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
CNRS; Willow Project-Team
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France
e-mail: sylvain.arlot@ens.fr

Abstract: In this paper, a new family of resampling-based penalization procedures for model selection is defined in a general framework. It generalizes several methods, including Efron's bootstrap penalization and the leave-one-out penalization recently proposed by Arlot (2008), to any exchangeable weighted bootstrap resampling scheme. In the heteroscedastic regression framework, assuming the models to have a particular structure, these resampling penalties are proved to satisfy a non-asymptotic oracle inequality with leading constant close to 1. In particular, they are asymptotically optimal. Resampling penalties are used for defining an estimator adapting simultaneously to the smoothness of the regression function and to the heteroscedasticity of the noise. This is remarkable because resampling penalties are general-purpose devices, which have not been built specifically to handle heteroscedastic data. Hence, resampling penalties naturally adapt to heteroscedasticity. A simulation study shows that resampling penalties improve on V-fold cross-validation in terms of final prediction error, 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, exchangeable weighted bootstrap, model selection, penalization, non-parametric regression, adaptivity, heteroscedastic data, regressogram, histogram selection.

Contents

*The author was financed in part by Univ Paris-Sud (Laboratoire de Mathematiques, CNRS-UMR 8628).

1. Introduction

In the last decades, model selection has received much interest. When the final goal is prediction, model selection can be seen more generally as the question of choosing between the outcomes of several prediction algorithms. With such a general formulation, a natural and classical answer is the following. First, estimate the prediction error for each model or algorithm; second, select the model minimizing this criterion. Model selection procedures mainly differ on the way of estimating the prediction error.

The empirical risk, also known as the apparent error or the resubstitution error, is a natural estimator of the prediction error. Nevertheless, minimizing the empirical risk can fail dramatically: the empirical risk is strongly biased for models involving a number of parameters growing with the sample size because the same data are used for building predictors and for comparing them.

In order to correct this drawback, *cross-validation* methods have been introduced [4, 65], relying on a data-splitting idea for estimating the prediction error with much less bias. In particular, V-fold cross-validation (VFCV, [36]) is a popular procedure in practice because it is both general and computationally tractable. A large number of papers exist about the properties of cross-validation methods, showing that they are efficient for a suitable choice of the way data are split (or V for VFCV). Asymptotic optimality results for leave-one-out cross-validation (that is the $V = n$ case) in regression have been proved for instance by Li [49] and by Shao [60]. However, when V is fixed, VFCV can be asymptotically suboptimal, as showed 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*. In short, penalization selects the model minimizing the sum of the empirical risk and of some measure of complexity[1] of the model (called penalty); see FPE [2], AIC [3], Mallows' C_p or C_L [51]. Model selection can target two different goals. On the one hand, a procedure is *efficient* (or asymptotically optimal) when its quadratic risk is asymptotically equivalent to the risk of the oracle. On the other hand, a procedure is *model consistent* when it selects the smallest true model asymptotically with probability one. This paper deals with *efficient* procedures, without assuming the existence of a true model. Therefore, the *ideal penalty* for prediction is the difference between the prediction error (the "true risk") and the empirical risk; penalties should be data-dependent estimates of the ideal penalty.

Many penalties or complexity measures have been proposed. Consider for instance regression and least-squares 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] is equal to $2n^{-1}\sigma^2 D$ for a model of dimension D and it can be modified according to the number of models [20, 58]. Mallows'

[1] Note that "complexity" here and in the following refers to the implicit modelization of a model or an algorithm, such as the number of estimated parameters. "Complexity" does not refer at all to the *computational* complexity of algorithms, which will always be called "computational complexity" in the following.

C_p-like penalties satisfy some optimality properties [61, 49, 14, 21] but they can fail when the data are heteroscedastic [7] because these penalties are linear functions of the dimension of the models.

In the binary supervised classification framework, several penalties have been proposed. First, VC-dimension-based penalties have the drawback of being independent of the underlying measure, so that they are adapted to the worst case. Second, global Rademacher complexities [45, 17] (generalized by Fromont with resampling ideas [33]) take into account the distribution of the data, but they are still too large to achieve fast rates of estimation when the margin condition [53] holds. Third, local Rademacher complexities [18, 46] are tighter estimates of the ideal penalty, but their computational cost is heavy and they involve huge (and sometimes unknown) constants. Therefore, easy-to-compute penalties that can achieve fast rates are still needed.

All the above penalties have serious drawbacks making them less often used in practice than cross-validation methods: AIC and Mallows' C_p rely on strong assumptions (such as homoscedasticity of the data and linearity of the models) and some mainly asymptotic arguments; VC-dimension-based penalties and global Rademacher complexities are far too pessimistic; local Rademacher complexities are computationally intractable, and their calibration is a serious issue. Another approach for designing penalties in the general framework may not suffer from these drawbacks: the *resampling* idea.

Efron's resampling heuristics [29] was first stated for the bootstrap, then generalized to the exchangeable weighted bootstrap by Mason and Newton [54] and by Præstgaard and Wellner [57]. In short, according to the resampling heuristics, the distribution of any function of the (unknown) distribution of the data and the sample can be estimated by drawing "resamples" from the initial sample. In particular, the resampling heuristics can be used to estimate the variance of an estimator [29], a prediction error [67, 32] or the ideal penalty (using the bootstrap [30, 31, 43], the M out of n bootstrap[2] [59] or a V-fold subsampling scheme [9]). The asymptotic optimality of Efron's bootstrap penalty for selecting among maximum likelihood estimators has been proved by Shibata [62]. Note also that global and local Rademacher complexities are using an i.i.d. Rademacher resampling scheme for estimating different upper bounds on the ideal penalty and Fromont's penalties [34] generalize the global Rademacher complexities to the exchangeable weighted bootstrap.

The first goal of this paper is to define and study *general-purpose penalties*, that is penalties well-defined in almost every framework and performing reasonably well in most of them, including regression and classification. The main interest of such penalties would be the ability to solve difficult problems (for instance heteroscedastic data, a non-smooth regression function or the fact that the oracle model achieves fast rates of estimation) *without knowing them in advance*. From the practical point of view, such a property is crucial.

To this aim, the resampling heuristics with the general exchangeable weighted bootstrap is used for estimating the ideal penalty (Section 2). This defines a

[2]Shao's goal in [59] was not efficiency but model consistency.

wide family of model selection procedures, called "Resampling Penalization" (RP), which includes Efron's and Shao's penalization methods [30, 59] as well as the leave-one-out penalization defined in [9]. To our knowledge, it has never been proposed with such general resampling schemes, so that the RP family contains a wide range of new procedures. Note that RP is well-defined in a general framework, including regression and classification, but also many other application fields (Section 7.2). Even if the main results are proved in the least-squares regression framework only, we obviously do not mean that RP should be restricted to this framework.

In this paper, the model selection efficiency of RP is studied with a *unified approach* for all the exchangeable resampling schemes. Therefore, comparing bootstrap with subsampling is quite straightforward (Section 5) which is not common in the resampling literature (except a few asymptotic results, see Barbe and Bertail [15]).

The point of view used in the paper is *non-asymptotic*, which has two major implications. First, non-asymptotic results allow to consider collections of models depending on the sample size n: in practice, it is usual to increase the number of explanatory variables with the number of observations. Considering models with a large number of parameters (for instance of order n^α for some $\alpha > 0$) is also particularly useful for designing adaptive estimators of a function which is only assumed to belong to some Hölderian ball (see Section 3.2). Thus, the non-asymptotic point of view allows not to assume that the regression function is described with a small number of parameters.

Second, several practical problems are "non-asymptotic" 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 by using the flexibility of penalization procedures. It is worth noting that such a non-asymptotic approach is not common in the model selection literature and few non-asymptotic results exist on general resampling methods.

Another important point is that the framework of the paper includes several kinds of *heteroscedastic data*. The observations $(X_i, Y_i)_{1 \leq i \leq n}$ are only assumed to be 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 zero mean and unit variance conditionally on X_i. In particular, the noise-level $\sigma(X)$ can strongly depend on X and the distribution of ϵ_i can depend on X_i. Such data are generally considered as difficult to handle because no information on σ is known, making irregularities of the signal difficult to distinguish from noise. As already mentioned, simple model selection procedures such as Mallows' C_p can fail in this framework [7] whereas it is natural to expect that resampling methods are robust to heteroscedasticity. In this article, both theoretical and simulation results confirm this fact (Sections 3 and 5).

The two main results of the paper are stated in Section 3. First, making mild assumptions on the distribution of the data, a non-asymptotic oracle in-

equality for RP is proved with leading constant close to 1 (Theorem 1). It holds 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 are highly heteroscedastic. For proving such a result, each model is assumed to be the vector space of piecewise constant functions (histograms) on some partition of the feature space. This is indeed a restriction, but we conjecture that it is mainly technical and that RP remains efficient in a much more general framework (see Section 7.2). Moreover, studying extensively the toy model of histograms allows to derive precise heuristics for the general framework. A major goal of the paper is to help practicioners who would like to know how to use resampling for performing model selection (see in particular Sections 6 and 7.3).

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

Several results similar to Theorem 1 exist in the literature for other procedures such as Mallows' C_p (with homoscedastic data only), VFCV and leave-one-out cross-validation. Moreover, there exist several minimax adaptive estimators for heteroscedastic data with a smooth noise-level, for instance [28, 35], and the regression function and the noise level can be estimated simultaneously [37]. In comparison, 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]), as detailed in Section 7.1.

A simulation study is conducted in Section 5 with small sample sizes. RP is showed to be competitive with Mallows' C_p for "easy" problems, and much better for some harder ones (for instance with a variable noise-level). Moreover, a well-calibrated RP yields almost always better model selection performance than VFCV. Therefore, RP can be of great interest in situations where no *a priori* information is known about the data. RP can deal with difficult problems, and compete with procedures that are fitted for easier problems. In short, RP is an efficient alternative to VFCV.

This article is organized as follows. The framework and the Resampling Penalization (RP) family of procedures are defined in Section 2. The main results are stated in Section 3. The differences between the resampling weights are investigated in Section 4. Then, a simulation study is presented in Section 5. Practical issues concerning the implementation of RP are considered in Section 6. RP is compared to other penalization methods in Section 7.1 and the extension of RP to the general framework is discussed in Section 7.2. Finally, Section 8 is devoted to the proofs. Some additional material (other simulation experiments and proofs) is available in a technical Appendix [8].

2. The Resampling Penalization procedure

In order to simplify the presentation, we choose to focus on the particular framework of least-squares regression on models of piecewise constant functions (histograms), which is the framework of the main results of Section 3 and the simulation study of Section 5.

Nevertheless, the RP family is a *general-purpose method* which can easily be defined in the general prediction framework. The main interest of the histogram framework is to provide general heuristics about RP, so that the practicioner can make the best possible use of RP in the general framework. A discussion on RP in the general prediction framework is provided in Section 7.2, including a general definition of RP.

2.1. Framework

Suppose we observe some data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$, independent with common distribution P, where the feature space $\mathcal{X}$ is typically a compact set of $\mathbb{R}^k$. Let s denote the regression function, that is $s(x) = \mathbb{E}[Y \mid X = x]$. Then,

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{1}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms; the ϵ_i possibly depend on X_i, but they are have zero mean and unit variance conditionally on X_i.

The goal is to predict Y given X where $(X, Y) \sim P$ is independent of the data. The quality of a predictor $t : \mathcal{X} \mapsto \mathbb{R}$ is measured by the quadratic prediction loss $P\gamma(t) := \mathbb{E}_{(X,Y)}[\gamma(t, (X, Y))]$, where $(X, Y) \sim P$ and $\gamma(t, (x, y)) := (t(x) - y)^2$ is the least-squares contrast. Since $P\gamma(t)$ is minimal when $t = s$, the excess loss is defined as

$$\ell(s, t) := P\gamma(t) - P\gamma(s) = \mathbb{E}_{(X,Y)}(t(X) - s(X))^2 .$$

Given a particular set of predictors S_m (called a *model*), the best predictor over S_m is defined as

$$s_m := \arg\min_{t \in S_m} \{P\gamma(t)\},$$

with its empirical counterpart

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

(when it exists and is unique) where $P_n = n^{-1} \sum_{i=1}^n \delta_{(X_i,Y_i)}$ is the empirical distribution. The estimator $\widehat{s}_m$ is the well-known *empirical risk minimizer*, also called least-squares estimator since γ is the least-squares contrast.

In this article, we mainly consider histogram models S_m, that is of the following form. Let $(I_\lambda)_{\lambda \in \Lambda_m}$ be some fixed partition of $\mathcal{X}$. Then, S_m denotes the set of functions $\mathcal{X} \mapsto \mathbb{R}$ which are constant over I_λ for every $\lambda \in \Lambda_m$; S_m is a

vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbf{1}_{I_\lambda})_{\lambda\in\Lambda_m}$. The empirical risk minimizer $\widehat{s}_m$ over an histogram model S_m is often called a *regressogram*.

Explicit computations are easier with regressograms because $(\mathbf{1}_{I_\lambda})_{\lambda\in\Lambda_m}$ is an orthogonal basis of $L^2(\mu)$ for any probability measure μ on $\mathcal{X}$. In particular,

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

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

Note that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i, that is $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda > 0$.

Let us assume that a collection of models $(S_m)_{m\in\mathcal{M}_n}$ is given. Model selection consists in selecting some data-dependent $\widehat{m} \in \mathcal{M}_n$ such that $\ell\left(s, \widehat{s}_{\widehat{m}}\right)$ is as small as possible. General penalization procedures can be described as follows. Let $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$ be some penalty function, possibly data-dependent, and define

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

Since the goal is to minimize the loss $P\gamma\left(\widehat{s}_m\right)$, the *ideal penalty* is

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma\left(\widehat{s}_m\right), \tag{3}$$

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 the histogram framework, note that $\widehat{s}_m$ is not uniquely defined when $\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda = 0$; then, we consider that the model S_m cannot be chosen, which is formally equivalent to add $+\infty\mathbf{1}_{\min_{\lambda\in\Lambda_m} \widehat{p}_\lambda = 0}$ to the penalty $\mathrm{pen}(m)$.

When S_m is the histogram model associated with some partition $\left(I_\lambda\right)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$, the ideal penalty (3) can be computed explicitly:

$$\begin{aligned}
\mathrm{pen}_{\mathrm{id}}(m) &= (P - P_n)\gamma(s_m) + P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) + P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right) \\
&= (P - P_n)\gamma(s_m) + \sum_{\lambda\in\Lambda_m} \left[p_\lambda\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 + \widehat{p}_\lambda\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 \right]
\end{aligned} \tag{4}$$

where $p_\lambda := \mathbb{P}\left(X \in I_\lambda\right)$. The ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ is unknown because it depends on the true distribution P; therefore, resampling is a natural method for estimating $\mathrm{pen}_{\mathrm{id}}(m)$.

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 says that one can mimic the relationship between P and P_n by drawing a n-sample with common distribution P_n, called the "resample"; let P_n^W denote the empirical distribution

of the resample. Then, the conditional distribution of the pair (P_n, P_n^W) given P_n should be close to the distribution of the pair (P, P_n). Hence, 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 with respect to the resampling randomness only. Let us emphasize that $\mathrm{pen}_{\mathrm{id}}(m)$ has the form $F(P, P_n)$.

Later on, this heuristics has been generalized to other resampling schemes, with the exchangeable weighted bootstrap [54, 57]. The empirical distribution of the resample then has the general form

$$P_n^W := \frac{1}{n}\sum_{i=1}^{n} W_i \delta_{(X_i, Y_i)}$$

where $W \in \mathbb{R}^n$ is an exchangeable[3] weight vector independent of the data and such that $\forall i$, $\mathbb{E}_W[W_i] = 1$. In this article, W is also assumed to satisfy $\forall i$, $W_i \geq 0$ a.s. and $\mathbb{E}_W[W_i^2] < \infty$.

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): $W_i = (n/q)\mathbf{1}_{i\in I}$ where I is a 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, Efr, Rad, Poi, Rho and Loo respectively denote the above resampling weight vector distributions with the "classical" value of the parameter.

Remark 1. The above terminology explicitly links the weight vector distributions with some classical resampling schemes. See [54, 40, 66] for more details about classical resampling weight names, as well as other classical examples.

- The name "Efron" comes from the classical choice $M = n$ for which Efron weights actually are the bootstrap weights. When $M < n$, Efron(M) is the M out of n bootstrap, used for instance by Shao [59].
- 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. For instance, global and local Rademacher complexities use this resampling scheme to estimate different upper bounds on $\mathrm{pen}_{\mathrm{id}}(m)$ (see Section 7.2.4).
- Poisson weights are often used as approximations to Efron weights, via the so-called "Poissonization" technique (see [66, Chapter 3.5] and [33]). They

[3] W is said to be *exchangeable* when its distribution is invariant by any permutation of its coordinates.

are known to be efficient for estimating several non-smooth functionals (see [15, Chapter 3] and [52, Section 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 [66, Example 3.6.14], more often called *subsampling weights* (see for instance the book by Politis, Romano and Wolf [56] on subsampling). They are close to the idea of splitting the data into a training set and a validation set (for instance, 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 complement, there is a one-to-one correspondence between subsampling weights and data splitting.

2.3. Resampling Penalization

Applying directly the resampling heuristics of Section 2.2 for estimating the ideal penalty (3), we would get the penalty

$$\mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m^W \right) - P_n^W \gamma \left(\widehat{s}_m^W \right) \right], \tag{5}$$

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

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

Two problems have to be solved before defining properly the Resampling Penalization procedure. Here, we focus on the histogram framework; the general framework will be considered in Section 7.2.

