigma_{\max})$ *holds true, but* $(\mathbf{A}\delta)$ *is replaced by* $(\mathbf{As}_{\max})$ $\|s\|_\infty \leq A$*, then, for every* $x \geq 0$*, there is an event of probability at least* $1 - e^{-x}$ *on which*

$$\left|\overline{\delta}(m)\right| \leq L_{a_{g\epsilon},\xi_{g\epsilon},A,\sigma_{\max}} n^{-1/2} x^{\xi_{g\epsilon}+1/2} \ . \tag{75}$$

proof of Lemma 14. From Lemma 8.18 of (6) (which is for instance a consequence of (55), Sect. 5.3.5), we have

$$\begin{aligned}
\left\|\overline{\delta}(m)\right\|_q &\leq \frac{2\sqrt{\kappa}\sqrt{q}}{\sqrt{n}} \left\|F_m - \mathbb{E}[F_m]\right\|_q \\
\text{with} \quad F_m &:= (Y - s_m(X))^2 - (Y - s(X))^2 \\
&= (s_m(X) - s(X))^2 - 2\epsilon\sigma(X)(s_m(X) - s(X)) \ .
\end{aligned}$$

Notice that $\epsilon\sigma(X)(s_m(X) - s(X))$ is centered conditionally to $X \in I_\lambda$ for all $\lambda \in \Lambda_m$. We thus have

$$\left\|\overline{\delta}(m)\right\|_q \leq \frac{2\sqrt{\kappa}\sqrt{q}}{\sqrt{n}} \left(\|s - s_m\|_\infty^2 + 2\sigma_{\max} \|s - s_m\|_\infty \|\epsilon\|_q \right) \ . \tag{76}$$

We now use assumptions $(\mathbf{A}_{\mathbf{g},\epsilon})$ and $(\mathbf{A}\delta)$. Then, for all $q \geq 2$,

$$\begin{aligned}
\left\|\overline{\delta}(m)\right\|_q &\leq 2\sqrt{\kappa}\sqrt{q} \left((c^g_{\Delta,m})^2 \ell(s, s_m) + 2c^g_{\Delta,m} \sqrt{\ell(s, s_m)} P^{g\epsilon}(q) \sigma_{\max} \right) \frac{1}{\sqrt{n}} \\
&\leq L_{c^g_{\Delta,m}} \sqrt{q} D_m^{-1/2} \ell(s, s_m) + L_{a_{g\epsilon},\xi_{g\epsilon},c^g_{\Delta,m}} q^{\xi_{g\epsilon}+1/2} \frac{\sigma^2_{\max}\sqrt{D_m}}{n} \ .
\end{aligned}$$

We take $\theta = D_m^{-1/2}$ and deduce (74) with the classical link between moments and concentration. For the second statement, start back from (76) and use that $\|s - s_m\|_\infty \leq 2A$. □

9.10. Technical lemmas

The three following lemmas are needed in our proofs. They are proven in the technical appendix (7).

Empirical and expected frequencies

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 15. *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 inequality holds:*

$$\min_{\lambda\in\Lambda_m} \{n\widehat{p}_\lambda\} \geq \frac{\min_{\lambda\in\Lambda_m}\{np_\lambda\}}{2} - 2(\gamma+1)\ln(n) \tag{77}$$

Proof. This relies mainly on Bernstein's inequality for each $n\widehat{p}_\lambda$. See Lemma 12 in (9) for a detailed proof. □

Bounds for $Q_m^{(p)}$

Besides the straightforward lower bound $\sigma_{\min}^2$ when (**An**), holds, we have the following.

Lemma 16. *Recall that*

$$Q_m^{(p)} := \frac{1}{D_m} \sum_{\lambda \in \Lambda_m} \sigma_\lambda^2 \ .$$

If $\mathcal{X} \subset \mathbb{R}^k$, $(\mathbf{Ar}_\mathbf{u}^\mathbf{d})$ $\max_{\lambda \in \Lambda_m} \{ \operatorname{diam}(I_\lambda) \} \leq c_{\mathrm{r,u}}^d D_m^{-\alpha_d} \operatorname{diam}(X)$, $(\mathbf{Ar_u})$ $\max_{\lambda \in \Lambda_m} \{ \operatorname{Leb}(I_\lambda) \} \leq c_{\mathrm{r,u}} D_m^{-1} \operatorname{Leb}(\mathcal{X})$, *and* σ *is piecewise* K_σ*-Lipschitz with at most* J_σ *jumps* $(\mathbf{A}\sigma)$, *then*

$$Q_m^{(p)} \geq \frac{\|\sigma\|_{L^2(\mathrm{Leb})}^2}{2c_{\mathrm{r,u}}} - \frac{K_\sigma^2 \left(c_{\mathrm{r,u}}^d \right)^2 \operatorname{diam}(\mathcal{X})^2}{D_m^{2\alpha_d}} - \frac{J_\sigma \|\sigma(X)\|_\infty^2}{2D_m} \ .$$

Remark 12. Since $\|\sigma(X)\|_2 > 0$ and σ is piecewise Lipschitz, $\|\sigma\|_{L^2(\mathrm{Leb})} > 0$. Thus, the lower bound for $Q_m^{(p)}$ is positive when D_m is large enough.

Sufficient condition for $(\mathbf{A}\delta)$

Lemma 17. *Assume that* $\mathcal{X} \subset \mathbb{R}$ *is bounded and:*

(**Al**) $s : \mathcal{X} \mapsto \mathbb{R}$ *is* B*-Lipschitz, piecewise* C^1 *and non-constant (*i.e. $\pm s' \geq B_0 > 0$ *on some interval* $J \subset \mathcal{X}$ *with* $\operatorname{Leb}(J) \geq c_J \operatorname{Leb}(\mathcal{X})$, *with* $c_J > 0$*).*

($\mathbf{Ar}_{\ell,\mathbf{u}}$) *Regularity of the partitions for* Leb*:*

$$\forall \lambda \in \Lambda_m, \quad c_{\mathrm{r},\ell} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \leq \operatorname{Leb}(I_\lambda) \leq c_{\mathrm{r,u}} D_m^{-1} \operatorname{Leb}(\mathcal{X}) \ .$$

($\mathbf{Ad}_\ell$) *Density bounded from below:* $\exists c_X^{\min} > 0$, $\forall I \subset \mathcal{X}$, $P(X \in I) \geq c_X^{\min} \operatorname{Leb}(I) \operatorname{Leb}(\mathcal{X})^{-1}$.

Then, $(\mathbf{A}\delta)$ *holds true,* i.e.*, for every model* m *of dimension* $D_m \geq D_0$,

$$\|s - s_m\|_\infty \leq c_{\Delta,m}^g \|s(X) - s_m(X)\|_2$$

$$\text{with} \quad c_{\Delta,m}^g = \left(\frac{c_{\mathrm{r,u}}}{c_{\mathrm{r},\ell}} \right)^{3/2} \frac{B\sqrt{24}}{B_0 \sqrt{c_X^{\min} c_J}} \quad \text{and} \quad D_0 := 4 c_{\mathrm{r,u}} c_J^{-1} \ .$$

9.11. Expectations of inverses

We prove in this subsection the lemmas of Sect. 4.4.3. Notice that Sect. 2 in (8) explains how to generalize (20) to a wide class of random variables. We will use two results that can be found there: the general lower bound

$$e_Z^+ \geq \mathbb{P}(Z > 0) \ , \tag{78}$$

which comes from Jensen inequality. Defining

$$e_{\mathcal{L}(Z)}^0 := \mathbb{E}\left[Z \right] \mathbb{E}\left[Z^{-1} \mathbb{1}_{Z>0} \right] = e_Z^+ \mathbb{P}(Z > 0) \ , \tag{79}$$

we have the following upper bound, which holds as soon as $\mathbb{P}(c_Z > Z > 0) = 0$.