First, (5) is not well-defined because $\widehat{s}_m^W$ is not unique if $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda^W = 0$. Hence, even when $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda > 0$, the problem occurs as soon as $\widehat{W}_\lambda = 0$ for some $\lambda \in \Lambda_m$, which has a positive probability (except when $D_m = 1$) for most of the resampling schemes since $\mathbb{P}_W (\forall i \geq 2, W_i = 0) > 0$. In order to make (5) well-defined, let us rewrite the resampling penalty as the resampling estimate of (4), that is

$$\mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m^W \right) - P_n^W \gamma \left(\widehat{s}_m^W \right) \right] = \widehat{p}_0(m) + \widehat{p}_1(m) + \widehat{p}_2(m)$$

where

$$\widehat{p}_0(m) := \mathbb{E}_W \left[(P_n - P_n^W) \gamma(\widehat{s}_m) \right] = \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}_W \left[1 - W_i \right] \gamma \left(\widehat{s}_m ; (X_i, Y_i) \right) \right) = 0$$

TABLE 1
C_W for several resampling schemes (see Section 3.4.1)

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

because $\mathbb{E}_W[W_i] = 1$ for every i,

$$\widehat{p}_1(m) := \sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W \left(\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right) \right)$$
$$\text{and} \quad \widehat{p}_2(m) := \sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W \left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right] \right).$$

With the convention $\widehat{p}_\lambda^W (\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda)^2 = 0$ when $\widehat{p}_\lambda^W = 0$, $\widehat{p}_2(m)$ is well-defined since $\widehat{\beta}_\lambda^W$ is well-defined when $\widehat{p}_\lambda^W > 0$. It remains to define properly $\widehat{p}_1(m)$. We suggest to replace the expectation over all the resampling weights by an expectation conditional on $\widehat{W}_\lambda > 0$, *separately for each* $m \in \mathcal{M}_n$ *and* $\lambda \in \Lambda_m$, which ensures that we only remove a small proportion of the possible resampling weights. To summarize, (5) is replaced by

$$\sum_{\lambda\in\Lambda_m} \left(\mathbb{E}_W \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \,\middle|\, \widehat{W}_\lambda > 0 \right] + \mathbb{E}_W \left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right] \right). \tag{6}$$

Second, (6) is strongly biased as an estimate of $\mathrm{pen}_{\mathrm{id}}$ when $\mathrm{var}(W_1)$ is small, because P_n^W is then much closer to P_n than P_n is close to P. Assuming the S_m to be histogram models, we will prove in Section 3.4.1 (see Propositions 1 and 2) that the bias can be corrected by multiplying (6) by a constant C_W which only depends on the distribution of W. The values of C_W for the classical weights are reported in Table 1. Remark that $C_W = 1$ in the bootstrap case (Efr), as well as for Rad, Poi and Rho.

We are now in position to define properly the Resampling Penalization (RP) procedure for selecting among histogram models. See Section 7.2 for the definition of RP in the general framework (Procedure 3).

Procedure 1 (Resampling Penalization for histograms)**.**

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

5. Select $\widehat{m} \in \arg\min_{m \in \widehat{\mathcal{M}}_n} \{ P_n \gamma (\widehat{s}_m) + \mathrm{pen}(m) \}$.

Remark 2.
1. At step 1, we remove more models than those for which $\widehat{s}_m$ is not uniquely defined. When $n\widehat{p}_\lambda = 1$ for some $\lambda \in \Lambda_m$, estimating the quality of estimation of $\widehat{\beta}_\lambda$ with only one data-point is hopeless with no assumption on the noise-level σ. The reason why we remove also models for which $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} = 2$ is that the oracle inequalities of Section 3 require it for some of the weights; nevertheless, such models generally have a poor prediction performance, so that step 1 is reasonable.
2. At step 3, C can be larger than C_W because overpenalizing can be fruitful from the non-asymptotic point of view, in particular when the sample size n is small or the noise level σ is large. The simulation study of Section 5 provides experimental evidence for this fact (see also Section 6.3.2).
3. RP (Procedure 1) generalizes several model selection procedures. With a bootstrap resampling scheme (Efr) and $C = 1$, RP is Efron's bootstrap penalization [30], which has also been called EIC in the log-likelihood framework [43]. With an M out of n bootstrap resampling scheme (Efr(M)) and $C = 1$, RP has been proposed and studied by Shao [59] in the context of model identification. Note that $C_W \neq 1$ for Efr(M) weights if $M \neq n$; this crucial point will be discussed in Section 3.4.1. RP with a (non-exchangeable) V-fold subsampling scheme has also been proposed recently in [9].
4. When W are the "leave-one-out" weights, RP is not the classical leave-one-out model selection procedure. Nevertheless, according to [9], when $C = n - 1$, it is identical to Burman's n-fold corrected cross-validation [22], hence close to the uncorrected one.

3. Main results

In this section, we state some non-asymptotic properties of Resampling Penalization (Procedure 1) for model selection. First, Theorem 1 is an oracle inequality with leading constant close to 1. In particular, Theorem 1 implies the asymptotic optimality of RP. Second, Theorem 2 is an adaptivity result for an estimator built upon RP, when the regression function belongs to some Hölderian ball. A remarkable point is that both results remain valid under mild assumptions on the distribution of the noise, which can be non-Gaussian and highly heteroscedastic.

Throughout this section, we assume the existence of non-negative constants $\alpha_{\mathcal{M}}$, $c_{\mathcal{M}}$, c_{rich} such that:

(**P1**) Polynomial size 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 weight vector W is chosen among Efr, Rad, Poi, Rho and Loo (defined in Section 2.2, with the classical value of their parameter).

(**P1**) is a natural restriction since RP plugs an estimator of the ideal penalty into (2). When $\mathrm{Card}(\mathcal{M}_n)$ is larger, say proportional to e^{an} for some $a > 0$,

Birgé and Massart [21] proved that penalties estimating the ideal penalty cannot be asymptotically optimal. (**P2**) is merely technical. (**P3**) can be relaxed, as explained in Section 4.2.

3.1. Oracle inequality

Theorem 1. *Assume that the data $(X_i, Y_i)_{1\leq i\leq n}$ 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**) *Polynomially decreasing bias: there exist* $\beta_1 \geq \beta_2 > 0$ *and* $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ *such that*

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

- ($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{D}(X)$*: there exists* $c_{\mathrm{r},\ell}^X > 0$ *such that*

$$\forall m \in \mathcal{M}_n, \quad D_m \min_{\lambda\in\Lambda_m} p_\lambda \geq c_{\mathrm{r},\ell}^X.$$

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

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

Moreover,

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

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* $\left(\ln(n)\right)^{-1/5}$*;* ε_n *can also be made smaller than* $n^{-\delta}$ *for any* $0 < \delta < \delta_0(\beta_1, \beta_2)$ *at the price of enlarging* K_1*.*

Theorem 1 is proved in Section 8.3. The non-asymptotic oracle inequality (8) implies that Procedure 1 is *a.s. asymptotically optimal* in this framework if $\lim_{n\to\infty}(C/C_W) = 1$. When W are Efr weights, the asymptotic optimality of RP was proved by Shibata [62] for selecting among maximum likelihood estimators, assuming that the distribution P belongs to some parametric family of densities (see also Remark 6 in Section 3.4.1).

Resampling Penalization yields an estimator with an excess loss as small as the one of the oracle without requiring any knowledge about P such as the smoothness of s or the variations of the noise-level σ. Therefore, RP is a *naturally adaptive procedure*. Note that (8) is even stronger than an adaptivity result because of the leading constant close to one, whereas adaptive estimators only achieve the correct estimation rate up to a possibly large absolute constant. Hence, one can expect that an estimator obtained with RP and a well chosen collection of models is almost optimal.

We now comment on the assumptions of Theorem 1:

1. The constant C can differ from C_W. For instance, when a constant $\eta > 1$ exists such that $C \in [C_W; \eta C_W]$, the oracle inequalities (8) and (9) hold with leading constant $2\eta - 1 + \varepsilon_n$ instead of $1 + \varepsilon_n$.
2. (**Ab**) and (**An**) are rather mild and neither A nor $\sigma_{\min}$ need to be known by the statistician. In particular, quite general heteroscedastic noises are allowed; (**Ab**) and (**An**) can even be relaxed as explained in Section 3.3.2.
3. When X has a lower bounded density with respect to Leb, ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied for "almost piecewise regular" histograms, including all those considered in the simulation study of Section 5.
4. The upper bound in (**Ap**) holds with $\beta = 2\alpha k^{-1}$ when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular on $\mathcal{X} \subset \mathbb{R}^k$ and s is α-Hölderian with $\alpha > 0$. The lower bound in (**Ap**) is discussed extensively in Section 3.3.1.

3.2. An adaptive estimator

A natural framework in which Theorem 1 can be applied is when $\mathcal{X}$ is a compact subset of $\mathbb{R}^k$, X has a lower bounded density with respect to the Lebesgue measure and s is α-Hölderian with $\alpha \in (0, 1]$. Indeed, the latter condition ensures that regular histograms can approximate s well. In this subsection, we show that Resampling Penalization can be used to build an estimator adaptive to the smoothness of s in this framework.

We first define the estimator. For the sake of simplicity[4], $\mathcal{X}$ is assumed to be a closed ball of $(\mathbb{R}^k, \|\cdot\|_\infty)$, say $[0,1]^k$.

Procedure 2 (Resampling Penalization with regular histograms)**.** For every $T \in \mathbb{N}\setminus\{0\}$, let $S_{m(T)}$ be the model of regular[5] histograms with T^k bins, that is the histogram model associated with the partition

$$(I_\lambda)_{\lambda\in\Lambda_{m(T)}} := \left(\prod_{i=1}^k \left[\frac{j_i}{T}; \frac{j_i+1}{T}\right)\right)_{0\le j_1,\dots,j_k\le T-1}.$$

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

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

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

1. Choose a resampling scheme $\mathcal{D}(W)$ among Efr, Rad, Poi, Rho and Loo.
2. Take the constant $C = C_W$ as defined in Table 1.
3. For each $m \in \widehat{\mathcal{M}}_n$, compute the resampling penalty $\mathrm{pen}(m)$ defined by (7).

[4] If $\mathcal{X}$ has a smooth boundary, Procedure 2 can be modified so that the proof of Theorem 2 remains valid.

[5] When $\mathcal{X}$ has a general shape, assume that both $\mathrm{Leb}(\mathcal{X})$ and $\mathrm{diam}(\mathcal{X})$ for $\|\cdot\|_\infty$ are finite. Then, a partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ is *regular with* T^k *bins* when $\mathrm{Card}(\Lambda_m) = T^k$ and there exist positive constants c_1, c_2, c_3, c_4 such that for every $\lambda \in \Lambda_m$, $c_1 T^{-k} \le \mathrm{Leb}(I_\lambda) \le c_2 T^{-k}$ and $c_3 T^{-1} \le \mathrm{diam}(I_\lambda) \le c_4 T^{-1}$.

4. Select $\widehat{m} \in \arg\min_{m \in \widehat{\mathcal{M}}_n} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$.
5. Define $\widetilde{s} := \widehat{s}_{\widehat{m}}$.

Theorem 2. *Let $\mathcal{X} = [0,1]^k$. Assume that the data $(X_i, Y_i)_{1 \leq i \leq n}$ 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.*
(**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 exist $\alpha \in (0;1]$ and $R > 0$ such that*

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

Let $\widetilde{s}$ be the estimator defined by Procedure 2 and $\sigma_{\max} := \sup_{\mathcal{X}} |\sigma| \leq 2A$. Then, there exist positive constants K_2 and K_3 such that,

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

If moreover the noise-level is smooth, that is

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

then, assumption (**An**) *can be removed and* (10) *holds with $\sigma_{\max}$ replaced by $\|\sigma\|_{L^2(\mathrm{Leb})} := [(\mathrm{Leb}(\mathcal{X}))^{-1} \int_{\mathcal{X}} \sigma^2(t) dt]^{1/2}$.*

For both results, K_2 may only depend on α and k. The constant K_3 may only depend on k, A, $c_X^{\min}$, R, α (and $\sigma_{\min}$ for (10)*); K_σ and J_σ for the latter result).*

Theorem 2 is proved in Section 8.5. The upper bounds given by Theorem 2 coincide with several classical minimax lower bounds on the estimation of functions in $\mathcal{H}(\alpha, R)$ with $\alpha \in (0, 1]$, up to an absolute constant. In the homoscedastic case, lower bounds have been proved by Stone [63] and generalized by several authors among which Korostelev and Tsybakov [47] and Yang and Barron [69]. Up to a multiplicative factor independent of n, R and σ the best achievable rate is

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

Hence, (10) shows that Procedure 2 achieves the right estimation rate in terms of n, R and σ, without using the knowledge of α, R or σ.

Moreover, (10) still holds in a wide heteroscedastic framework, without using any information on the noise-level $\sigma(\cdot)$. Then, up to a multiplicative constant independent of n and R (but possibly of the order of some power of $\sigma_{\max}/\sigma_{\min}$), the upper bound (10) is the best possible estimation rate.

Minimax lower bounds proved in the heteroscedastic case (see for instance [28, 35] and references therein) show that when $k = \alpha = 1$ and the noise-level is smooth enough, the best achievable estimation rate depends on σ through the multiplicative factor $\|\sigma\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+k}}$. Therefore, the upper bound given by Theorem 2

under assumption $(\mathbf{A}\sigma)$ is tight, even through its dependence on the noise-level. Up to our best knowledge, such an upper bound had never been obtained when $\alpha \in (0,1)$ and $k > 1$, even with estimators using the knowledge of α, σ and R.

Theorem 2 shows that Procedure 2 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 result is quite strong. Although similar properties have already been proved for "*ad hoc*" estimators (see [28, 35] and Section 7.1.3), *Resampling Penalization has not been designed specifically to have such a property*. Therefore, exchangeable resampling penalties are *naturally adaptive* to the smoothness of s and to the heteroscedasticity of the data.

Remark 3.

1. The proof of Theorem 2 shows that $\widehat{s}_{\widehat{m}}$ achieves the minimax rate of estimation on an event of probability larger than $1 - K_3' n^{-2}$. In particular, with probability one,
$$\limsup_{n\to\infty} \left(\ell\left(s, \widetilde{s}\right) R^{\frac{-2k}{2\alpha+k}} n^{\frac{2\alpha}{2\alpha+k}} \left\|\sigma\right\|_{L^2(\mathrm{Leb})}^{\frac{-4\alpha}{2\alpha+k}} \right) \leq K_2(\alpha, k).$$
2. If s is piecewise α-Hölderian with at most J_s jumps (each jump of height bounded by $2A$), then (10) holds with K_3 depending also on J_s.
3. As for Theorem 1, the boundedness of the data and the lower bound on the noise level can be replaced by other assumptions (see Section 3.3.2).

3.3. Discussion on some assumptions

The aim of this subsection is to discuss some of the main assumptions made in Theorems 1 and 2. We first tackle the lower bound in $(\mathbf{Ap})$ which is required in Theorem 1. Then, two alternative assumption sets to Theorems 1 and 2 are provided, allowing the noise level to vanish or the data to be unbounded.

3.3.1. Lower bound in $(\mathbf{Ap})$

The lower bound $\ell\left(s, s_m\right) \geq C_{\mathrm{b}}^{-} D_m^{-\beta_1}$ in $(\mathbf{Ap})$ may seem unintuitive because it means that s is not too well approximated by the models S_m. Assuming that $\inf_{m\in\mathcal{M}_n} \ell\left(s, s_m\right) > 0$ is classical for proving the asymptotic optimality of Mallows' C_p [61, 49, 21].

Let us explain why $(\mathbf{Ap})$ is used for proving Theorem 1. According to Remark 8 in Section 8.2, when the lower bound in $(\mathbf{Ap})$ is no longer assumed, (8) holds with two modifications on its right-hand side: the infimum is restricted to models of dimension larger than $\left(\ln(n)\right)^{\gamma_1}$ and a remainder term $\left(\ln(n)\right)^{\gamma_2} n^{-1}$ is added (where γ_1 and γ_2 are absolute constants). This is essentially the same as (8) unless there exists a model of small dimension with a small bias; the lower bound in $(\mathbf{Ap})$ is sufficient to ensure this does not happen. Note that assumption $(\mathbf{Ap})$ was made in the density estimation framework [64, 23] for the same technical reasons.

As showed in [8], (**Ap**) is at least satisfied with

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

in the following case: $(I_\lambda)_{\lambda\in\Lambda_m}$ is "regular" (as defined in Procedure 2 below), X has a lower-bounded density with respect to the Lebesgue measure Leb on $\mathcal{X} \subset \mathbb{R}^k$ and s is non-constant and α-Hölderian (with respect to $\|\cdot\|_\infty$).

The general formulation of (**Ap**) is crucial to make Theorem 1 valid *whatever the distribution of* X which can be useful in some practical problems. Indeed, when X has a general distribution, a collection $(S_m)_{m\in\mathcal{M}_n}$ satisfying (**P1**), (**P2**), ($\mathbf{Ar}_\ell^{\mathbf{X}}$) and (**Ap**) can always be chosen either thanks to prior knowledge on $\mathcal{D}(X)$ or to unlabeled data. In the latter case, classical density estimation procedures can be applied for estimating $\mathcal{D}(X)$ from unlabeled data (see for instance [26] on density estimation). Assumption (**Ap**) then means that the collection of models has good approximation properties, uniformly over some appropriate function space (depending on $\mathcal{D}(X)$) to which s belongs.

3.3.2. Two alternative assumption sets

Theorems 1 and 2 are corollaries of a more general result, called Lemma 7 in Section 8.2. The assumptions of Theorems 1 and 2, in particular (**Ab**) and (**An**) on the distribution of the noise $\sigma(X)\epsilon$, are only sufficient conditions for the assumptions of Lemma 7 to hold. The following two alternative sufficient conditions are proved to be valid in Section 8.4.

First, one can have $\sigma_{\min} = 0$ in (**An**) if moreover $\mathbb{E}\left[\sigma(X)^2\right] > 0$, $\mathcal{X} \subset \mathbb{R}^k$ is bounded and

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

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

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

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

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

Second, the Y_i can be unbounded (assuming now that $\sigma_{\min} > 0$ in (**An**)) if moreover $\mathcal{X} \subset \mathbb{R}$ is bounded measurable and

($\mathbf{A}_{\mathrm{gauss}}$) The noise is sub-Gaussian: $\exists c_{\mathrm{gauss}} > 0$ 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 regression 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

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

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

Third, it is possible to have simultaneously $\sigma_{\min} = 0$ in ($\mathbf{An}$) and unbounded data, see [8] for details.

The above results mean that Theorem 1 holds for most "reasonably" difficult problems. Actually, Proposition 3 and Remark 7 show that the resampling penalties are much closer to $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ than $\mathrm{pen}_{\mathrm{id}}(m)$ itself, provided that the concentration inequalities for $\mathrm{pen}_{\mathrm{id}}$ are tight (Proposition 10). Therefore, up to differences within ε_n, RP with $C = C_W$ and the "ideal" deterministic penalization procedure $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ perform equally well on a set of probability $1 - K_1 n^{-2}$. For every assumption set such that the proof of Theorem 1 gives an oracle inequality for the penalty $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$, the same proof gives a similar oracle inequality for RP.

3.4. Probabilistic tools

Theorems 1 and 2 rely on several probabilistic tools of independent interest: precise computation of the expectations of resampling penalties (Propositions 1 and 2), concentration inequalities for resampling penalties (Proposition 3) and bounds on expectations of the inverses of several classical random variables (Lemma 4–6). Their originality comes from their non-asymptotic nature: explicit bounds on the deviations or the remainder terms are provided for finite sample sizes.

3.4.1. Expectations of resampling penalties

Using only the exchangeability of the weights, the resampling penalty can be computed explicitly (Lemma 16 in Section 8.8). This can be used to compare the expectations of the resampling penalties and the ideal penalty. First, Proposition 1 is valid for *general exchangeable weights*.