$$\begin{aligned}\forall \alpha > 0, \quad e_Z^0 &= \mathbb{E}\left[Z^{-1} \mathbb{1}_{\alpha \mathbb{E}[Z] > Z > 0} \right] \mathbb{E}[Z] + \mathbb{E}\left[Z^{-1} \mathbb{1}_{Z \geq \alpha \mathbb{E}[Z]} \right] \mathbb{E}[Z] \\ &\leq \mathbb{P}\left(\alpha \mathbb{E}[Z] > Z > 0 \right) \mathbb{E}[Z] c_Z^{-1} + \alpha^{-1} \enspace . \end{aligned} \tag{80}$$

9.11.1. Binomial case (proof of Lemma 4)

We only have to prove (21). When $n \geq 9$, the upper bound follows from (79) together with Lemma 4.1 of (37) (showing that $e^0_{\mathcal{B}(n,p)} \leq 2n/(n+1)$). When $n \leq 8$, $e^+_{\mathcal{B}(n,1/2)} \leq 1.21$ (see for instance Sect. 8.7 of (6)). For the lower bound, the crucial point is that $Z \sim \mathcal{B}\left(n, \frac{1}{2}\right)$ is nonnegative and symmetric, *i.e.* $\mathcal{L}(Z) = \mathcal{L}(n - Z)$. Using only this property, and defining $p_0 = \mathbb{P}(Z = 0) = \mathbb{P}(Z = n) = 2^{-n}$, we have

$$\begin{aligned} e_Z^+ &= \frac{\mathbb{P}\left(Z = n \,|\, Z > 0 \right)}{2} + \mathbb{E}\left[\frac{1}{Z} \,\middle|\, 0 < Z < 2 \right] \frac{n}{2} \frac{\mathbb{P}(0 < Z < n)}{\mathbb{P}(Z > 0)} \\ &= \frac{p_0}{2(1-p_0)} + \frac{1 - 2p_0}{1 - p_0} \frac{n}{2} \mathbb{E}\left[\frac{1}{2} \left(\frac{1}{Z} + \frac{1}{n - Z} \right) \,\middle|\, 0 < Z < n \right] \\ &= \frac{p_0}{2(1-p_0)} + \frac{1 - 2p_0}{1 - p_0} \left(1 + \frac{n}{2} \mathbb{E}\left[\frac{\left(Z - \frac{n}{2} \right)^2}{Z(n-Z)} \,\middle|\, 0 < Z < n \right] \right) \enspace . \end{aligned} \tag{81}$$

Since Z is binomial with parameters $(n, 1/2)$, we have

$$\frac{n(1-2p_0)}{2} \mathbb{E}\left[\frac{\left(Z - \frac{n}{2} \right)^2}{Z(n-Z)} \,\middle|\, 0 < Z < n \right] \geq \mathbb{P}\left(Z = 1 \text{ or } Z = n - 1 \right) \frac{(n-2)^2}{4(n-1)}$$

if $n \geq 3$. Putting this into (81), we obtain:

$$e^+_{\mathcal{B}\left(n, \frac{1}{2}\right)} \geq \frac{1}{1 - 2^{-n}} \left(2^{-n-1} + 1 - 2^{1-n} + \frac{n(n-2)^2}{2^{n+1}(n-1)} \right) \geq 1 \enspace . \quad \square$$

9.11.2. Hypergeometric case (proof of Lemma 5)

Let $Z \sim \mathcal{H}(n, r, q)$. It has an expectation $\mathbb{E}[Z] = \frac{qr}{n}$.

General lower bound We first use (78) and

$$\mathbb{P}(Z = 0) \leq \left(1 - \frac{r}{n} \right)^q \leq \exp\left(-\frac{qr}{n} \right) \enspace .$$

Moreover, if $r \geq n - q + 1$, $\mathbb{P}(Z > 0) = 1$.