Proposition 1. *Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ and $W \in [0,\infty)^n$ be an exchangeable random vector independent of the data. Define $\mathrm{pen}_{\mathrm{id}}(m)$ by (3) and $\mathrm{pen}(m)$ by (7). Let $\mathbb{E}^{\Lambda_m}[\cdot]$ denote expectations conditionally on $(\mathbf{1}_{X_i\in I_\lambda})_{1\leq i\leq n,\,\lambda\in\Lambda_m}$. Then, if $\min_{\lambda\in\Lambda_m}\widehat{p}_\lambda > 0$,*

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

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

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

$$R_{1,W}(n, \widehat{p}_\lambda) := \mathbb{E}\left[\frac{(W_1 - \widehat{W}_\lambda)^2}{\widehat{W}_\lambda^2} \,\middle|\, X_1 \in I_\lambda, \widehat{W}_\lambda > 0 \right], \tag{13}$$

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

In particular,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) \right] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(2 + \delta_{n,p_\lambda} \right) \sigma_\lambda^2 \tag{15}$$

where $\delta_{n,p}$ only depends on (n,p) and satisfies $|\delta_{n,p}| \leq L_1 (np)^{-1/4}$ for some absolute constant L_1.

Proposition 1 is proved in Section 8.8.

Remark 4.
- In order to make the expectation in (15) well-defined, a convention for $\mathrm{pen}_{\mathrm{id}}(m)$ has to be chosen when $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda = 0$. See Section 8.1 for details.
- Combining Proposition 1 with [6, Lemma 8.4], a similar result holds for non-exchangeable weights (with only a modification of the definitions of $R_{1,W}$ and $R_{2,W}$).

In the general heteroscedastic framework (1), Proposition 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 Section 7.1.1.

If $R_{1,W}(n, \widehat{p}_\lambda) + R_{2,W}(n, \widehat{p}_\lambda)$ does not depend too much on $\widehat{p}_\lambda$ (at least when $n\widehat{p}_\lambda$ is large), Proposition 1 shows that $\mathrm{pen}(m)$ estimates unbiasedly $\mathrm{pen}_{\mathrm{id}}(m)$ as soon as[6]

$$C = C_W \approx \frac{2}{R_{1,W}(n,1) + R_{2,W}(n,1)} = \frac{1}{\mathbb{E}\left[(W_1 - 1)^2 \right]}.$$

In particular, all the examples of resampling weights given in Section 2.2 satisfy that $R_{1,W}(n, \widehat{p}_\lambda) \approx R_{2,W}(n, \widehat{p}_\lambda)$ does not depend on $\widehat{p}_\lambda$ when $n\widehat{p}_\lambda$ is large, which leads to Proposition 2 below (see Table 2 for exact expressions of $R_{2,W}$ and C_W).

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 Table 2.*

[6] The definition of C_W actually used in this paper is slightly different for Efron(M) and Poisson(μ) weights (see Table 2). We arbitrarily choosed the simplest possible expression making C_W asymptotically equivalent to $1/\mathbb{E}[(W_1 - 1)^2]$. The results of the paper also hold when $C_W = 1/\mathbb{E}[(W_1 - 1)^2]$.

TABLE 2
$R_{2,W}(n,\widehat{p}_\lambda)$ and C_W for several resampling schemes. The formulas for $R_{2,W}$ come from Lemma 17

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

Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ and pen(m) *be defined by (7). Then, there exist real numbers $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ depending only on n, $\widehat{p}_\lambda$ and the resampling scheme $\mathcal{D}(W)$ such that*

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] = \frac{C}{C_W n}\sum_{\lambda\in\Lambda_m}\left(2+\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}\right)\sigma^2_\lambda. \tag{16}$$

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 an absolute constant (for Rho($\lfloor n/2\rfloor$) and Loo) or depends respectively on B (Efr), p (Rad) or μ (Poi). More precise bounds for each weight distribution are given by (62)–(66) in Section 8.9.

Proposition 2 is proved in Section 8.9.

Remark 5. Proposition 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 Section 8.9).

Remark 6. Combined with the explicit expressions of C_W for several resampling weights (Table 2), Proposition 2 helps to understand several known results.

- In the maximum likelihood framework, Shibata [62] showed the asymptotical equivalence of two bootstrap penalization methods. The first penalty, denoted by B_1, is Efron's bootstrap penalty [30], which is defined by (5) with Efr weights. The second penalty, denoted by B_2, was proposed by Cavanaugh and Shumway [25]; it transposes

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

  into the maximum likelihood framework. In the least-squares regression framework (with histogram models), the proofs of Propositions 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 Section 8.10 show that this remains true without expectations.

- With Efron (M_n) weights (and a bootstrap selection procedure close to RP, but with $C = 1$), Shao [59] showed that $M_n = n$ leads to an inconsistent model selection procedure for identification. On the contrary, when $M_n \to \infty$ and $M_n \ll n$, Shao's bootstrap selection procedure is model consistent. Proposition 2 shows that these assumptions on M_n can be rewritten $C = 1 \gg C_W = M_n/n$. Therefore, the rationale behind Shao's result may mostly be that identification needs overpenalization within a factor tending to infinity with n.

3.4.2. Concentration inequalities for resampling penalties

From (4), the ideal penalty can be written

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(s_m) + \sum_{\lambda\in\Lambda_m} \frac{p_\lambda + \widehat{p}_\lambda}{(n\widehat{p}_\lambda)^2} \left(\sum_{X_i \in I_\lambda} (Y_i - \beta_\lambda) \right)^2 .$$

Hence, $\mathrm{pen}_{\mathrm{id}}(m)$ is a U-statistics of order 2 conditionally on $(\mathbf{1}_{X_i\in I_\lambda})_{(i,\lambda\in\Lambda_m)}$, which is sufficient to prove that resampling yields a consistent estimate of $\mathrm{pen}_{\mathrm{id}}(m)$ (Arcones and Giné [5] considered the bootstrap case; Hušková and Janssen [42] extended it to the exchangeable weighted bootstrap).

In the non-asymptotic framework, that is when the models S_m can depend on n, the following concentration inequality is needed.

Proposition 3. *Let $\gamma > 0$, $A_n \geq 2$ and W be an exchangeable weight vector. Let S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$ and $\mathrm{pen}(m)$ be defined by (7). Assume that two positive constants a_ℓ and ξ_ℓ exist such that for every $q \geq 2$,*

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

Let $\Omega_m(A_n)$ denote the event $\{\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq A_n\}$. Then, there exist constants $K_4, K_5 > 0$ and an event of probability at least $1 - K_4 n^{-\gamma}$ on which

$$\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right]\right| \mathbf{1}_{\Omega_m(A_n)} \leq C K_5 \times \sup_{np\geq A_n} \left\{ R_{1,W}(n,p) + R_{2,W}(n,p) \right\} \frac{(\ln(n))^{\xi_\ell+1}}{\sqrt{A_n D_m}} \mathbb{E}\left[p_2(m)\right]$$

where $R_{1,W}$ and $R_{2,W}$ are defined by (13) and (14). The constant K_4 is absolute and K_5 may only depend on a_ℓ, ξ_ℓ and γ.

If moreover W satisfies the assumptions of the second part of Proposition 2 and C_W is defined as in Table 2, then a constant $K_W > 0$ exists such that

$$\left|\mathrm{pen}(m) - \mathbb{E}^{\Lambda_m}[\mathrm{pen}(m)]\right| \mathbf{1}_{\Omega_m(A_n)} \leq \frac{C K_5 K_W (\ln(n))^{\xi_\ell+1}}{C_W \sqrt{A_n D_m}} \mathbb{E}\left[p_2(m)\right] . \tag{17}$$

For the Rad(p) weights, K_W is smaller than $(1-p)^{-1}$ multiplied by an absolute constant. For the other weights, K_W is an absolute constant.

Proposition 3 is proved in Section 8.10.1. Note that the moment condition holds under the assumptions of Theorem 1 as well as the alternative assumptions of Section 3.3.2. It is here stated in its most general form.

Remark 7. Since the $A_n^{-1/2}$ factor should tend to infinity with n for most reasonable models, Proposition 3 gives better bounds for resampling penalties than what could be obtained for ideal penalties with Proposition 10 in the same framework.

Although we do not know how tight are the bounds of Proposition 3, such a phenomenon is classical with bootstrap and can be understood from the asymptotic point of view through Edgeworth expansions [39]. In a non-asymptotic Gaussian framework, [10, Section 2.3] shows the same property for resampling estimators, which concentrate at the rate N^{-1} instead of $N^{-1/2}$ (N being the amount of data). Since A_n plays the role of N, the gain $A_n^{-1/2}$ can reasonably be conjectured to be unimprovable without some more assumptions.

Let us emphasize that if resampling penalties estimate $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ instead of $\mathrm{pen}_{\mathrm{id}}(m)$, RP with $C = C_W$ cannot take into account the fact that $\mathrm{pen}_{\mathrm{id}}(m)$ may be far from its expectation.

3.4.3. Expectations of inverses

For any non-negative random variable Z, we define

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

This quantity appears in the explicit formulas for $R_{1,W}$ when W is among the examples of resampling weights of Section 2.2 (see Lemma 17). Therefore, in order to prove Proposition 2, non-asymptotic bounds on e_Z^+ are needed when Z has a binomial, hypergeometric or Poisson distribution.

Former results concerning e_Z^+ can be found in papers by Lew [48] (for general Z), by Jones and Zhigljavsky [44] (for the Poisson case) and by Žnidarič [70] (for the binomial and Poisson case), but they are either asymptotic or not precise enough. Lemmas 4–6 solve this issue.

In the rest of the paper, for any $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of a and b and $a \vee b$ denotes 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 $np \geq 1$,*

$$\kappa_2 \wedge \left(1 + \kappa_1 (np)^{-1/4}\right) \geq e_{\mathcal{B}(n,p)}^+ \geq 1 - e^{-np} \tag{18}$$

$$\text{and} \qquad 2 + 3\times 10^{-4} \geq e_{\mathcal{B}\left(n,\frac{1}{2}\right)}^+ \geq \mathbf{1}_{n\geq 3}. \tag{19}$$

The first bounds (18) were first stated in [9, Lemma 3] where they are proved. The second ones (19) are proved in Section 8.11.1. Lemma 4 implies in particular that $e_{\mathcal{B}(n,p)}^+ \to 1$ when $np \to \infty$, which can be derived from [70].

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

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

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:*

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

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

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

$$\text{Then,} \qquad 1 + \kappa_3(\epsilon)\frac{n}{q}\sqrt{\frac{\ln(r)}{r}} \geq e^+_{\mathcal{H}(n,r,q)}. \tag{20}$$

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

$$14.3 \geq \sup_{r\geq 1}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2}\rfloor)}\right\} \qquad \text{and} \qquad 3 \geq \sup_{r\geq 26}\left\{e^+_{\mathcal{H}(n,r,\lfloor \frac{n}{2}\rfloor)}\right\}. \tag{21}$$

4. *"Loo" case:*

$$1 + \frac{\mathbf{1}_{r\geq 2}}{n(r-1)} \geq e^+_{\mathcal{H}(n,r,n-1)} = 1 + \frac{1}{n}\left(\frac{(n-1)r}{n(r-1)}\mathbf{1}_{r\geq 2} - 1\right) \geq 1 - \frac{\mathbf{1}_{r=1}}{n}. \tag{22}$$

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

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

Lemma 5 is proved in Section 8.11.2. It implies in particular that

$$e^+_{\mathcal{H}(n_k,r_k,q_k)} \xrightarrow[k\to\infty]{} 1 \quad \text{if } n_k \geq r_k \xrightarrow[k\to\infty]{} +\infty$$

and $\sup_k \left\{n_k q_k^{-1}\right\} < +\infty$.

Poisson case

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

$$\left(2 - 2e^{-2\mu}\right) \wedge \left(1 + \frac{2(1+e^{-3})}{(\mu-2)\mathbf{1}_{\mu>2}}\right) \geq e^+_{\mathcal{P}(\mu)} \geq 1 - \mathbf{1}_{\mu<1.61}e^{-\mu}.$$

Lemma 6 is proved in Section 8.11.3. It implies in particular that $e^+_{\mathcal{P}(\mu)} \to 1$ when $\mu \to \infty$, which can be derived from [44, 70].

4. Comparison of the weights

We investigate in this section how the loss of the final estimator may depend on the distribution of the exchangeable weight vector W. First, we consider in Section 4.1 the most classical ones, that is Efr, Rad, Poi, Rho and Loo. Then, we discuss in Section 4.2 whether Theorem 1 can be extended to general exchangeable weights.

4.1. Comparison of the classical weights

According to Theorem 1, any resampling scheme among Efr, Rad, Poi, Rho and Loo leads to an asymptotically optimal procedure. Even from the non-asymptotic point of view, it is not quite clear to distinguish between these weights with the results of Section 3. Indeed, the resampling penalties are equal in expectation at first order (Proposition 2), and their deviations are negligible in front of their expectations (Proposition 3).

Therefore, 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 subsection second-order terms in the expectations of the penalties (that is, differences between second-order terms in (15) and (16)), for a fixed sample size. Asymptotic considerations can be found in the book by Barbe and Bertail [15, Chapter 2] where Edgeworth expansions are used to compare the accuracy of estimation with many exchangeable weights. The asymptotic results mentioned in Section 3.4.3 may also be useful.

Propositions 1 and 2 show that $\mathrm{pen}_{\mathrm{id}}(m)$ and $\mathrm{pen}(m)$ have the same expectation, up to the small terms δ_{n,p_λ} and $\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$. More precisely,

$$\mathbb{E}\left[\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\overline{\delta}^{(\mathrm{penW})}_{n,p_\lambda} - \delta_{n,p_\lambda}\right)(\sigma_\lambda)^2$$
$$\text{with} \qquad \overline{\delta}^{(\mathrm{penW})}_{n,p_\lambda} := \mathbb{E}\left[\delta^{(\mathrm{penW})}_{n,\widehat{p}_\lambda} \,\middle|\, \widehat{p}_\lambda > 0\right].$$

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

It follows that Loo weights are the most accurate ones, even when np is small. On the contrary, Rho $(n/2)$ and Rad tend to overestimate $\mathrm{pen}_{\mathrm{id}}$ since $\overline{\delta}^{(\mathrm{penW})}_{n,p} > \delta_{n,p}$ (except when np is small, where the inequality is reversed). It also seems that the bias of Rho (q) is a decreasing function of q, as illustrated by Figures 3–4. Finally, Efr and Poi are strongly underestimating the ideal penalty, mostly because of the $1-(n\widehat{p}_\lambda)^{-1}$ term in $R_{1,W}(n,\widehat{p}_\lambda)$ and $R_{2,W}(n,\widehat{p}_\lambda)$.

This can be summed up as follows:

$$\mathrm{penRad} \approx \mathrm{penRho} > \mathrm{penLoo} \approx \mathrm{pen}_{\mathrm{id}} >> \mathrm{penEfr} \approx \mathrm{penPoi}, \tag{23}$$

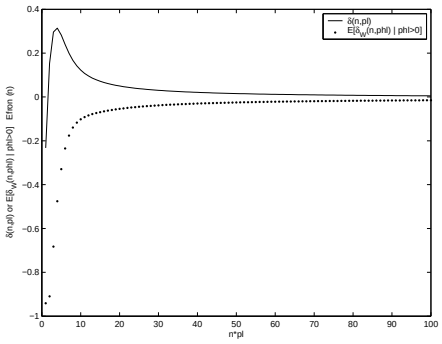

FIG 1. $\delta_{n,p} > 0 > \overline{\delta}_{n,p}^{(\mathrm{penEfr(n)})}$.

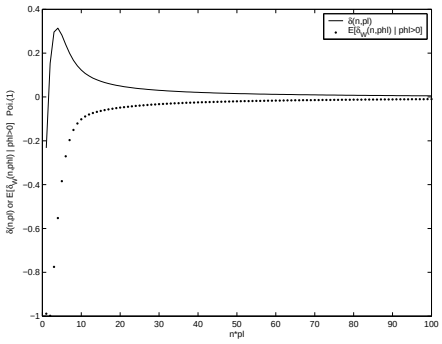

FIG 2. $\delta_{n,p} > 0 > \overline{\delta}_{n,p}^{(\mathrm{penPoi(1)})}$.

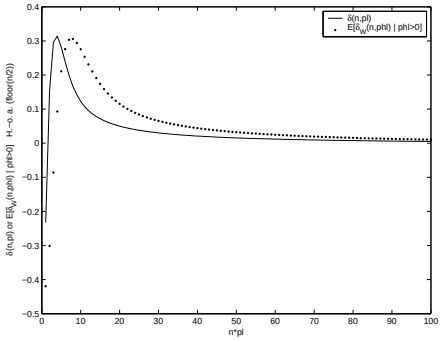

FIG 3. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRho}(n/2))}$ *for* $np \geq 6$.

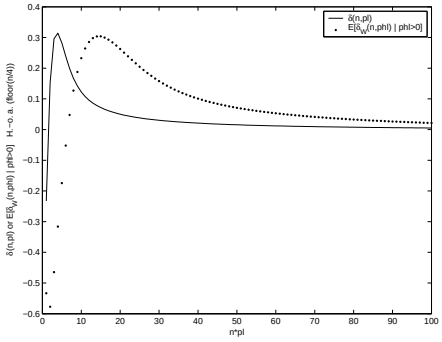

FIG 4. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRho}(n/4))}$ *for* $np \geq 9$.

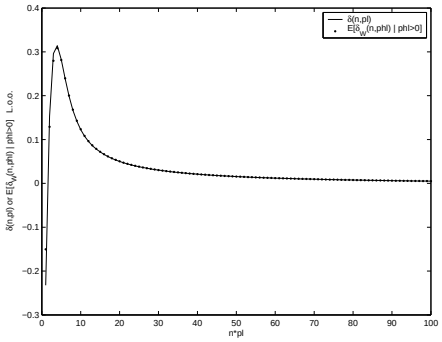

FIG 5. $\delta_{n,p} \approx \overline{\delta}_{n,p}^{(\mathrm{penLoo})}$.