A general upper bound According to (79) and the lower bound for $\mathbb{P}(Z>0)$ above, it is sufficient to upper bound $e^0_{\mathcal{H}(n,r,q)}$. We first prove the following general result, that holds for every $n \geq r, q \geq 1$:

$$e^+_{\mathcal{H}(n,r,q)} \leq \frac{\inf_{\frac{q}{n}>\beta\geq\frac{2}{r}} \left\{ \frac{qr}{n} \exp\left[-\frac{2(\beta r-1)^2}{r+1} \right] + \frac{1}{1-\frac{n\beta}{q}} \right\}}{1-\exp\left(-\frac{qr}{n}\right)} \tag{82}$$

The idea of the proof is to use (80) with $c_Z = 1$, $\mathbb{E}[Z] = qrn^{-1}$. For this, we need the following concentration result by Hush and Scovel (41): for all $x \geq 2$,

$$\mathbb{P}\left(\mathbb{E}(Z) - Z > x\right) < \exp\left(-2(x-1)^2 \left[\left(\frac{1}{r+1} + \frac{1}{n-r+1} \right) \vee \left(\frac{1}{q+1} + \frac{1}{n-q+1} \right) \right] \right) .$$

Taking $\alpha = 1 - \frac{n\beta}{q}$ with $\frac{q}{n} > \beta \geq \frac{2}{r}$, we obtain

$$e^0_{\mathcal{H}(n,r,q)} \leq \frac{qr}{n} \exp\left[-\frac{2(\beta r-1)^2}{r+1} \right] + \frac{1}{1-\frac{n\beta}{q}} .$$

As a consequence, (82) holds.

Back to (23) With the supplementary conditions on n, r and q, we can take $\beta = \frac{1+\sqrt{\frac{3}{4}\ln(r)(r+1)}}{r}$ in (82). Hence

$$e^0_{\mathcal{H}(n,r,q)} \leq \frac{1}{2\sqrt{r}} + \frac{1}{1-\frac{n}{q}\left(\frac{1+\sqrt{\frac{3}{4}\ln(r)(r+1)}}{r} \right)} \leq 1 + \frac{n}{q} K(\epsilon) \sqrt{\frac{\ln(r)}{r}}$$

$$\text{with} \quad K(\epsilon) = \frac{1}{2\sqrt{\ln(2)}} + \frac{1}{\epsilon^2}\left(\sqrt{\frac{\ln(3)}{3}} + \frac{3}{4} \right) .$$

Using (79) and the upper bound on $\mathbb{P}(Z=0)$, we deduce (23) since $r \geq 2$ and

$$\kappa_3(\epsilon) = 0.9 + 1.4 \times \epsilon^{-2} \geq 1.02 \times K(\epsilon) + 0.03 .$$

"Rho" case We now assume that $q = \lfloor \frac{n}{2} \rfloor$ so that $\frac{n}{q} = 2 + \frac{1}{\lfloor \frac{n}{2} \rfloor} \leq 3$ and converges to 2 when n goes to infinity.

For $r \geq 6$, we can take $\beta = \frac{2}{r}$ in (82) and we obtain:

$$e^+_{\mathcal{H}(n,6,q)} \leq 9.68 \qquad e^+_{\mathcal{H}(n,7,q)} \leq 7.61 \qquad e^+_{\mathcal{H}(n,8,q)} \leq 7.46 \qquad e^+_{\mathcal{H}(n,9,q)} \leq 7.32$$

For $r \geq 10$, taking $\beta = \frac{1}{4} + \frac{1}{r}$ in (82), we derive

$$\sup_{r\geq 10} e^+_{\mathcal{H}(n,r,q)} \leq 7.49 \qquad \sup_{r\geq 26} e^+_{\mathcal{H}(n,r,q)} \leq 3 .$$

Small values of r must be treated appart. For $r=1$, it is easy to compute $e^+_{\mathcal{H}(n,1,q)} = qn^{-1} \leq 1$. When $n=r$, we have $e^+_{\mathcal{H}(n,n,q)} = 1$. Otherwise, using the fact that for every $n \geq r+1$, $\frac{n!}{(n-r)!} \geq \frac{(r+1)!}{(r+1)^r} n^r$,

$$e^0_{\mathcal{H}(n,r,q)} \leq \frac{r}{R} \frac{(r+1)^r}{(r+1)! R^r} \left(\sum_{k=1}^{r} \binom{r}{k} \frac{(R-1)^{r-k}}{k} \right)$$

with $R = \frac{n}{q} \in [1; +\infty)$. For $r=2$, this upper bound is lower than 1.6. If $\frac{n}{q} \leq 3$ (which holds in the "Rho" case),

$$e^+_{\mathcal{H}(n,3,q)} \leq 4.67 \qquad e^+_{\mathcal{H}(n,4,q)} \leq 8.15 \qquad e^+_{\mathcal{H}(n,5,q)} \leq 14.29 \ .$$

"Loo" case We now have $q = n-1$. We first consider $r=1$. The conditioning make Z deterministic and equal to 1, so that

$$e^+_{\mathcal{H}(n,1,n-1)} = \mathbb{E}[Z] = 1 - \frac{1}{n} \ .$$

Now, if $r \geq 2$, $Z > 0$ holds a.s. since it only take two values:

$$\mathbb{P}\left(Z = r-1 \right) = \frac{r}{n} \quad \text{and} \quad \mathbb{P}\left(Z = r \right) = \frac{n-r}{n} \ .$$

As a consequence,

$$e^+_{\mathcal{H}(n,r,n-1)} = \frac{(n-1)r}{n} \left(\frac{r}{(r-1)n} + \frac{n-r}{nr} \right) = 1 + \frac{1}{n} \left(\frac{(n-1)r}{n(r-1)} - 1 \right) \ .$$

The lower bound is straightforward since $n \geq r$.

"Lpo" case As noticed in Lemma 10, we have

$$\forall r \geq p+1, \qquad e^+_{\mathcal{H}(n,r,n-p)} \geq 1 \ .$$

Moreover, when $r \geq p+1$, $\mathcal{H}(n,r,n-p)$ has its support in $\{r-p, \dots, r\}$ and thus

$$\begin{aligned} e^+_{\mathcal{H}(n,r,n-p)} &= \frac{(n-p)r}{n} \sum_{j=r-p}^{r} \frac{\binom{r}{j}\binom{n-r}{n-p-j}}{j\binom{n}{n-p}} \\ &= \frac{(n-p)r}{n} \sum_{k=(p+r-n)\vee 0}^{p} \frac{\binom{r}{k}\binom{n-r}{p-k}}{(r-k)\binom{n}{p}} \ . \end{aligned}$$

More precisely, the k-th term of the sum is equal to

$$\frac{(n-p)r}{n} \frac{\binom{r}{k}\binom{n-r}{p-k}}{(r-k)\binom{n}{p}} \leq \left(\frac{r}{n} \right)^k \left(1 - \frac{r}{n} \right)^{p-k} \binom{p}{k} \frac{r}{r-p} \frac{n^p}{n \cdots (n-p+1)} \ ,$$

so that

$$e^+_{\mathcal{H}(n,r,n-p)} \leq \frac{r n^p}{(r-p) n \cdots (n-p+1)} \ .$$

The result follows. □

Remark 13 *(Asymptotics).* If for some $\alpha > 0$, $q_k r_k^{1/2-\alpha} n_k^{-1} \xrightarrow[k\to+\infty]{} +\infty$ and $n_k \geq r_k \to +\infty$, then $e^+_{\mathcal{H}(n_k,r_k,q_k)} \to 1$ when $k \to \infty$. The upper bound is obtained by taking

$$\beta = \frac{1 + \sqrt{(r+1)\ln\left(\frac{qr}{n}\right)}}{r}$$

in (82) (it's possible for r sufficiently large). The lower bound is straightforward.