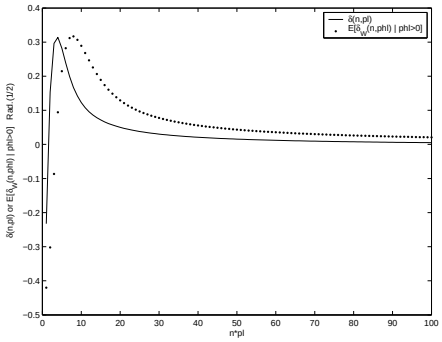

FIG 6. $\delta_{n,p} > \overline{\delta}_{n,p}^{(\mathrm{penRad}(1/2))}$ *for* $np \geq 6$.

where ">>" means a comparatively large gap, but still negligible at first order. Hence, we can expect that the Loo penalty is the most efficient, closely followed by Rad and by Rho. However, from the non-asymptotic point of view, it turns out that smaller prediction loss is obtained by overpenalizing slightly

(and sometimes strongly, see the simulations of Section 5 and the discussion of Section 6.3.2). Then, the ordering of (23) may also be the one of the prediction performances of RP, the best performances being obtained with Rad and Rho. This is confirmed by the simulation study of Section 5.

Another interesting point is that $\overline{\delta}_{n,p}^{(\mathrm{penRho})} \propto \delta_{n,p}$ when np is large enough. Then, provided that histograms with too small bins are removed from the collection, penLoo and penRho are almost equivalent, up to the choice of the factor C. If a wise tuning of C is possible, it remains to choose between Loo and Rho according to computational issues (see the discussion of Section 6.2).

4.2. Other exchangeable weights

The oracle inequality of Theorem 1 is only stated for the five "classical" exchangeable weights of Section 2.2. Nevertheless, replacing the threshold 3 by some $T \geq 2$ at step 1 of Procedure 1, the proof of Theorem 1 can be extended to any resampling weight vector W satisfying:

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 Proposition 2),
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 15).

In particular, the first two conditions hold for all the exchangeable weights considered in Proposition 2. The third one is satisfied for most of them as soon as T is large enough (see Lemma 15 in Section 8.6).

5. Simulation study

As an illustration of the results of Section 3, the prediction performances of Procedure 1 (with several resampling schemes), Mallows' C_p and V-fold cross-validation are compared on some simulated data.

5.1. Experimental setup

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

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

where $(X_i)_{1\leq i\leq n}$ are independent with uniform distribution over $\mathcal{X} = [0;1]$ and $(\epsilon_i)_{1\leq i\leq n}$ are independent standard Gaussian variables independent of $(X_i)_{1\leq i\leq n}$. The experiments differ from the regression function s (smooth for S, see Figure 7; smooth with jumps for HS, see Figure 8), the noise type (homoscedastic for S1 and HSd1, heteroscedastic for S2 and HSd2) and the sample size n (see Table 3). Instances of data sets are plotted on Figures 9–12.

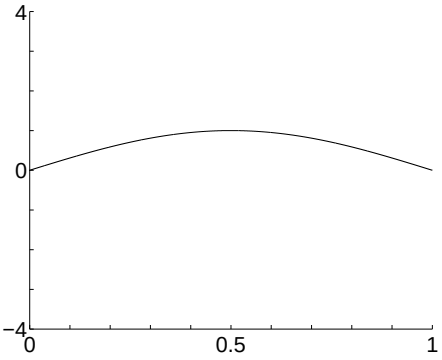

FIG 7. $s(x) = \sin(\pi x)$.

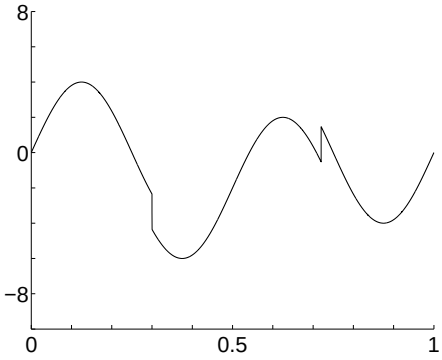

FIG 8. $s(x) = \text{HeaviSine}(x)$ *(see [27]).*

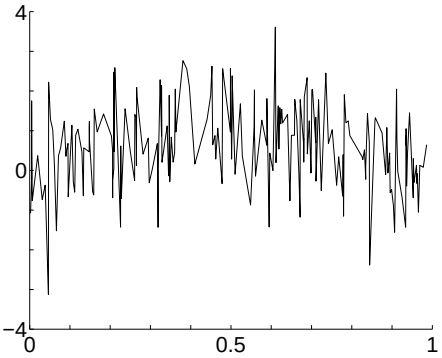

FIG 9. *S1:* $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$.

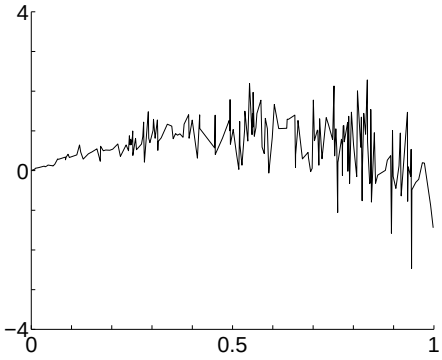

FIG 10. *S2:* $s(x) = \sin(\pi x)$, $\sigma(x) = x$, $n = 200$.

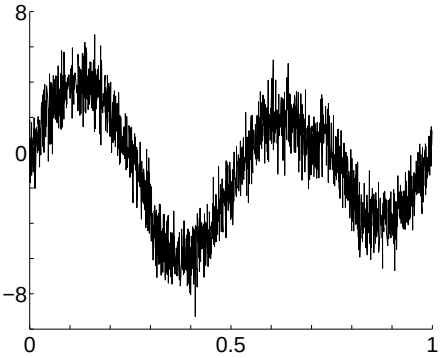

FIG 11. *HSd1: HeaviSine,* $\sigma \equiv 1$, $n = 2048$.

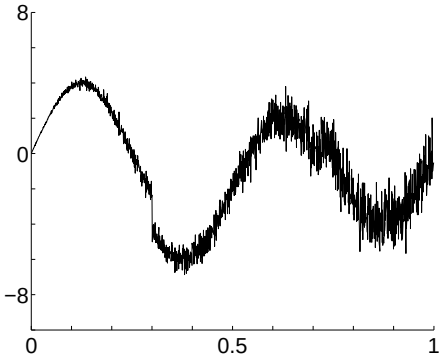

FIG 12. *HSd2: HeaviSine,* $\sigma(x) = x$, $n = 2048$.

The collections of histogram models also differ according to the experiments. Define

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

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

For every $m \in (\mathbb{N}\backslash\{0\})\cup(\mathbb{N}\backslash\{0\})^2$, let S_m be the histogram model associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Then, for each experiment, the collection of models is $(S_m)_{\mathcal{M}_n}$ with different index sets $\mathcal{M}_n$:

S1 regular histograms with $1 \le D \le n\,(\ln(n))^{-1}$ pieces, that is

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

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

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

HSd1 dyadic regular histograms with 2^k pieces, $0 \le k \le \ln_2(n) - 1$, that is

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

HSd2 dyadic histograms regular on $[0;1/2]$ (resp. on $[1/2;1]$) with bin sizes 2^{-k_1} (resp. 2^{-k_2}), $0 \le k_1, k_2 \le \ln_2(n) - 2$ (dyadic version of S2). The model of constant functions is added to $\mathcal{M}_n$, that is

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

Note that the collections of models used in experiments S2 and HSd2 can adapt to s and $\sigma(\cdot)$. Therefore, the oracle model is generally quite efficient so that the model selection problem is more challenging.

The following procedures[7] are compared:

Mal Mallows' C_p penalty: $\mathrm{pen}(m) = 2\widehat{\sigma}^2 D_m n^{-1}$ where $\widehat{\sigma}^2$ is the classical variance estimator defined as

$$\widehat{\sigma}^2 = \frac{d^2\left(Y_{1\dots n}, S_{\lfloor n/2\rfloor}\right)}{n - \lfloor n/2\rfloor}, \tag{24}$$

where $Y_{1\dots n} = (Y_i)_{1\le i\le n} \in \mathbb{R}^n$, $S_{\lfloor n/2\rfloor}$ is any model of dimension $\lfloor n/2\rfloor$ (only assumed to have a bias negligible in front of σ^2) and d is the

[7] The code used for computing resampling penalties is available on the author's webpage at http://www.di.ens.fr/~arlot/index.htm.

Euclidean distance on $\mathbb{R}^n$. The non-asymptotic validity of this model selection procedure in homoscedastic regression has been assessed by Baraud [13].

$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$ Expectation of the ideal penalty: $\mathrm{pen}(m) = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$, which witnesses what is a good performance in each experiment.

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

LOO Leave-one-out (that is VFCV with $V = n$).

penEfr Efron (n) penalty (7) with $C = C_W = 1$.

penRad Rademacher $(1/2)$ penalty (7) with $C = C_W = 1$.

penRho Random hold-out $(n/2)$ penalty (7) with $C = C_W = 1$.

penLoo Leave-one-out penalty (7) with $C = C_W = n - 1$.

For each of these, the same penalties multiplied by $5/4$ are also considered (and they are denoted 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 Section 6.3.2).

In each experiment, for each simulated data set, first the models with 2 data points or less in one piece of their associated partition are removed. Then, the least-squares estimators $\widehat{s}_m$ are computed for each $m \in \widehat{\mathcal{M}}_n$. Finally, $\widehat{m} \in \widehat{\mathcal{M}}_n$ is selected using each procedure and its true excess loss $\ell\left(s, \widehat{s}_{\widehat{m}}\right)$ is computed as well as the excess loss of the oracle $\inf_{m \in \mathcal{M}_n} \ell\left(s, \widehat{s}_m\right)$. $N = 1000$ data sets are simulated, thanks to which the model selection performance of each procedure is estimated 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). Since C_{or} and $C_{\mathrm{path-or}}$ approximatively give the same rankings between procedures, Table 3 only reports C_{or}; the values of $C_{\mathrm{path-or}}$ are reported in [8].

5.2. Results and comments

First, the above experiments show the interest of both Resampling Penalization (RP) and VFCV in several difficult frameworks, with relatively small sample sizes. Although RP and VFCV cannot compete with simple procedures such as Mallows' C_p from the computational point of view, they are much more efficient when the noise is heteroscedastic (S2 and HSd2). In these difficult frameworks, the prediction performances of RP and VFCV are comparable to those of $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$. Note that in HSd2, penRad and penRho give smaller losses than any penalty proportional to the dimension of the models (see Section 7.1.2). Moreover, penRad and penRho perform slighlty worse than Mallows' C_p for the easiest problems (S1 and HSd1), which can be interpretated as the unavoidable price for robustness.

TABLE 3
Accuracy indices C_{or} for each procedure in four experiments, $\pm$ a rough estimate of uncertainty of the value reported (that is the empirical standard deviation divided by $\sqrt{N}$; $N = 1000$). In each column, the more accurate procedures (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
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	1.919 ± 0.03	2.296 ± 0.05	1.028 ± 0.004	$\mathbf{1.102 \pm 0.004}$
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]+$	$\mathbf{1.792 \pm 0.03}$	$\mathbf{2.028 \pm 0.04}$	$\mathbf{1.003 \pm 0.003}$	$\mathbf{1.089 \pm 0.004}$
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	$\mathbf{1.097 \pm 0.004}$

Second, in the four experiments, the best procedures always are the overpenalizing ones: many of them even beat the perfectly unbiased $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$, showing the crucial need to overpenalize. This phenomenon disappears for small σ and large n [8, Experiments S0.1 and S1000], hence it is certainly due to the small signal-to-noise ratio. We would like to insist on the importance of the overpenalization 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.

Let us now compare RP and VFCV. According to the four experiments of Table 3, RP with Rad or Rho resampling schemes clearly outperforms VFCV for any V, even without overpenalizing. The only exception to this is HSd1 where 2-fold cross-validation yields a particularly good model selection performance.

This can be interpretated thanks to the non-asymptotic study of the performance of V-fold cross-validation provided in [9]. In short, VFCV overpenalizes within a factor $1 + 1/(2(V-1))$, while the V-fold criterion has a variance decreasing with V.

Then, when overpenalization is necessary (for instance in S1, S2 or HSd1), small values of V can outperform the leave-one-out ($V = n$). Nevertheless, RP with the right overpenalization level C/C_W leads to a smaller prediction loss than VFCV, because RP provides a less variable model selection criterion than VFCV. The reason why penRad and penRho also perform slightly better without overpenalization is that they naturally overpenalize when $C = C_W = 1$ (see Section 4).

Let us now consider the model selection 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 performs slightly worse (but not always significantly) and Efr much worse. Looking carefully at the values of the penalties, it appears that Rad and Rho slightly overpenalize, 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 [8]). Note that this comparison can also be derived from theoretical computations (see Section 4). Since overpenalization is benefic in the four experiments of Table 3, this explains why penRad and penRho slightly outperform penLoo. In the case of Efron's boostrap penalty, underpenalizing implies overfitting which explains the comparatively bad performances reported in Table 3.

We conclude this section with remarks concerning some particular points of the simulation study.

- On the same data sets, Mallows' C_p and its overpenalized version Mal+ were performed 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 yielded worse model selection performance for all experiments but S2, in which $C_{\mathrm{or}}(\mathrm{Mal}) = 2.657 \pm 0.06$ and $C_{\mathrm{or}}(\mathrm{Mal+}) = 2.437 \pm 0.05$. Therefore, overpenalization is crucial in experiment S2, more than the shape[8] of the penalty itself. Moreover, the overpenalization level being fixed, resampling penalties remain significantly better than Mallows' C_p. Hence, the performances of Mallows' C_p in Table 3 are not only due to a bad estimation of the mean noise-level (see also Section 7.1).
- Eight additional experiments are reported in [8], showing similar results with various n, σ and s (although the assumptions of Theorem 1 are not always satisfied).
- Resampling penalties with a V-fold subsampling scheme have also been studied in [9, Section 4] on the same simulated data: exchangeable resampling schemes always give better model selection performance than non-exchangeable ones (significantly when V is small), except for Efr and Poi which tend to underestimate the ideal penalty.

[8] The shape of a penalty is defined as the way $\mathrm{pen}(m)$ depends on m up to a linear transformation.

6. Practical implementation

This section tackles three main issues for using Procedure 1 in practice: how to compute the resampling penalty (7)? how to choose the weights W? how to choose the constant C?

6.1. Computational cost

An exact computation of resampling penalties with exchangeable weights (without using formula (50) for histograms) would be either impossible or computationally expensive. We suggest two possible ways to fix this problem.

First, one can use a classical Monte-Carlo approximation, that is draw a small number B of independent weight vectors instead of considering each element of the support of $\mathcal{D}(W)$. Practical Monte-Carlo methods for the boostrap are proposed for instance by Hall [39, Appendix II]. Moreover, a non-asymptotic estimation of the accuracy of Monte-Carlo approximation can be obtained via McDiarmid's inequality (see Arlot, Blanchard and Roquain [10, Proposition 2.7] for a precise result using the same idea in another framework). This would provide a practical way of quantifying what is lost by making a Monte-Carlo 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 $\mathcal{D}(W)$ is much smaller than n. A case-example is *V-fold subsampling*: 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 of the data, we define

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

The resulting resampling penalties —called V-fold penalties— have been introduced and studied in [9]. They are computationally similar to VFCV while being more flexible, since the overpenalization factor is decoupled from the choice of V; hence, like resampling penalties, V-fold penalties select an estimator with smaller prediction loss than the one selected by VFCV.

Both Monte-Carlo approximation of RP and V-fold penalization have been tested on the simulated data of Section 5. The detailed results are given in [8].

6.2. Choice of the weights

The influence of the weights has been investigated from the theoretical point of view in Section 4 with focus on second-order terms in expectation. However, deviations of $\mathrm{pen}(m)$ around its expectation are likely to depend on the weight vector W since the upper bound in (17) may not be tight. The simulation study of Section 5 allows to take into account both phenomena in the comparison between the resampling weights.

In terms of model selection efficiency, Table 3 shows that the best weights (for accuracy of prediction and for the variability[9] of this accuracy) are Rho and Rad, whereas Loo perform slightly worse. On the contrary, from both accuracy and variability points of view, Efron's bootstrap weights perform worse than Rho, Rad and Loo, mainly because they lead to underpenalization.

Note however that this comparison strongly depends on the precise definition[10] of C_W, which makes all penalties unbiased at first order but possibly under or over-penalizing at second order. Then, different prediction performances may be observed on data which do not require overpenalization. Nevertheless, the computations of Section 4 show that Efron's bootstrap weights have a real drawback which cannot be fixed only by changing C_W.

When computing the penalties exactly, Loo weights are the only computationally tractable ones, while being almost as accurate as Rho and Rad. Hence, we suggest their use, enlarging the constant C when needed (see Section 6.3.2 on overpenalization).

However, computing n empirical risk minimizers (or the outputs of computationally more expensive algorithms) for each model is not always possible. In such a case, one should avoid using the Leave-one-out with a Monte-Carlo approximation, which would give a large importance to a small number of data points. Rho or Rad weights are much safer in this situation. Alternatively, one may consider the use of V-fold penalties [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, the deviations of resampling 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, the optimal choice of V for V-fold cross-validation is known to be different between least-squares regression and binary classification [9, Section 2.3]. Such differences are expected to arise for choosing between exchangeable resampling weights.

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 the approach of this paper is to propose the use of other exchangeable resampling schemes instead of the boostrap so that the bias of resampling penalties no longer has to be corrected.

[9] The variability of the accuracy 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 a procedure performing always equally well can be preferred to a procedure with better mean efficiency but poor performances on a small probability event.

[10] However, it is quite unclear how to change C_W in order to optimize each penalty in the general case. This is why C_W has been chosen as "simple" as possible in Table 2.

6.3. Choice of the constant C

6.3.1. Optimal constant for bias

From the asymptotic point of view, the optimal $C = C^\star$ for prediction is generally the one for which pen estimates the ideal penalty $\mathrm{pen}_{\mathrm{id}}$ unbiasedly (at least for collections of models of polynomial size). This is how C_W is defined in the histogram framework and Theorem 1 implies that $C = C_W$ is asymptotically optimal for prediction. Hence[11], $C^\star$ is asymptotically equivalent to C_W.

As showed by Arlot and Massart [11], $C^\star$ can also be estimated directly from data for general penalties, in particular for RP. Hence, the knowledge of C_W is not necessary, which can be useful in the general prediction framework (see Section 7.2).

6.3.2. Overpenalization

A careful look at the proof of Theorem 1 shows that a similar oracle inequality holds for any $C > 4C_W/5$, the leading constant remaining close to one when $C \sim C_W$ *asymptotically*. In other words, when the sample size n is small, the optimal constant $C^\star$ may not be exactly equal to C_W. The simulations of Section 5 also support this fact: *Overpenalization*, that is, 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, when σ is large or when s is non-smooth.

This problem would appear even if the "optimal" constant $C^\star$ such that pen is non-asymptotically unbiased was known. On Figure 13, the estimated model selection performance of the penalty $C_{\mathrm{ov}} \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ is plotted as a function of C_{ov}, for experiment S2 of Section 5. It appears that the optimal overpenalization constant $C^\star_{\mathrm{ov}} \in (1.5; 2.35)$ for this particular problem. More generally, the drawback of using $C = C^\star$ 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 deviations, the constant C must be slightly enlarged. A major issue remains: How to estimate $C^\star_{\mathrm{ov}}$ from data only, since it strongly depends on n, on σ, on the smoothness of s and on the number of models in $\mathcal{M}_n$?