9.11.3. Poisson case (proof of Lemma 6)

Let $Z \sim \mathcal{P}(\mu)$, and define $g : [0;\infty) \mapsto \mathbb{R}$ by $g(0) = 0$ and for every $\mu > 0$

$$g(\mu) := e^+_{\mathcal{P}(\mu)} = \mu \mathbb{E}\left[Z^{-1} \,\middle|\, Z > 0 \right] = \frac{\mu e^{-\mu}}{1-e^{-\mu}} \sum_{k=1}^{+\infty} \frac{\mu^k}{k \times k!} = \frac{\mu}{e^\mu - 1} \int_0^\mu \frac{e^x - 1}{x} dx \ .$$

The function g is continuous at 0 and has a first derivative $g'(0) = 1$. For every $x \geq 0$, we define

$$h(x) = \frac{e^x - 1}{x} \qquad H(x) = \int_0^x h(t)dt \qquad a(x) = \frac{h'(x)}{h(x)} = 1 - \frac{e^x - 1 - x}{x(e^x - 1)} \ .$$

where the last equality holds if $x > 0$, and $a(0) = 1/2$. Then, $g(u) = H(u)/h(u)$ satisfies the following ordinary differential equation:

$$g(0) = 0 \qquad \forall u \geq 0, \quad g'(u) = 1 - a(u)g(u) \ .$$

Since

$$\forall u \geq 0, \quad \frac{1}{2} \leq a(u) \leq 1 \qquad \text{and} \qquad \lim_{u\to+\infty} a(u) = 1 \ ,$$

g satisfies a differential inequation

$$1 - \frac{g}{2} \leq g' \leq 1 - g \qquad g(0) = 0 \ .$$

Then, for every $x \geq x_0 \geq 0$,

$$2\left[1 - e^{2(x_0-x)}\left(1 - \frac{g(x_0)}{2}\right)\right] \geq g(x) \geq 1 + (g(x_0) - 1)e^{x_0 - x} \ . \qquad (83)$$

Lower bound The general lower bound (78) gives

$$g(\mu) \geq \mathbb{P}(Z > 0) = 1 - e^{-\mu} \ .$$

We can do better: remark that if $g(x_0) \geq 1$, (83) shows that $g(x) \geq 1$ for every $x \geq x_0$. Since $g = H/h$ and for every $u \geq 0$,

$$H(u) \geq u + \frac{u^2}{4} + \frac{u^3}{18} \ , \text{ we deduce that } \quad g(u) \geq \frac{u\left(u + \frac{u^2}{4} + \frac{u^3}{18}\right)}{e^u - 1} \ .$$

Then, $g(1.61) \geq 1$, so that $g(x) \geq 1$ for every $x \geq 1.61$.

Upper bound Using (83) with $x_0 = 0$ gives

$$\forall x \geq 0, \quad g(x) \leq 2 - 2e^{-2x} \leq 2.$$

Moreover, for every $\epsilon \in (0;1)$, $1 - \epsilon \leq a(x) \leq 1$ as soon as $x \geq \epsilon^{-1}$. Then, on $[\epsilon^{-1}; \infty)$, g satisfies the differential inequation

$$g' \geq 1 - (1-\epsilon)g \ .$$

Integrating this between ϵ^{-1} and $2\epsilon^{-1}$, we obtain that

$$g(2\epsilon^{-1}) \leq \frac{1}{1-\epsilon}\left[1 + \left(g(\epsilon^{-1})(1-\epsilon) - 1\right)\exp\left(-\epsilon^{-1}(1-\epsilon)^{-1}\right)\right] \ .$$

For every $x > 2$, $\epsilon = 2x^{-1} \in (0;1)$ so that

$$g(x) \leq 1 + \frac{2 + (x-4)\exp\left(-\frac{x^2}{2(x-2)}\right)}{x-2} \leq 1 + \frac{2(1+e^{-3})}{x-2} \ .$$

The result follows. □

Acknowledgments

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

References

[1] Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing the fit of a parametric function. *J. Amer. Statist. Assoc.* **94**, 447, 869–879. MRMR1723323 (2000g:62173)

[2] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math. 22*, 203–217. MRMR0286233 (44 #3447)

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest, 267–281. MRMR0483125 (58 #3144)

[4] Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics 16*, 125–127. MRMR0343481 (49 #8222)

[5] Arcones, M. A. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist. Wiley, New York, 13–47. MRMR1197777 (94b:62031)

[6] Arlot, S. (2007). Resampling and model selection. Ph.D. thesis, University Paris-Sud 11. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[7] Arlot, S. (2008a). Technical appendix to "Model selection by resampling penalization".
[8] Arlot, S. (2008b). Technical appendix to "V-fold cross-validation improved: V-fold penalization". hal-00239182.
[9] Arlot, S. (2008c). V-fold cross-validation improved: V-fold penalization. arXiv:0802.0566.
[10] Arlot, S., Blanchard, G., and Roquain, É. (2007). Non-asymptotic resampling-based confidence regions and multiple tests in high dimension. Long version of the COLT 2007 paper. arXiv:math.ST/07120775.
[11] Arlot, S. and Massart, P. (2008). Slope heuristics for heteroscedastic regression on a random design. arXiv:0802.0837.
[12] Audibert, J.-Y. (2004). Théorie statistique de l'apprentissage : une approche PAC-bayésienne. Ph.D. thesis, Université Paris VI.
[13] Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117**, 4, 467–493. MRMR1777129 (2001i:62048)
[14] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist. 6*, 127–146 (electronic). MRMR1918295 (2003h:62062)
[15] Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*. Lecture Notes in Statistics, Vol. **98**. Springer-Verlag, New York. MRMR2195545
[16] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 3, 301–413. MRMR1679028 (2000k:62049)
[17] Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning 48*, 85–113.
[18] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33**, 4, 1497–1537. MRMR2166554 (2006h:62042)
[19] Bartlett, P. L., Mendelson, S., and Philips, P. (2004). Local complexities for empirical risk minimization. In *Learning theory*. Lecture Notes in Comput. Sci., Vol. **3120**. Springer, Berlin, 270–284. MRMR2177915 (2006h:62033)
[20] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3**, 3, 203–268. MR2002i:62072
[21] Birgé, L. and Massart, P. (2006). Minimal penalties for gaussian model selection. *Probab. Theory Related Fields* **134**, 3.
[22] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. MRMR726392 (86b:62101)
[23] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 3, 503–514. MRMR1040644 (91e:62080)
[24] Burman, P. (2002). Estimation of equifrequency histograms. *Statist. Probab. Lett.* **56**, 3, 227–238. MRMR1892984 (2002m:62065)
[25] Catoni, O. (2007). *Pac-Bayesian Supervised Classification: The Ther-*

modynamics of Statistical Learning. IMS Lecture Notes Monograph Series, Vol. **56**. Inst. Math. Statist. doi:10.1214/074921707000000391.

[26] Cavanaugh, J. E. and Shumway, R. H. (1997). A bootstrap variant of AIC for state-space model selection. *Statist. Sinica* **7**, 2, 473–496. MRMR1466691

[27] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 432, 1200–1224. MRMR1379464 (96k:62093)

[28] Efromovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6**, 4, 925–942. MRMR1422411 (98b:62060)

[29] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1, 1–26. MRMR515681 (80b:62021)

[30] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 382, 316–331. MRMR711106 (84k:62039)

[31] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 394, 461–470. MRMR845884 (88a:62190)

[32] Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92**, 438, 548–560. MRMR1467848 (98c:62083)

[33] Fromont, M. (2004). Model selection by bootstrap penalization for classification. In *Learning theory*. Lecture Notes in Comput. Sci., Vol. **3120**. Springer, Berlin, 285–299. MRMR2177916

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

[35] Galtchouk, L. and Pergamenshchikov, S. (2005). Efficient adaptive nonparametric estimation in heteroscedastic models. Université Louis Pasteur, IRMA, Preprint.