One can think of choosing C_{ov} by V-fold cross-validation, but this would lead to a computationally intractable procedure. An alternative idea is to use resampling for building a simultaneous confidence region on $\left(\mathrm{pen}_{\mathrm{id}}(m)\right)_{m \in \mathcal{M}_n}$ instead of estimating $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ only (see [10] on confidence regions built with general exchangeable resampling schemes). Then, the uncertainty on the estimation of $\mathrm{pen}_{\mathrm{id}}(m)$ can be taken into account for choosing a model, similarly to model selection procedures built upon relative bounds [12, 24]. Finally, the choice of the overpenalization factor would be replaced by the choice of a confidence level which should be made by the practicioner. See also [6, Section 11.3.3] for a discussion on a data-driven choice of the overpenalization factor.

[11] See the proof of Theorem 1 in [9] to prove that asymptotic optimality requires $C^\star / C_W \xrightarrow[n \to \infty]{} 1$ as soon as there are enough models close to the oracle.

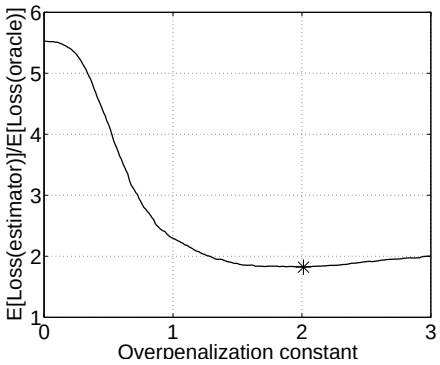

FIG 13. *The non-asymptotic need for overpenalization: the prediction performance* C_{or} *(defined in Section 5.1) of the model selection procedure* (2) *with* $\mathrm{pen}(m) = C_{\mathrm{ov}}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ *is represented as a function of* C_{ov}. *Data and models are the ones of experiment S2:* $n = 200$, $\sigma(x) = x$, $s(x) = \sin(\pi x)$. *See Section 5 for details.*

7. Discussion

7.1. Comparison with other procedures

In this article, the Resampling Penalization (RP) family of model selection procedures is defined and showed to satisfy some optimality properties under mild assumptions on the data (Theorems 1 and 2). In particular, RP is robust to the heteroscedasticity of the noise according to both theoretical and experimental results. The price for robustness is that the computational cost of RP is generally larger than simple procedures like Mallows' C_p, even with the suggestions of Section 6.1. The purpose of this subsection is to identify the "easy" problems, for which the computational cost of RP can be reduced by using C_p-like penalties without enlarging the prediction loss too much.

7.1.1. Mallows' C_p

Mallows' C_p penalty is equal to $2\sigma^2 D_m n^{-1}$ for a model S_m of dimension D_m, when the noise-level σ is constant. Non-asymptotic results about C_p-like penalties can be found in [16, 13, 14, 21]. They imply that Mallows' C_p is asymptotically optimal in the homoscedastic framework, when the size of $\mathcal{M}_n$ is polynomial in n.

When the mean noise-level is unknown, it must be estimated. A classical estimator of $\mathbb{E}\left[\sigma^2(X)\right]$ is defined by (24). Baraud [13, 14] showed that the resulting data-driven model selection procedure satisfies a non-asymptotic oracle inequality with leading constant close to one.

Assume for the sake of simplicity that n is even and let $S_{n/2}$ be a model such that each piece of the associated partition contains exactly two data points.

Reordering the (X_i, Y_i) according to X_i,

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

so that

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{Mallows}}(m)\right] \approx \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(D_m \widehat{p}_\lambda\right) \left(\sigma_\lambda^r\right)^2 + \frac{2D_m}{n^2} \sum_{i=1}^{n/2} \left(s(X_{2i}) - s(X_{2i-1})\right)^2 \tag{25}$$

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

This should be compared with the result of Proposition 1:

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^d\right)^2\right) \tag{26}$$

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

Although 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 variations of s through $(s(X_{2i}) - s(X_{2i-1}))^2$ or $(\sigma_\lambda^d)^2$), they differ on at least two points.

First, when s is smooth and $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \}$ is large, the "bias" term in (25) is negligible in front of the one of (26), which means that Mallows' C_p underpenalizes when the "bias" component of $\mathrm{pen}_{\mathrm{id}}$ is large. Second, the "variance" component of $\mathrm{pen}_{\mathrm{id}}$, which is the main one in general, is distorted in Mallows' C_p: the part of the penalty corresponding to I_λ is multiplied by $D_m\widehat{p}_\lambda$ which is not close to 1 when the partition $(I_\lambda)_{\lambda \in \Lambda_m}$ is not regular with respect to $\mathcal{D}(X)$. This happens for instance in experiments S2 and HSd2 of Section 5. Therefore, there are at least three possibly "hard" problem classes:

- heteroscedastic noise, with irregular histograms and X uniform (for instance S2, HSd2 in Section 5, or Svar2 in [8]),
- heteroscedastic noise, with regular histograms and X highly non-uniform on $\mathcal{X}$,
- regression function s with jumps (such as HeaviSine[12]) or large non-smooth areas (such as Doppler in [8]).

In either of these cases, one should avoid the use of C_p-like penalties, and we suggest resampling penalties as an efficient alternative. As explained in Section 7.1.2 below, the first class of problems can make any penalty proportional to the dimension D_m suboptimal.

[12] 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.

7.1.2. Linear penalties

Mallows' C_p is simple because it is a linear function of the dimension D_m of S_m:

$$\mathrm{pen}(m) = \widehat{K} D_m \tag{27}$$

and $\widehat{K}$ is the only constant to determine. Depending on what is known on the mean variance level, the constant $\widehat{K}_{\mathrm{Mallows}}$ can be defined as

$$2\mathbb{E}\left[\sigma(X)^2\right] n^{-1} \quad \text{or} \quad 2\widehat{\sigma}^2 n^{-1}.$$

Refined versions of Mallows' C_p have also been proposed [16, 14, 21] but they are still linear or very close to linearity.

However, according to (11), the ideal penalty is not linear in general, even in expectation. Moreover, there exist some frameworks in which any penalty of the form (27) is suboptimal when data are heteroscedastic [7], that is, it cannot satisfy any oracle inequality with leading constant smaller than some absolute constant $\kappa > 1$. In other words, the *optimal linear penalization procedure* $\mathrm{pen}_{\mathrm{opt,lin}}(m) := \widehat{K}^\star D_m$ is suboptimal, where

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

As showed by Theorem 1, RP does not suffer from this drawback.

On the one hand, the optimal linear penalization procedure has a better model selection performance than RP for S1, S2 and HSd1, which is not surprising for the "easy" problems where Mallows' C_p is almost optimal (S1, HSd1). It is less intuitive for S2 where data are heteroscedastic. 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.

On the other hand, in experiment HSd2, the optimal linear penalization has a model selection performance $C_{\mathrm{or}} = 1.18 \pm 0.01$, which is worse than the one of RP ($C_{\mathrm{or}} \leq 1.11$). Thus, the most difficult problem of Section 5 (with a large collection of models, heteroscedastic data and bias) gives an example where linear penalties are definitely not adapted, in addition to the ones of [7].

7.1.3. Ad hoc *procedures*

One of the main advances with Theorems 1 and 2 is that RP is proved to work in the heteroscedastic framework contrary to Mallows' C_p. Nevertheless, in a framework such as the one of experiment S2, Mallows' C_p can be adapted to heteroscedasticity by splitting $\mathcal{X}$ into several parts where σ is almost constant, and performing the histogram selection procedure with Mallows' C_p separately on each part of $\mathcal{X}$.

More generally, Efromovich and Pinsker [28] and Galtchouk and Pergamenschikov [35] (among several others) defined estimators of s that are minimax

adaptive in the heteroscedastic framework, the latter by model selection. In the Gaussian regression framework, Gendre [37] proposed a model selection method for estimating simultaneously the regression function and the noise level.

All these procedures may perform slightly better than RP in terms of prediction loss. They are called "*ad hoc*" because they have been specially designed for the heteroscedastic framework (and a particular collection of estimators for [35, 37]). On the contrary, RP is a general-purpose device: It was neither built to be adaptive to heteroscedasticity nor to take advantage of a specific model, and RP has exactly the same definition in the general prediction framework (see Section 7.2).

When no information is available on the data or when no model selection procedure is known for using such information, we suggest the use of RP. Moreover, available information can be partial or wrong. Then, using an *ad hoc* procedure would be disastrous whereas a general device like RP would still work. In short, choose RP if you have no useful information or if you do not trust them.

7.1.4. 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 can have some tricky behavior, in particular for choosing V [68, 9]. This can also be showed in the simulation experiments of Section 5 (see Table 3): In HSd1, $V = 2$ performs better than $V \in \{5, 10, 20\}$, a phenomenon explained in [9] by analyzing how the bias of the V-fold criterion depends on V.

V-fold penalization, that is, RP with a V-fold subsampling scheme, was proposed in [9] where it was showed to improve significantly the model selection performance of VFCV. In this paper and in [8], RP with several exchangeable resampling schemes —generalizing the $V = n$ case— is proved to perform at least as well as V-fold penalization and often better.

Several penalization procedures use the bootstrap for estimating the ideal penalty [30, 25, 62]. As noticed in Remark 6, the penalization procedures studied by Shibata [62] are quite close to RP, although they are restricted to bootstrap weights, which are the worst ones in the framework of the present paper (see Sections 4.1 and 6.2). Moreover, they do not consider useful to multiply the penalty by a factor C possibly different from one, contrary to what is suggested in RP. The factor C is crucial because it disconnects the choice of the weights from the overpenalization problem.

In order to select the correct model asymptotically with probability one, Shao [59] proposed to use RP with the M_n out of n bootstrap and provided a sufficient condition on M_n to achieve model consistency. Thanks to the unified approach for all the exchangeable resampling weights provided in this paper, Shao's condition can be rewritten as $C = 1 \gg C_W$ (see Remark 6), which corresponds to the known fact that model consistency requires overpenalization within a factor tending to infinity with n [1]. Hence, we conjecture that RP with a constant $C \gg C_W$ is model consistent for most exchangeable W, which

may improve Shao's penalties since $\mathrm{Efr}(M_n)$ weights are probably not the best weights in terms of accuracy (see Section 4) and variability[13].

7.2. Resampling Penalization in the general prediction framework

As mentioned in Section 2.1, Resampling Penalization is a general-purpose method which is definitely not restricted to the histogram selection problem. The purpose of this subsection is to define properly RP in the general prediction framework and to discuss briefly what differences can be expected compared to the histogram selection framework.

7.2.1. Framework

Suppose we observe some data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ independent with common distribution P. The goal is to predict Y given X where $(X, Y) \sim P$ is independent of the data. The quality of a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$ is measured by the prediction loss $P\gamma(t) := \mathbb{E}_{(X,Y)}\left[\gamma(t, (X, Y))\right]$ where $(X, Y) \sim P$ and γ is a given contrast function. Typically, $\gamma(t, (x, y))$ measures the discrepancy between $t(x)$ and y. The excess loss is defined as $\ell(s, t) := P\gamma(t) - \inf_{t:\mathcal{X}\mapsto\mathcal{Y}} P\gamma(t)$, even if $s = \arg\min_t \{P\gamma(t)\}$ is not well-defined. Classical examples are least-squares regression where $\mathcal{Y} = \mathbb{R}$ and $\gamma(t, (x, y)) = (t(x) - y)^2$ and binary supervised classification where $\mathcal{Y} = \{0, 1\}$ and $\gamma(t, (x, y)) = \mathbf{1}_{t(x)\neq y}$ is the 0-1 contrast.

A general prediction algorithm $\widehat{s}$ is then defined as a function associating a predictor to any data sample. In order to simplify the presentation, algorithms are assumed to depend only on the empirical distribution $P_n = n^{-1}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ as an input[14]. For instance, the empirical risk minimizer over a set S_m of predictors is defined as $\widehat{s}_m(P_n) := \arg\min_{t\in S_m} P_n\gamma(t)$, provided the minimum in S_m exists and is unique.

Let us assume that a collection of algorithms $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ is given. The goal is to select some data-dependent $\widehat{m} \in \mathcal{M}_n$ minimizing the prediction loss $P\gamma(\widehat{s}_m(P_n))$. The penalization method consists in selecting

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

where $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}$ is a penalty function, possibly data-dependent. Since the goal is to minimize the prediction loss, the *ideal penalty* is

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma\left(\widehat{s}_m(P_n)\right) = F_m(P, P_n)$$

which cannot be used because it depends on the unknown distribution P. When $\mathcal{M}_n$ is not too large (for instance, when $\mathrm{Card}(\mathcal{M}_n) \leq Cn^{\alpha}$ for some positive constants C, α), a natural strategy is to define $\mathrm{pen}(m)$ as an estimator of $\mathrm{pen}_{\mathrm{id}}(m)$ with a bias as small as possible.

[13] Taking into account all the data for computing the resampling penalty with $\mathrm{Efr}(M_n)$ weights is computationally costly when n/M_n is large.

[14] Otherwise, we can consider algorithms whose input is any weighted sample.

7.2.2. Definition of Resampling Penalization

As detailed in Section 2.2, the resampling heuristics can be used for estimating $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[F_m(P, P_n)\right]$, leading to the following procedure.

Procedure 3 (Resampling Penalization)**.**

1. Replace $\mathcal{M}_n$ by
$$\widehat{\mathcal{M}}_n = \left\{m \in \mathcal{M}_n \text{ s.t. } \widehat{s}_m(P_n) \text{ is well-defined } \right\}.$$
2. Choose a resampling scheme, that is the distribution $\mathcal{D}(W)$ of a weight vector W.
3. Choose a constant $C \geq C_W$.
4. Compute the following resampling penalty for each $m \in \mathcal{M}_n$:
$$\mathrm{pen}(m) = C\mathbb{E}_W\left[P_n\gamma\left(\widehat{s}_m\left(P_n^W\right)\right) - P_n^W\gamma\left(\widehat{s}_m\left(P_n^W\right)\right)\right], \qquad (28)$$
where $P_n^W := n^{-1}\sum_{i=1}^n W_i\delta_{(X_i,Y_i)}$.
5. Select $\widehat{m} \in \arg\min_{m\in\widehat{\mathcal{M}}_n}\left\{P_n\gamma\left(\widehat{s}_m\left(P_n\right)\right) + \mathrm{pen}(m)\right\}$.

As for the histogram selection problem, two possible problems have to be solved. First, $\widehat{s}_m(P_n^W)$ may not be well-defined for a.e. W even if $m \in \widehat{\mathcal{M}}_n$. A way to define properly the resampling penalty for every $m \in \widehat{\mathcal{M}}_n$ such that $\widehat{s}_m(P_n^W)$ is well-defined for every $W \in (0,+\infty)^n$ is suggested in [6, Section 8.1]. This assumption is satisfied by regressograms (hence, in the framework of the rest of the paper) for which the suggest of [6, Section 8.1] yields exactly the penalty (7).

Second, the constant C_W such that (28) estimates unbiasedly $\mathrm{pen}_{\mathrm{id}}(m)$ when $C = C_W$ is required in Procedure 3. For the histogram selection problem, the explicit expression of C_W follows from Propositions 1 and 2. In general, the asymptotic theory of exchangeable bootstrap empirical processes [66, Theorem 3.6.13] suggests that $C_W = 1$ if $\mathrm{var}(W_1) \ll 1$, which holds for the classical weights Efr, Rad, Poi and Rho; nevertheless, asymptotic control on the bias is not sufficient when the collection of algorithms is allowed to depend on the sample size n, as in the histogram selection problem. Therefore, further theoretical investigations would be useful to compute the theoretical value of C_W to be used in Procedure 3. From the practical point of view, the data-driven calibration algorithm of [11] can be used for choosing the constant C in front of the resampling penalty.

7.2.3. Model selection properties of Resampling Penalization

The theoretical validity of Procedure 3 is only proved for histogram model selection in this paper, because precise non-asymptotic controls of the ideal penalty and its resampling counterpart are needed. To our knowledge, the only known result about model selection with Resampling Penalization was that RP with the classical bootstrap weights (Efr) is asymptotically optimal for selecting among

maximum likelihood estimators in [62], assuming that the distribution P belongs to some parametric family of densities.

RP can be conjectured to enjoy adaptivity properties for a wide class of model selection problems for two main reasons. First, RP relies on the resampling idea which is known to be robust in a wide variety of frameworks; Theorems 1 and 2 have confirmed the robustness of RP to heteroscedasticity, whereas RP has not been designed specifically for least-squares regression with heteroscedastic data. Second, several of the key concentration inequalities used to prove Theorems 1 and 2 have been extended in [11, Propositions 8 and 10] to a general framework including bounded regression and binary classification.

As mentioned at the end of Section 7.2.1, Procedure 3 should be restricted to choosing among a number of algorithms at most polynomial in n. Indeed, when $\mathrm{Card}(\mathcal{M}_n)$ is larger, estimating unbiasedly $\mathrm{pen}_{\mathrm{id}}$ can yield strong overfitting [21]. Therefore, RP must be modified for large collections $\mathcal{M}_n$. We suggest to group algorithms according to some modelling complexity index C_m, such as the dimension of S_m if $\widehat{s}_m$ is the empirical risk minimizer over some vector space S_m; then, for every $C \in \mathcal{C}_n = \{ C_m \text{ s.t. } m \in \mathcal{M}_n \}$, define $\widehat{s}_C := \widehat{s}_{\widehat{m}(C)}$ where $\widehat{m}(C) \in \arg\min_{C_m = C} P_n \gamma \left(\widehat{s}_m(P_n) \right)$; finally, apply Procedure 3 to the collection $\left(\widehat{s}_C \right)_{C \in \mathcal{C}_n}$, assuming that $\mathrm{Card}\left(\mathcal{C}_n \right)$ is at most polynomial in n.

7.2.4. Related penalties for classification

In the classification framework, RP should be compared to several classical resampling-based penalization methods. First, RP with Efr weights was first introduced by Efron [30] and called bootstrap penalization; its main drawback is its bias (as for the histogram selection problem), which can be corrected in several ways, using for instance the double bootstrap penalization or the .632 bootstrap [30]. Nevertheless, the computational cost of the double bootstrap is heavy and the general validity of the .632 bootstrap is questionable because of its poor theoretical grounds.