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

[37] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York. MRMR1920390 (2003g:62006)

[38] Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York. MRMR1145237 (93h:62029)

[39] Hall, P. and Mammen, E. (1994). On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.* **22**, 4, 2011–2030. MRMR1329180 (96d:62071)

[40] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction. MRMR1851606 (2002k:62048)

[41] Hush, D. and Scovel, C. (2005). Concentration of the hypergeometric distribution. *Statist. Probab. Lett.* **75**, 2, 127–132. MRMR2206293

[42] Hušková, M. and Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *Ann. Statist.* **21**, 4, 1811–1823.

MRMR1245770 (94i:62020)

[43] Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.* **49**, 3, 411–434. MRMR1482365 (99c:62040)

[44] Jones, C. M. and Zhigljavsky, A. A. (2004). Approximating the negative moments of the Poisson distribution. *Statist. Probab. Lett.* **66**, 2, 171–181. MRMR2029732 (2004j:62025)

[45] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory* **47**, 5, 1902–1914. MRMR1842526 (2002e:62064)

[46] Koltchinskii, V. (2006). 2004 IMS Medallion Lecture: Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.* **34**, 6.

[47] Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*. Lecture Notes in Statistics, Vol. **82**. Springer-Verlag, New York. MRMR1226450 (95a:62028)

[48] Lew, R. A. (1976). Bounds on negative moments. *SIAM J. Appl. Math.* **30**, 4, 728–731. MRMR0501260 (58 #18663)

[49] Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 3, 958–975. MRMR902239 (89c:62112)

[50] Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32**, 4, 1679–1697. MRMR2089138 (2005h:62169)

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

[52] Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations*. Lecture Notes in Statistics, Vol. **77**. Springer.

[53] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 6, 1808–1829. MRMR1765618 (2001i:62074)

[54] Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20**, 3, 1611–1624. MRMR1186268 (93h:62091)

[55] Massart, P. (2007). *Concentration inequalities and model selection*. Lecture Notes in Mathematics, Vol. **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MRMR2319879

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

[57] Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York. MRMR1707286 (2001d:62047)

[58] Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 4, 2053–2086. MRMR1245301 (94k:60054)

[59] Sauvé, M. (2006). Histogram selection in non gaussian regression. Tech. Rep. 5911, INRIA. may.

[60] Sauvé, M. and Tuleau, C. (2006). Variable selection through cart. Tech. rep., INRIA. May.

[61] Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91**, 434, 655–665. MRMR1395733 (97f:62149)

[62] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 2, 221–264. With comments and a rejoinder by the author. MRMR1466682 (99m:62104)

[63] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 1, 45–54. MRMR614940 (84a:62103a)

[64] Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica* **7**, 2, 375–394. MRMR1466687 (98k:62083)

[65] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 6, 1348–1360. MRMR594650 (83d:62068)

[66] Stone, C. J. (1985). An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 513–520. MRMR822050 (87d:62098)

[67] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B 36*, 111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors. MRMR0356377 (50 #8847)

[68] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MRMR1385671 (97g:60035)

[69] Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 4, 1261–1350. With discussion and a rejoinder by the author. MRMR868303 (88f:62106)

[70] Yang, Y. (2006). Comparing learning methods for classification. *Statist. Sinica* **16**, 2, 635–657. MRMR2267253

[71] Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. Accepted by Annals of Statistics.

[72] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 5, 1564–1599. MRMR1742500 (2001g:62006)

[73] Žnidarič, M. (2005). Asymptotic expansions for inverse moments of binomial and poisson distributions. arXiv:math.ST/0511226.