Second, the global Rademacher complexities were introduced in order to obtain theoretically validated model selection procedures in classification [45, 17]. They are resampling estimates of

$$\mathrm{pen}_{\mathrm{id,g}}(m) := \sup_{t \in S_m} \left\{ (P - P_n)\gamma(t) \right\} \geq (P - P_n)\gamma\left(\widehat{s}_m(P_n) \right) = \mathrm{pen}_{\mathrm{id}}(m),$$

with Rad weights; more recently, Fromont [34] generalized global Rademacher complexities to a wide family of exchangeable resampling weights and obtained non-asymptotic oracle inequalities. Nevertheless, global complexities (that is, estimates of $\mathrm{pen}_{\mathrm{id,g}}$) are too large compared to $\mathrm{pen}_{\mathrm{id}}$ so that they cannot achieve fast rates of estimation when the margin condition [53] holds.

Therefore, localized penalties taking into account the closeness between $\widehat{s}_m(P_n)$ and s have been introduced, in particular local Rademacher complexities [50, 18, 19, 46]; these papers proved sufficiently tight oracle inequalities to ensure that the final prediction loss can achieve fast rates. Nevertheless, local Rademacher

complexities are computationally heavy and depend on several constants which are difficult to calibrate.

RP aims at combining the advantages of these three approaches in classification. From the practical point of view, RP is computationally tractable (see Section 6.1) and reasonably easy to calibrate (see Section 6.3). Compared to global Rademacher complexities, resampling penalties estimate directly $\mathrm{pen}_{\mathrm{id}}$, so that RP should be able to achieve fast rates of estimation when the margin condition holds. Finally, contrary to the bootstrap penalty, RP can be used with several resampling weights including i.i.d. Rademacher weights (Rad), so that the bias of RP may not have to be corrected.

7.3. Conclusion

This article intends to help the practicioner to answer the following question: When should 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, in particular 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 procedures such as V-fold cross-validation or V-fold penalties [9].
- possibly outperformed by Mallows' C_p (for easy problems) or *ad hoc* procedures (in some particular frameworks, when some information on the data is available).

8. Proofs

8.1. Notation

Before starting the proofs, we introduce some additional notation and conventions:

- The letter L denotes "some positive absolute 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 $L_{c_1,\ldots,c_k}$; if $(\mathbf{A})$ denotes a set of assumptions, $L_{(\mathbf{A})}$ denotes any positive constant depending on the parameters appearing in $(\mathbf{A})$.

- By convention, $\infty \mathbf{1}_E$ and $\mathbf{1}_E/0$ are both equal to zero when the event E does not hold.
- For any $x \in \mathbb{R}$, $x_+ := x \vee 0 = \max(x, 0)$ and $x_- := (-x) \vee 0$.
- For any non-negative random variable Z, $e^0_{\mathcal{D}(Z)} := \mathbb{E}\left[Z\right] \mathbb{E}\left[Z^{-1} \mathbf{1}_{Z>0}\right]$.
- For any model $m \in \mathcal{M}_n$,

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

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

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

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

$$\widetilde{p_1}(m) := \widetilde{p_1}^{(0)}(m) + \sum_{\lambda \in \Lambda_m} p_\lambda \left(\sigma_\lambda\right)^2 \mathbf{1}_{\widehat{p_\lambda} = 0}$$
$$\text{with} \quad \widetilde{p_1}^{(0)}(m) := \sum_{\lambda \in \Lambda_m} \frac{p_\lambda \mathbf{1}_{\widehat{p_\lambda} > 0}}{\left(n \widehat{p}_\lambda\right)^2} S_{\lambda,1}^2$$
$$\text{and} \quad \widetilde{p}_2(m) := p_2(m) + \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(\sigma_\lambda\right)^2 \mathbf{1}_{n\widehat{p_\lambda} = 0}$$

  Note that $p_1(m) = \widetilde{p_1}^{(0)}(m) = \widetilde{p_1}(m)$ and $p_2(m) = \widetilde{p}_2(m)$ are well-defined when $\widehat{s}_m$ is uniquely defined, and other models are always removed from $\mathcal{M}_n$. The above convention is only important when writing expectations, so it is merely technical. In the following, $\widetilde{p_1}$ (resp. $\widetilde{p}_2$) will often be written simply p_1 (resp. p_2).

Using the above notations, $p_1(m)$ and $p_2(m)$ can now be computed explicitly for histogram models. For any $m \in \mathcal{M}_n$ such that $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda > 0$,

$$p_1(m) = \sum_{\lambda \in \Lambda_m} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2 = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(\frac{p_\lambda}{\widehat{p}_\lambda} \frac{S_{\lambda,1}^2}{n\widehat{p}_\lambda}\right) \tag{29}$$

$$p_2(m) = \sum_{\lambda \in \Lambda_m} \widehat{p}_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2 = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(\mathbf{1}_{n\widehat{p_\lambda} > 0} \frac{S_{\lambda,1}^2}{n\widehat{p}_\lambda}\right) \tag{30}$$

since $\widehat{\beta}_\lambda - \beta_\lambda = S_{\lambda,1}/(n\widehat{p}_\lambda)$.

8.2. General framework

The main results (Theorems 1 and 2) actually are corollaries of a more general oracle inequality (Lemma 7). First, two different assumption sets under which Lemma 7 holds are stated in this subsection. The first one **(Bg)** deals with bounded data, the second one **(Ug)** with unbounded data.

8.2.1. Bounded assumption set *(Bg)*

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

(**P1**) Polynomial size 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 weight vector W is 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 exist $a_\ell, \xi_\ell, D_0 \geq 0$ such that for every $q \geq 2$, for every $m \in \mathcal{M}_n$ such that $D_m \geq D_0$,

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

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

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

($\mathbf{A_Q}$) There exist $c_Q^- > 0$ and $D_0 \geq 0$ such that for every $m \in \mathcal{M}_n$ with $D_m \geq D_0$,

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

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

8.2.2. Unbounded assumption set *(Ug)*

(**Ab**) is replaced in (**Bg**) by

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

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

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

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

(**A**δ) Global moment assumption for the bias: there exists $c^g_{\Delta,m} > 0$ such that for every $m \in \mathcal{M}_n$ with $D_m \geq D_0$,

$$\|s - s_m\|_\infty \leq c^g_{\Delta,m} \|s(X) - s_m(X)\|_2 .$$

8.2.3. General result

Lemma 7. *Let $n \in \mathbb{N}\setminus\{0\}$, $\gamma_0 > 0$ and $\widehat{m}$ be defined by Procedure 1. Assume that either* (**Bg**) *or* (**Ug**) *holds with constants independent of 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*

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left[2\eta - 1 + \left(\ln(n)\right)^{-1/5}\right] \inf_{m\in\mathcal{M}_n} \left\{\ell\left(s, \widehat{s}_m\right)\right\} \tag{31}$$

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

Lemma 7 is proved in Section 8.7.

Remark 8. If the lower bound in (**Ap**) is removed from the assumption set, then there exist constants $\gamma_1, \gamma_2 > 0$ (depending only on ξ_ℓ, resp. on ξ_ℓ and $\xi_{g\epsilon}$) and an event of probability at least $1 - K_1 n^{-\gamma_0}$ on which

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left[2\eta - 1 + \left(\ln(n)\right)^{-1/5}\right] \inf_{\substack{m\in\mathcal{M}_n \\ D_m \geq \left(\ln(n)\right)^{\gamma_1}}} \left\{\ell\left(s, \widehat{s}_m\right)\right\} + \frac{\left(\ln(n)\right)^{\gamma_2}}{n}. \tag{32}$$

This assertion is proved in Section 8.7.3.

Remark 9. In the infimum in (31), $\widehat{s}_m$ may not be well-defined for some $m \in \mathcal{M}_n$. By convention $\ell\left(s, \widehat{s}_m\right)$ is defined as $+\infty$ for these m.

From the proof of Lemma 7, there exists a constant $c > 0$ (depending on $\alpha_{\mathcal{M}}$, γ_0 and $c^X_{\mathrm{r},\ell}$) such that every model of dimension smaller than $cn\left(\ln(n)\right)^{-1}$ belongs to $\widehat{\mathcal{M}}_n$ on the event where (31) 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 the infimum can be restricted to models of dimension smaller than $cn\left(\ln(n)\right)^{-1}$ with any of these conventions for $\ell\left(s, \widehat{s}_m\right)$.

The main results of the paper (Theorems 1 and 2) can now be proved, which is done in Sections 8.3–8.5.

First, the assumptions of Theorem 1 imply (**Bg**). Second, the alternative assumption sets stated in Section 3.3.2 imply (**Bg**). Third, the assumptions of Theorem 2 imply (**Bg**) except the lower bound in (**Ap**), so that Remark 8 can be used instead of Lemma 7.

8.3. Proof of Theorem 1

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

$$P_m^{\ell}(q) = \frac{\sqrt{\sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4}}{\sqrt{D_m} Q_m^{(p)}} \leq \frac{\|Y - s_m(X)\|_\infty^2}{Q_m^{(p)}} \leq \frac{4A^2}{Q_m^{(p)}} \tag{33}$$
$$Q_m^{(p)} := \frac{1}{D_m} \sum_{\lambda\in\Lambda_m} \left[\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^d\right)^2 \right] \geq \sigma_{\min}^2.$$

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

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

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

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

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

8.4. Proof of Theorem 1: alternative assumptions

In this section, the statements of Section 3.3.2 are proved.

8.4.1. No uniform lower bound on the noise-level

When $\sigma_{\min} = 0$ in $(\mathbf{An})$, Lemma 8 below proves that $(\mathbf{A_Q})$ also holds with $D_0 = L_{(\mathbf{Bg})}$. Therefore, using (33), $(\mathbf{A_{m},}_{\ell})$ holds with the same D_0. □

Lemma 8. *Let $\mathcal{X} \subset \mathbb{R}^k$, $m \in \mathcal{M}_n$, and assume that positive constants $c_{\mathrm{r,u}}^d, \alpha_d, c_{\mathrm{r,u}}, K_\sigma, J_\sigma$ exist such that*

$(\mathbf{Ar_u^d})$ $\max_{\lambda\in\Lambda_m}\{\mathrm{diam}(I_\lambda)\} \leq c_{\mathrm{r,u}}^d D_m^{-\alpha_d} \mathrm{diam}(X)$,
$(\mathbf{Ar_u})$ $\max_{\lambda\in\Lambda_m}\{\mathrm{Leb}(I_\lambda)\} \leq c_{\mathrm{r,u}} D_m^{-1}$ *and*
$(\mathbf{A}\sigma)$ *σ is piecewise K_σ-Lipschitz with at most J_σ jumps.*

Then,

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

Lemma 8 is proved in the technical appendix [8].

Remark 10. Since $\|\sigma(X)\|_2 > 0$ and σ is piecewise Lipschitz, $\|\sigma\|_{L^2(\mathrm{Leb})} > 0$. Thus, the lower bound on $Q_m^{(p)}$ is positive when D_m is large enough.

8.4.2. Unbounded data

We still use Lemma 7, but the proof is a little longer and requires the following Lemma 9 which is proved in the technical appendix [8].

Lemma 9. *Assume that $\mathcal{X} \subset \mathbb{R}$ is bounded and the following:*

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

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

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

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

*Then, (**A**δ) holds true, that is, for every model S_m of dimension $D_m \geq D_0$,*

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

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

Pathwise oracle inequality We prove that (8) holds with probability $1 - K_1 n^{-\gamma_0}$ for a general γ_0, since it will be required for proving a classical oracle inequality below. First, (**A**$_{\mathbf{m},\ell}$), (**A**$_{\mathbf{Q}}$) and (**A**$_{\mathbf{g},\epsilon}$) hold since for every $m \in \mathcal{M}_n$

$$P_m^\ell(q) = \frac{\sqrt{\sum_{\lambda \in \Lambda_m} m_{q,\lambda}^4}}{\sqrt{D_m} Q_m^{(p)}} \leq \frac{\left(2A + c_{\mathrm{gauss}} \sqrt{q} \sigma_{\max}\right)^2}{Q_m^{(p)}} \leq \frac{q L_{c_{\mathrm{gauss}}, \sigma_{\max}, A}}{Q_m^{(p)}}$$

$$Q_m^{(p)} \geq \sigma_{\min}^2$$

$$P^{g\epsilon}(q) \leq \sigma_{\max} c_{\mathrm{gauss}} \sqrt{q}.$$

Second, Lemma 9 (with (**Al**), (**Ar**$_{\ell,\mathbf{u}}$) and (**Ad**$_\ell$)) shows that (**A**δ) holds with $c^g_{\Delta,m} = L_{(\mathbf{Ug})}$ and $D_0 = L_{(\mathbf{Ug})}$.

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, it suffices to upper bound

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

For every $m \in \widehat{\mathcal{M}}_n$, a bound on $\mathbb{E}^{\Lambda_m}[(p_1(m))^2]$ is required. Starting from (29),

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

since

$$\mathbb{E}^{\Lambda_m}\left[\frac{S_{\lambda,1}^4}{(n\widehat{p}_\lambda)^2}\right] = \mathbb{E}^{\Lambda_m}\left[\frac{\left(\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\right)^4}{(n\widehat{p}_\lambda)^2}\right] = \frac{m_{4,\lambda}^4}{n\widehat{p}_\lambda} + \frac{6(n\widehat{p}_\lambda-1)m_{2,\lambda}^4}{n\widehat{p}_\lambda}$$

$$\text{and}\quad D_m\sum_{\lambda\in\Lambda_m} m_{q,\lambda}^4 \leq \left(a_\ell q^{\xi_\ell}\right)^2\left(\sigma_{\max}^2+(2A)^2\right)^2.$$

Hence, using that $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$,

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

which proves (9). □

8.5. *Proof of Theorem 2*

In this proof, (**H**) denotes the set of assumptions made in Theorem 2. (**H**) implies all the assumptions of Theorem 1 except maybe the lower bound in (**Ap**); indeed, ($\mathbf{Ad}_\ell$) and the fact that all the models are "regular" imply ($\mathbf{Ar}_\ell^{\mathbf{X}}$). Therefore, we can start from (32) in Remark 8 below Lemma 7 which does not require the lower bound in (**Ap**) to hold. The constants γ_i are absolute because the data are bounded.

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

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

If $n \geq L_{(\mathbf{H}),c}$, T_0^k is larger than $(\ln(n))^{\gamma_1}$ and smaller than $cn\,(\ln(n))^{-1}$. Hence, from the proof of Lemma 7, $m(T_0) \in \widehat{\mathcal{M}}_n$ and $m(T_0)$ has a finite excess loss on the large probability event of Lemma 7. Moreover,

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

when $n \geq L_{(\mathbf{H})}$. Since $\ell\left(s,s_{m(T_0)}\right) \leq R^2 T_0^{-2\alpha}$ and

$$\mathbb{E}\left[\widetilde{p_1}^{(0)}(m(T_0))\right] \leq \left(\sup_{np\geq 0} e^0_{\mathcal{B}(n,p)}\right)\frac{1}{n}\sum_{\lambda\in\Lambda_{m(T_0)}}\left(\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^d\right)^2\right)$$

$$\leq \frac{2R^2 T_0^{1-2\alpha}}{n} + \frac{2\sigma_{\max}^2 D_{m(T_0)}}{n}$$

(the bound $e^0_{\mathcal{B}(n,p)} \leq 2$ coming from [38, Lemma 4.1]), an event of probability at least $1 - K_1' n^{-2}$ exists on which

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

where K_2 may only depend on k and α. Note that the constant K_1 has been replaced by $K_1' \geq K_1$ so that the probability bound $1 - K_1' n^{-2}$ is nonpositive when n is too small. Enlarging K_1' once more, the term $\left(\ln(n)\right)^{\gamma_2} n^{-1}$ can be dropped off by adding 1 to the constant K_2. Then, taking expectations as in the proof of Theorem 1, (10) holds.

When $(\mathbf{A}\sigma)$ holds, $\sigma_{\max}$ can be replaced by $\|\sigma\|_{L^2(\mathrm{Leb})}$ in the definition of $m(T_0)$. Then, for every $\lambda \in \Lambda_{m(T_0)}$ such that σ does not jump on I_λ,

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

for every $\theta > 0$ (since $\mathrm{Leb}(\mathcal{X}) = 1$). If σ jumps on I_λ (and there exist at most J_σ such λ), $\max_{I_\lambda} \sigma^2 \leq \sigma_{\max}^2$. Hence, taking $\theta = T_0^{-1}$,

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

and the end of the proof does not change. In this second case, $(\mathbf{An})$ can also be removed because all the assumptions stated in the first part of Section 3.3.2 are satisfied. □

8.6. Additional probabilistic tools

Several probabilistic results are needed in addition to the ones of Section 3.4 for proving Lemma 7. First, Proposition 10 below deals with concentration properties of p_1 and p_2. Remark that concentration inequalities for p_2 can be obtained in a general framework [11, Proposition 10]. On the contrary, we do not know any other non-asymptotic bound on the two-sided deviations of p_1.

Proposition 10. *Let $\gamma > 0$ and S_m be the model of histograms associated with some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$. Assume that $\min_{\lambda \in \Lambda_m}\{np_\lambda\} \geq B_n$ and that*

positive constants a_ℓ, ξ_ℓ exist such that $(\mathbf{A_m},\ell)$ $\forall q \geq 2$, $P_m^\ell(q) \leq a_\ell q^{\xi_\ell}$. *Then, if* $B_n \geq 1$, *an event of probability at least* $1 - Ln^{-\gamma}$ *exists on which*

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{a_\ell,\xi_\ell,\gamma}\left[\frac{(\ln(n))^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right]$$
$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{a_\ell,\xi_\ell,\gamma}\left[\frac{(\ln(n))^{\xi_\ell+2}}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \quad (34)$$
$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{a_\ell,\xi_\ell,\gamma}\frac{(\ln(n))^{\xi_\ell+1}}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right].$$

Moreover, if $B_n > 0$, *an event of probability at least* $1 - Ln^{-\gamma}$ *exists on which*

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + \frac{(\gamma+1)\ln(n)}{B_n}} - L_{a_\ell,\xi_\ell,\gamma}\left[\frac{(\ln(n))^{\xi_\ell+2}}{\sqrt{D_m}} + e^{-LB_n}\right]\right)\mathbb{E}\left[\widetilde{p_2}(m)\right]. \quad (35)$$

Proposition 10 is proved in [8]. Second, Lemmas 11 and 12 below provide concentration inequalities for $\overline{\delta}(m)$, when the data are either bounded or unbounded.

Lemma 11. *Assume that* $\|Y\|_\infty \leq A < \infty$. *Recall that for every* $m \in \mathcal{M}_n$, $\overline{\delta}(m) = (P_n - P)(\gamma(s_m) - \gamma(s))$. *Then for every* $x \geq 0$, *an event of probability at least* $1 - 2e^{-x}$ *exists on which*

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta\ell(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right)\frac{A^2 x}{n}. \quad (36)$$

In particular,

$$\left|\overline{\delta}(m)\right| \leq \frac{\ell(s, s_m)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q_m^{(p)}}\frac{\mathbb{E}^{\Lambda_m}[p_2(m)]}{\sqrt{D_m}}x. \quad (37)$$

Proof of Lemma 11. (36) essentially relies on Bernstein's inequality and is proved in details in [11, Proposition 8]. Then, (37) follows from (36) with $\eta = D_m^{-1/2}$ and the definition of $Q_m^{(p)}$. □

Lemma 12. *Assume that positive constants* $a_{g\epsilon}$, $\xi_{g\epsilon}$, $\sigma_{\max}$ *and* $c^g_{\Delta,m}$ *exist such that*

$(\mathbf{A_{g,\epsilon}})$ $\forall q \geq 2$, $P^{g\epsilon}(q) \leq a_{g\epsilon}q^{\xi_{g\epsilon}}$,
$(\mathbf{A}\sigma_{\max})$ $\|\sigma(X)\|_\infty \leq \sigma_{\max}$,
$(\mathbf{A}\delta)$ $\|s - s_m\|_\infty \leq c^g_{\Delta,m}\|s(X) - s_m(X)\|_2$.

Then, for every $x \geq 0$, *an event of probability at least* $1 - e^{-x}$ *exists on which*

$$\left|\overline{\delta}(m)\right| \leq \frac{L_{a_{g\epsilon},\xi_{g\epsilon},c^g_{\Delta,m}}x^{\xi_{g\epsilon}+1/2}}{\sqrt{D_m}}\left[\ell(s, s_m) + \frac{\sigma^2_{\max}}{Q_m^{(p)}}\mathbb{E}\left[p_2(m)\right]\right]. \quad (38)$$

Moreover, 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* $x \geq 0$*, an event of probability at least* $1 - e^{-x}$ *exists on which*

$$\left|\overline{\delta}(m)\right| \leq L_{a_{g\epsilon},\xi_{g\epsilon},A,\sigma_{\max}} n^{-1/2} x^{\xi_{g\epsilon}+1/2}. \tag{39}$$

Lemma 12 is proved in Section 8.10. Third, Lemma 13 ensures that empirical frequencies $n\widehat{p}_\lambda$ are not too far from the expected ones np_λ.

Lemma 13. *Let* $(p_\lambda)_{\lambda\in\Lambda_m}$ *be non-negative real numbers of sum 1,* $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ *be a multinomial vector of parameters* $(n;(p_\lambda)_{\lambda\in\Lambda_m})$ *and* $\gamma > 0$*. Assume that* $\mathrm{Card}(\Lambda_m) \leq n$ *and* $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n > 0$*. Then, an event of probability at least* $1 - Ln^{-\gamma}$ *exists on which*

$$\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq \frac{\min_{\lambda\in\Lambda_m}\{np_\lambda\}}{2} - 2(\gamma+1)\ln(n). \tag{40}$$

Proof of Lemma 13. First, for every $\lambda \in \Lambda_m$, Bernstein's inequality [55, Proposition 2.9] applied to $n\widehat{p}_\lambda$ shows that an event of probability at least $1-2n^{-(\gamma+1)}$ exists on which

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

Since $\sqrt{2np_\lambda(\gamma+1)\ln(n)} \leq (np_\lambda)/2 + (\gamma+1)\ln(n)$, (40) holds on an event of probability at least $1 - 2\,\mathrm{Card}(\Lambda_m)n^{-(\gamma+1)} \geq 1 - 2n^{-\gamma}$. □

Finally, Lemmas 14 and 15 below are useful to compare the expectations of p_1 and p_2 on the one hand, and the expectations of pen and $\mathrm{pen}_{\mathrm{id}}$ for possibly large models on the other hand.

Lemma 14 (Lemma 7 of [9])**.** *If* $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B \geq 1$*,*

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

where $\delta_{n,p}$ *is the same as in* (15)*. A similar result holds with* p_2 *instead of* $\widetilde{p}_2$ *inside the expectation.*

Lemma 15. *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* $C = C_W$ *(see Table 2). Then, if* $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq 3$*,*

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m)\right] \geq \frac{5}{4}\mathbb{E}^{\Lambda_m}\left[p_2(m)\right]. \tag{41}$$

If $\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq T$ *for some positive* T*,* (41) *still holds for weight vectors among:*

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

Lemma 15 is proved in Section 8.9.

8.7. 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.

8.7.1. 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 $\widehat{m}$, for every $m \in \widehat{\mathcal{M}}_n$,

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m})\right) \leq \ell\left(s, \widehat{s}_m\right) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m)\right). \quad (42)$$

The proof of Lemma 7 is divided into three main parts:

1. With a large probability, $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ is negligible in front of $\ell\left(s, \widehat{s}_m\right)$ uniformly over models S_m of "intermediate" dimension, that is $\left(\ln(n)\right)^{\gamma_1} \leq D_m \leq cn\left(\ln(n)\right)^{-1}$ for some constants $c, \gamma_1 > 0$. This relies on the concentration inequalities and comparisons of expectations stated in Sections 3.4 and 8.6.
2. The model $\widehat{m}$ selected by Resampling Penalization has an "intermediate" dimension. In order to prove this, a lower bound on

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

   is proved for large and small models, and this bound is showed to be larger than $\mathrm{crit}''(m_0)$, where S_{m_0} is the model of intermediate dimension belonging to the collection $\left(S_m\right)_{m\in\mathcal{M}_n}$ according to assumption (**P2**). Lemma 15 is crucial at this point.
3. The oracle model (that is the one minimizing $\ell\left(s, \widehat{s}_m\right)$) is also of "intermediate" dimension, which is proven similarly to point 2 with $\mathrm{crit}''(m)$ replaced by $\ell\left(s, \widehat{s}_m\right)$.

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

$$A_n(m) := \min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\} \qquad \text{and} \qquad B_n(m) = \min_{\lambda\in\Lambda_m}\left\{np_\lambda\right\}.$$

Let Ω_{n,γ_0} be the event on which the concentration inequalities of Propositions 3 and 10 and Lemmas 11 and 13 hold for every $m \in \mathcal{M}_n$ with $\gamma = \alpha_{\mathcal{M}} + \gamma_0$ (or similarly $x = (\alpha_{\mathcal{M}} + \gamma_0)\ln(n)$ in Lemma 11). Using assumption (**P1**), the union bound gives $\mathbb{P}\left(\Omega_{n,\gamma_0}\right) \geq 1 - L_{c_{\mathcal{M}}} n^{-\gamma_0}$.

1. pen is close to $\mathrm{pen}'_{\mathrm{id}}$ for intermediate models Let $c, \gamma_1 > 0$ be two constants to be chosen later, and consider $\widetilde{\mathcal{M}}_n$, the set of $m \in \mathcal{M}_n$ such that $\left(\ln(n)\right)^{\gamma_1} \leq D_m \leq cn\left(\ln(n)\right)^{-1}$. According to ($\mathbf{Ar}_\ell^{\mathbf{X}}$), for every $m \in \widetilde{\mathcal{M}}_n$,

$B_n(m) \geq c_{\mathrm{r},\ell}^X c^{-1} \ln(n)$ so that (40) ensures that $A_n(m) \geq \ln(n)$ on Ω_{n,γ_0} if $c \leq L_{c_{\mathrm{r},\ell}^X,\alpha_{\mathcal{M}},\gamma_0}$. In particular, $\widetilde{\mathcal{M}}_n \subset \widehat{\mathcal{M}}_n$ on Ω_{n,γ_0}.

Assume also 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,

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

is smaller than $L_{(\mathbf{Bg})}\left(\ln(n)\right)^{-1}\left(\ell\left(s,s_m\right) + \mathbb{E}\left[p_2(m)\right]\right)$ on Ω_{n,γ_0} provided that $c \leq L_{c_{\mathrm{r},\ell}^X,\gamma}$ (to ensure that $B_n(m)$ is large enough) and $\gamma_1 \geq 2\xi_\ell + 6$. Fix now $c = L_{c_{\mathrm{r},\ell}^X,\gamma} > 0$ and $\gamma_1 = L_{\xi_\ell}$ satisfying these conditions. Using Proposition 2, Lemma 14 and the lower bound on $B_n(m)$, we have for every $m \in \widetilde{\mathcal{M}}_n$

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

as soon as $n \geq L_{(\mathbf{Bg})}$ (this restriction is necessary because the bounds are in terms of $\ell\left(s,\widehat{s}_m\right)$ instead of $\ell\left(s,s_m\right) + \mathbb{E}\left[p_2\right]$). Combined with (42), this gives: if $n \geq L_{(\mathbf{Bg})}$

$$\ell\left(s,\widehat{s}_{\widehat{m}}\right)\mathbf{1}_{\widehat{m}\in\widetilde{\mathcal{M}}_n} \leq \left[2\eta - 1 + \frac{L_{(\mathbf{Bg})}}{\left(\ln(n)\right)^{1/4}}\right]\inf_{m\in\widetilde{\mathcal{M}}_n}\left\{\ell\left(s,\widehat{s}_m\right)\right\}. \tag{43}$$

2. $\widehat{m}$ has an "intermediate" dimension The penalized empirical criterion $\mathrm{crit}(m) = P_n\gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$ has the same minimizers as

$$\mathrm{crit}''(m) = \ell\left(s,\widehat{s}_m\right) + \mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m) = \ell\left(s,s_m\right) + \mathrm{pen}(m) - p_2(m) + \overline{\delta}(m)$$

over $\widehat{\mathcal{M}}_n$.

According to (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If $n \geq L_{(\mathbf{Bg})}$, $m_0 \in \widetilde{\mathcal{M}}_n$ so that (using (**Ap**) and the same inequalities as in the first part of the proof)

$$\mathrm{crit}''(m_0) \leq \ell\left(s,s_{m_0}\right) + \left|\overline{\delta}(m_0)\right| + \mathrm{pen}(m_0) \leq L_{(\mathbf{Bg})}\left(n^{-\beta_2/2} + n^{-1/2}\right). \tag{44}$$

Therefore, it remains to provide lower bounds on $\mathrm{crit}''(m)$ for $m \notin \widetilde{\mathcal{M}}_n$.

On the one hand, on Ω_{n,γ_0} if $D_m < \left(\ln(n)\right)^{\gamma_1}$,

$$\begin{aligned}\mathrm{crit}''(m) &\geq \ell\left(s,s_m\right) - \left|\overline{\delta}(m)\right| - p_2(m)\\ &\geq C_{\mathrm{b}}^-\left(\ln(n)\right)^{-\gamma_1\beta_1} - L_{A,\gamma_0}\sqrt{\frac{\ln(n)}{n}} - L_{(\mathbf{Bg})}\frac{\left(\ln(n)\right)^{1+\xi_\ell+\gamma_1}}{n}.\end{aligned} \tag{45}$$

On the other hand, if $D_m > cn\left(\ln(n)\right)^{-1}$ and $m \in \widehat{\mathcal{M}}_n$, by Lemma 15, $\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}(m) - p_2(m)\right] \geq \mathbb{E}^{\Lambda_m}\left[p_2(m)\right]/4$. Therefore, we have $\mathrm{pen}(m) - p_2(m) \geq (1 - L_{(\mathbf{Bg})}n^{-1/4})\mathbb{E}\left[p_2(m)\right]$ on Ω_{n,γ_0}, so that

$$\mathrm{crit}''(m) \geq \mathrm{pen}(m) - p_2(m) - \left|\overline{\delta}(m)\right| \geq L_{(\mathbf{Bg})}\left(\ln(n)\right)^{-1} \tag{46}$$

when $n \geq L_{(\mathbf{Bg})}$. Comparing (44), (45) and (46), it follows that any minimizer $\widehat{m}$ of crit over $\widehat{\mathcal{M}}_n$ belongs to $\widetilde{\mathcal{M}}_n$ on Ω_{n,γ_0}, provided that $n \geq L_{(\mathbf{Bg})}$.

3. the oracle has an "intermediate" dimension It remains to prove that the infimum can be extended to $\mathcal{M}_n$ on the right-hand side of (43), 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 (35) 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}. Moreover, for every $m \notin \widetilde{\mathcal{M}}_n$, either $D_m < (\ln(n))^{\gamma_1}$ and $\ell(s, \widehat{s}_m) \geq \ell(s, s_m) \geq L_{(\mathbf{Bg})} (\ln(n))^{-\gamma_1 \beta_1}$ or $D_m > cn (\ln(n))^{-1}$ and $\ell(s, \widehat{s}_m) \geq p_1(m) \geq L_{(\mathbf{Bg})} (\ln(n))^{-2}$ on Ω_{n,γ_0} by (35) as soon as $n \geq L_{(\mathbf{Bg})}$. Hence, if $n \geq L_{(\mathbf{Bg})}$, $m \notin \widetilde{\mathcal{M}}_n$ cannot contribute to the infimum in the right-hand side of (43). This concludes the proof of (31) in the bounded case. □

8.7.2. Unbounded case

The proof of the bounded case has to be slightly modified. In the definition of Ω_{n,γ_0}, the concentration inequalities of Lemma 11 are replaced by those of Lemma 12. Then, γ_1 has to be chosen such that $\gamma_1 \geq 2\xi_{g\epsilon} + 3$. The rest of the proof of (43) is unchanged.

In order to prove that $\widehat{m} \in \widetilde{\mathcal{M}}_n$, (45) has to be slightly changed because of the use of (39) instead of (36) to bound $\overline{\delta}(m)$. The final part of the proof is then modified similarly. □

8.7.3. Proof of Remark 8

We now prove the assertion made in Remark 8 below Lemma 7. Starting from (43), we can prove in the same way that $D_{\widehat{m}} \leq cn (\ln(n))^{-1}$, but $D_{\widehat{m}} < (\ln(n))^{\gamma_1}$ cannot be excluded.

Let $m \in \widehat{\mathcal{M}}_n$ such that $D_m < (\ln(n))^{\gamma_1}$. Assume first that

$$\ell(s, s_m) \geq \frac{2\eta - 1 + \varepsilon_n}{1 - (\ln(n))^{-1}} \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{(\ln(n))^{\xi_\ell + \gamma_1 + 2}}{\left(1 - (\ln(n))^{-1}\right) n}, \tag{47}$$

where $\varepsilon_n \leq L_{(\mathbf{Bg})} (\ln(n))^{-1/4}$ comes from (8). Then, on Ω_{n,γ_0}, using (36) with $\eta = (\ln(n))^{-1}$ and (47),

$$\begin{aligned}
\mathrm{crit}''(m) &\geq \ell(s, s_m) - \left|\overline{\delta}(m)\right| - p_2(m) \\
&\geq (2\eta - 1 + \varepsilon_n) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{(\ln(n))^{\xi_\ell + \gamma_1 + 1} \left(\ln(n) - L_{(\mathbf{Bg})} \right)}{n} \\
&\geq (2\eta - 1 + \varepsilon_n) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{(\ln(n))^{\xi_\ell + \gamma_1 + 2}}{2n}, \qquad (48)
\end{aligned}$$

provided that $n \geq L_{(\mathbf{Bg})}$. In addition, 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},

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

and this upper bound is smaller than the lower bound in (48).

Hence, on Ω_{n,γ_0}, if $D_{\widehat{m}} < (\ln(n))^{\gamma_1}$ (47) cannot be satisfied with $m = \widehat{m}$. Moreover, by (34), for every $m \in \mathcal{M}_n$ such that $D_m \leq cn\,(\ln(n))^{-1}$

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

on Ω_{n,γ_0}. Therefore,

$$\begin{aligned}
\ell\left(s, \widehat{s}_{\widehat{m}}\right) &= \ell\left(s, s_{\widehat{m}}\right) + \widetilde{p_1}(\widehat{m}) \\
&\leq \frac{2\eta - 1 + \varepsilon_n}{1 - (\ln(n))^{-1}} \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + L_{(\mathbf{Bg})} \frac{(\ln(n))^{\xi_\ell + \gamma_1 + 2}}{n} \\
&\leq \left(2\eta - 1 + (\ln(n))^{-1/5}\right) \inf_{m \in \widetilde{\mathcal{M}}_n} \{ \ell(s, \widehat{s}_m) \} + \frac{(\ln(n))^{\xi_\ell + \gamma_1 + 3}}{n} \qquad (49)
\end{aligned}$$

assuming that $n \geq L_{(\mathbf{Bg})}$.

When $D_{\widehat{m}} \geq (\ln(n))^{\gamma_1}$, (31) holds on Ω_{n,γ_0} which implies (49). Hence, (49) holds on Ω_{n,γ_0}.

Finally, with the same arguments as in Section 8.7.1, the infimum on the right-hand side of (49) can be extended to the set of $m \in \mathcal{M}_n$ such that $D_m \geq (\ln(n))^{\gamma_1}$, with the convention $\ell(s, \widehat{s}_m) = +\infty$ if $A_n(m) = 0$. Enlarging the constant K_1 to remove the condition $n \geq L_{(\mathbf{Bg})}$, (32) is proved to hold with $\gamma_2 = \gamma_1 + \xi_\ell + 3$. The proof is quite similar in the unbounded case. □

8.8. Expectations

Proof of Proposition 1. On the one hand, (11) and (15) are consequences of (29) and (30); note that (15) holds whatever the convention taken for p_1 and p_2 in Section 8.1.

On the other hand, (12) follows from Lemma 16 below which is slighlty more geenral since W is allowed to depend on $(\mathbf{1}_{X_i \in I_\lambda})_{(i,\lambda)}$. □

Lemma 16. *Let S_m be the model of histograms adapted to some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$, $W \in [0;\infty)^n$ be a random vector such that for every $\lambda \in \Lambda_m$, $(W_i)_{X_i \in I_\lambda}$ is exchangeable and independent of $(X_i, Y_i)_{X_i \in I_\lambda}$. Let $\mathrm{pen}(m)$ be defined by (7) and assume $\min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} \geq 1$. Then,*

$$\mathrm{pen}(m) = \frac{C}{n} \sum_{\lambda \in \Lambda_m} (R_{1,W}(n, \widehat{p}_\lambda) + R_{2,W}(n, \widehat{p}_\lambda)) \frac{n\widehat{p}_\lambda S_{\lambda,2} - S_{\lambda,1}^2}{n\widehat{p}_\lambda (n\widehat{p}_\lambda - 1)} \mathbf{1}_{n\widehat{p}_\lambda \geq 2}, \qquad (50)$$

where $R_{1,W}$ *and* $R_{2,W}$ *are defined by* (13) *and* (14)*, that is*

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

Proof of Lemma 16. First, as $\mathrm{pen}_{\mathrm{id}}(m)$ was split into $p_1(m)$ and $p_2(m)$ (plus a centered term), the resampling penalty (without the constant C) is split into two terms:

$$\widehat{p}_1(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\,\middle|\,\widehat{W}_\lambda>0\right] \tag{51}$$

$$\widehat{p}_2(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\right]. \tag{52}$$

A key quantity to compute is the following: for every $\lambda\in\Lambda_m$ and $\widehat{W}_\lambda>0$,

$$\begin{aligned}
&\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\,\middle|\,\widehat{W}_\lambda\right]\\
&=\mathbb{E}_W\left[\widehat{p}_\lambda\left(\frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\left(1-\frac{W_i}{\widehat{W}_\lambda}\right)\right)^2\,\middle|\,\widehat{W}_\lambda\right]\\
&=\frac{1}{n^2\widehat{p}_\lambda}\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2\,\mathbb{E}_W\left[\left(1-\frac{W_i}{\widehat{W}_\lambda}\right)^2\,\middle|\,\widehat{W}_\lambda\right] \qquad (53)\\
&\quad+\frac{1}{n^2\widehat{p}_\lambda}\sum_{\substack{i\neq j\\ X_i\in I_\lambda, X_j\in I_\lambda}}(Y_i-\beta_\lambda)(Y_j-\beta_\lambda)\mathbb{E}_W\left[\left(1-\frac{W_i}{\widehat{W}_\lambda}\right)\left(1-\frac{W_j}{\widehat{W}_\lambda}\right)\,\middle|\,\widehat{W}_\lambda\right].
\end{aligned}$$

Since the weights are exchangeable, $(W_i)_{X_i\in I_\lambda}$ is also exchangeable conditionally on $\widehat{W}_\lambda$ and $(X_i)_{1\le i\le n}$. Hence, the "variance" term

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

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

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

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

$$0=\mathbb{E}_W\left[\left(\sum_{X_i\in I_\lambda}\left(W_i-\widehat{W}_\lambda\right)\right)^2\,\middle|\,\widehat{W}_\lambda\right]$$

$$= n\widehat{p}_\lambda R_V(n, n\widehat{p}_\lambda, \widehat{W}_\lambda, \mathcal{D}(W)) + n\widehat{p}_\lambda \left(n\widehat{p}_\lambda - 1\right) R_C(n, n\widehat{p}_\lambda, \widehat{W}_\lambda, \mathcal{D}(W))$$

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

$$R_C(n, n\widehat{p}_\lambda, \widehat{W}_\lambda, W) = \frac{-1}{n\widehat{p}_\lambda - 1} R_V(n, n\widehat{p}_\lambda, \widehat{W}_\lambda, \mathcal{D}(W)) \tag{54}$$

and $R_V(n, 1, \widehat{W}_\lambda, \mathcal{D}(W)) = 0$. Then, (53) and (54) imply

$$\begin{aligned}\mathbb{E}_W\left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda\right)^2 \,\middle|\, \widehat{W}_\lambda\right] &= \frac{R_V(n, n\widehat{p}_\lambda, \widehat{W}_\lambda, \mathcal{D}(W))}{\widehat{W}_\lambda n^2 \widehat{p}_\lambda} \mathbf{1}_{n\widehat{p}_\lambda \geq 2} \\ &\quad \times \left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - 1} S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda - 1} S_{\lambda,1}^2\right]\end{aligned} \tag{55}$$

Finally, (50) follows from the combination of (51) and (52) with (55). □

8.9. Resampling constants

Some results relative to the exchangeable weights introduced in Section 2.2 are proved in this subsection. First, Lemma 17 below provides explicit formulas for $R_{1,W}(n, \widehat{p}_\lambda)$ and $R_{2,W}(n, \widehat{p}_\lambda)$ which appear in the explicit formula (50) for the resampling penalty.

Lemma 17. *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 $q \in \{1, \dots, n\}$,*

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

$$R_{1,\mathrm{Rad}(p)} = \frac{1}{p} e^+_{\mathcal{B}(n\widehat{p}_\lambda, p)} - 1 \qquad R_{2,\mathrm{Rad}(p)} = \frac{1}{p} - 1 \tag{57}$$

$$R_{1,\mathrm{Poi}(\mu)} = \frac{1}{\mu} e^+_{\mathcal{P}(n\widehat{p}_\lambda \mu)}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \qquad R_{2,\mathrm{Poi}(\mu)} = \frac{1}{\mu}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right) \tag{58}$$

$$R_{1,\mathrm{Rho}(q)} = \frac{n}{q} e^+_{\mathcal{H}(n, n\widehat{p}_\lambda, q)} - 1 \qquad R_{2,\mathrm{Rho}(q)} = \frac{n}{q} - 1 \tag{59}$$

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

where $\mathcal{B}$, $\mathcal{P}$ and $\mathcal{H}$ denote 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 17. Since W is independent of the data, the observations with $X_i \in I_\lambda$ can be assumed to be 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 exchangeable). Hence, by definition of $\widehat{W}_\lambda = (n\widehat{p}_\lambda)^{-1} \sum_{i=1}^{n\widehat{p}_\lambda} W_i$,

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

Then, the quantity

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

appearing both in $R_{1,W}$ and $R_{2,W}$ is the variance of the weight W_i conditionally on $\widehat{W}_\lambda$.

Exchangeable subsampling weights A *subsampling weight* is defined as any resampling weight W such that $W_i \in \{0, \kappa\}$ a.s. for every i. Such weights can be written $W_i = \kappa \mathbf{1}_{i \in I}$ for some random $I \subset \{1, \dots, n\}$. Rad and Rho are the two main examples of such weights and they are both exchangeable. This kind of weights are called "bootstrap without replacement weights" in [66, Example 3.6.14]. First, when W is an exchangeable subsampling weight, (60) implies

$$\widehat{W}_\lambda = \mathbb{E}_W\left[W_i \middle| \widehat{W}_\lambda \right] = \kappa \mathbb{P}\left(W_i = \kappa \middle| \widehat{W}_\lambda \right)$$

so that

$$\mathcal{D}\left(W_i \middle| \widehat{W}_\lambda \right) = \kappa \mathcal{B}(\kappa^{-1} \widehat{W}_\lambda) \qquad \text{and} \qquad R_V(\widehat{W}_\lambda) = \widehat{W}_\lambda(\kappa - \widehat{W}_\lambda).$$

Then, this result is applied to Rad with $\kappa = p^{-1}$ and $\mathcal{D}(\widehat{W}_\lambda) = (n\widehat{p}_\lambda p)^{-1} \times \mathcal{B}(n\widehat{p}_\lambda, p)$ which proves (57). In the Rho case, $\kappa = (n/q)$ and $\mathcal{D}(\widehat{W}_\lambda) = (q\widehat{p}_\lambda)^{-1} \times \mathcal{H}\left(n, n\widehat{p}_\lambda, q\right)$ so that (59) follows. The Loo is a particular case of Rho (with $q = n - 1$) and $e^+_{\mathcal{H}(n, n\widehat{p}_\lambda, n-1)}$ can be computed with (22) 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 independent random variables with uniform distribution over $\{1, \dots, n\}$. Therefore,

$$\mathcal{D}(\widehat{W}_\lambda) = (M\widehat{p}_\lambda)^{-1} \mathcal{B}(M, \widehat{p}_\lambda) \quad \text{and} \quad \mathcal{D}\left(W_i \middle| \widehat{W}_\lambda \right) = \frac{n}{M} \mathcal{B}\left(M\widehat{p}_\lambda \widehat{W}_\lambda, \frac{1}{n\widehat{p}_\lambda} \right)$$

so that

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

and (56) follows.

Poisson One can check that the weights defined by (61) with $M = N_n \sim \mathcal{P}(\mu n)$ independent of the $(U_j)_{j \geq 1}$, are actually Poisson (μ) weights; this is the classical poissonization trick [66, Chapter 3.5]. Moreover, conditionally on $\widehat{W}_\lambda$ and $N_n = M$, the same reasoning as for Efron(M) (with a multiplicative constant μ^{-1} instead of n/M) leads to (58). □

Proof of Proposition 2. From (50), (16) holds with

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

Combining Lemma 17 with Lemma 4 (for Efr and Rad), Lemma 5 (for Rho and Loo) and Lemma 6 (for Poi), the following non-asymptotic bounds are obtained:

1. Efron (M_n): let $\kappa_1 = 5.1$ and $\kappa_2 = 3.2$, then
$$(\kappa_2 - 1) \wedge \left(\frac{\kappa_1}{\left(B n \widehat{p}_\lambda \right)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penEfr}(M_n))} \geq \frac{-2}{n\widehat{p}_\lambda} - e^{-Bn\widehat{p}_\lambda}. \tag{62}$$

2. Rademacher (p):
$$\frac{2}{1-p} \left[(\kappa_2 - 1) \wedge \left(\frac{\kappa_1}{\left(np\widehat{p}_\lambda \right)^{1/4}} \right) \right] \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRad}(p))} \geq \frac{-2e^{-pn\widehat{p}_\lambda}}{1-p} \tag{63}$$
$$\left(1 + 3 \times 10^{-4}\right) \wedge \left(\frac{\kappa_1 \times 2^{1/4}}{\left(n\widehat{p}_\lambda \right)^{1/4}} \right) \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRad}(1/2))} \geq -\mathbf{1}_{n\widehat{p}_\lambda \leq 2}. \tag{64}$$

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

4. Random hold-out (q_n): on the one hand,
$$\delta_{n,\widehat{p}_\lambda}^{(\mathrm{penRho}(q_n))} = \frac{n}{n-q} \left(e^+_{\mathcal{H}(n, n\widehat{p}_\lambda, q_n)} - 1 \right) \geq \frac{e^{-n\widehat{p}_\lambda B_-}}{1 - B_+},$$
where the lower bounds assume 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 $q_n = \lfloor n/2 \rfloor$, this upper bound is combined with (21).

5. Leave-one-out:
$$\frac{\mathbf{1}_{n\widehat{p}_\lambda \geq 2}}{n\widehat{p}_\lambda - 1} \geq \delta_{n,\widehat{p}_\lambda}^{(\mathrm{penLoo})} \geq -\mathbf{1}_{n\widehat{p}_\lambda = 1}. \tag{66}$$

□

Proof of Lemma 15. Lemma 15 is a byproduct of the proof of Proposition 2 (combined with Lemma 14). □

8.10. Concentration inequalities

In this subsection, concentration inequalities are proved for the resampling penalty (Proposition 3) and for $\overline{\delta}(m)$ with unbounded data (Lemma 12).

8.10.1. Proof of Proposition 3

According to (50), $\mathrm{pen}(m)$ is a U-statistics of order 2 conditionally on $(\mathbf{1}_{X_i \in I_\lambda})_{(i,\lambda)}$. Then, [9, Lemma 5] with

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

implies that for every $q \geq 2$

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

Conditional concentration inequalities follow from the classical link between moments and concentration [6, Lemma 8.10], with a probability bound $1 - n^{-\gamma}$. Since $1 - n^{-\gamma}$ is deterministic, this implies unconditional concentration inequalities.

The second statement follows from the proof of Proposition 2 where non-asymptotic upper bounds on

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

can be found. □

8.10.2. Proof of Lemma 12

From [6, Lemma 8.18] which is stated and proved in [8],

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

Note that $\epsilon\sigma(X)(s_m(X) - s(X))$ is centered conditionally on $X \in I_\lambda$ for every $\lambda \in \Lambda_m$. Hence,

$$\left\| \overline{\delta}(m) \right\|_q \leq \frac{2\sqrt{\kappa}\sqrt{q}}{\sqrt{n}} \left(\left\| s - s_m \right\|_\infty^2 + 2\sigma_{\max} \left\| s - s_m \right\|_\infty \left\| \epsilon \right\|_q \right). \tag{67}$$

Using now assumptions $(\mathbf{A}_{\mathbf{g},\epsilon})$ and $(\mathbf{A}\delta)$, for every $q \geq 2$,

$$\left\| \overline{\delta}(m) \right\|_q \leq 2\sqrt{\kappa}\sqrt{q} \left((c^g_{\Delta,m})^2 \ell\left(s, s_m\right) + 2c^g_{\Delta,m} \sqrt{\ell\left(s, s_m\right)} P^{g\epsilon}(q)\sigma_{\max} \right) \frac{1}{\sqrt{n}}$$

$$\leq L_{c^g_{\Delta,m}} \sqrt{q} D_m^{-1/2} \ell\left(s, s_m\right) + L_{a_{g\epsilon}, \xi_{g\epsilon}, c^g_{\Delta,m}} q^{\xi_{g\epsilon}+1/2} \frac{\sigma^2_{\max} \sqrt{D_m}}{n}.$$

Taking $\theta = D_m^{-1/2}$, (38) follows from the classical link between moments and concentration inequalities [6, Lemma 8.10]. For the second statement, start back from (67) and use that $\|s - s_m\|_\infty \leq 2A$. □

8.11. Expectations of inverses

This subsection is devoted to the proofs of the lemmas of Section 3.4.3. Note that [9, Section 2 of the Technical appendix] explains how to generalize (18) to a wide class of random variables. Two useful results can be found in [9, Technical appendix]: first, the general lower bound

$$e_Z^+ \geq \mathbb{P}(Z > 0), \tag{68}$$

comes from Jensen inequality. Second, defining

$$e^0_{\mathcal{D}(Z)} := \mathbb{E}\left[Z\right] \mathbb{E}\left[Z^{-1} \mathbf{1}_{Z>0}\right] = e_Z^+ \mathbb{P}(Z > 0), \tag{69}$$

the following upper bound holds as soon as $\mathbb{P}(c_Z > Z > 0) = 0$:

$$\begin{aligned} \forall \alpha > 0, \quad e^0_Z &= \mathbb{E}\left[Z^{-1} \mathbf{1}_{\alpha \mathbb{E}[Z] > Z > 0}\right] \mathbb{E}[Z] + \mathbb{E}\left[Z^{-1} \mathbf{1}_{Z \geq \alpha \mathbb{E}[Z]}\right] \mathbb{E}[Z] \\ &\leq \mathbb{P}\left(\alpha \mathbb{E}[Z] > Z > 0\right) \mathbb{E}[Z] c_Z^{-1} + \alpha^{-1}. \end{aligned} \tag{70}$$

8.11.1. Binomial case (proof of (19) in Lemma 4)

When $n \geq 9$, the upper bound follows from (69) together with Lemma 4.1 of [38] (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 [6, Section 8.7]). For the lower bound, the crucial point is that $Z \sim \mathcal{B}\left(n, \frac{1}{2}\right)$ is nonnegative and symmetric, that is, $\mathcal{D}(Z) = \mathcal{D}(n - Z)$. Using only this property and defining $p_0 = \mathbb{P}(Z = 0) = \mathbb{P}(Z = n) = 2^{-n}$, we have

$$\begin{aligned} e_Z^+ &= \frac{\mathbb{P}\left(Z = n \,|\, Z > 0\right)}{2} + \mathbb{E}\left[\frac{1}{Z} \,\middle|\, 0 < Z < 2\right] \frac{n}{2} \frac{\mathbb{P}(0 < Z < n)}{\mathbb{P}(Z > 0)} \\ &= \frac{p_0}{2(1-p_0)} + \frac{1-2p_0}{1-p_0} \frac{n}{2} \mathbb{E}\left[\frac{1}{2}\left(\frac{1}{Z} + \frac{1}{n-Z}\right) \,\middle|\, 0 < Z < n\right] \\ &= \frac{p_0}{2(1-p_0)} + \frac{1-2p_0}{1-p_0}\left(1 + \frac{n}{2} \mathbb{E}\left[\frac{\left(Z - \frac{n}{2}\right)^2}{Z(n-Z)} \,\middle|\, 0 < Z < n\right]\right). \end{aligned} \tag{71}$$

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

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

if $n \geq 3$. Putting this into (71), we obtain:

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

8.11.2. Hypergeometric case (proof of Lemma 5)

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

General lower bound It follows from (68),

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

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

A general upper bound According to (69) and the lower bound for $\mathbb{P}(Z>0)$ above, an upper bound on $e^+_{\mathcal{H}(n,r,q)}$ can be derived from an upper bound on $e^0_{\mathcal{H}(n,r,q)}$. Recall the following concentration result by Hush and Scovel [41]: for every $x \geq 2$,

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

Combined with the above concentration inequality, (70) with $c_Z=1$, $\mathbb{E}[Z] = qrn^{-1}$ and $\alpha = 1-\frac{n\beta}{q}$ for any $\frac{q}{n} > \beta \geq \frac{2}{r}$ yields

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

Therefore,

$$e^+_{\mathcal{H}(n,r,q)} \leq \frac{\inf_{\frac{q}{n}>\beta\geq\frac{2}{r}}\left\{\frac{qr}{n}\exp\left[-\frac{2(\beta r-1)^2}{r+1}\right]+\frac{1}{1-\frac{n\beta}{q}}\right\}}{1-\exp\left(-\frac{qr}{n}\right)} \tag{72}$$

holds for every $n \geq r, q \geq 1$.

End of the proof of (20) With the additional conditions on n, r and q, β can be taken equal to $\frac{1+\sqrt{\frac{3}{4}\ln(r)(r+1)}}{r}$ in (72) so that

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

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

Using (69) and the upper bound on $\mathbb{P}(Z=0)$, (20) follows 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 Assume now that $q = \lfloor \frac{n}{2} \rfloor$ so that $\frac{n}{q} = 2 + \frac{1}{\lfloor \frac{n}{2} \rfloor} \leq 3$ and tends to 2 when n tends to infinity.

For $r \geq 6$, $\beta = \frac{2}{r}$ in (72) yields

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

For $r \geq 10$, $\beta = \frac{1}{4} + \frac{1}{r}$ in (72) yields

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

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 Assume now $q = n-1$. On the one hand, if $r = 1$, the conditioning makes Z deterministic and equal to 1 so that

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

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

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

Hence,

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

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

"Lpo" case As noticed in Lemma 17,

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

Moreover, when $r \geq p+1$ the support of $\mathcal{H}(n, r, n-p)$ is $\{r-p, \dots, r\}$ and

$$e^+_{\mathcal{H}(n,r,n-p)} = \frac{(n-p)r}{n} \sum_{j=r-p}^{r} \frac{\binom{r}{j}\binom{n-r}{n-p-j}}{j\binom{n}{n-p}}$$

$$= \frac{(n-p)r}{n} \sum_{k=(p+r-n)\vee 0}^{p} \frac{\binom{r}{k}\binom{n-r}{p-k}}{(r-k)\binom{n}{p}}.$$

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

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

The result follows. □

Remark 11 (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

$$\beta = \frac{1+\sqrt{(r_k+1)\ln\left(\frac{q_k r_k}{n_k}\right)}}{r_k}$$

in (72), which is possible for k sufficiently large. The lower bound is straightforward.

8.11.3. 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 $x \geq 0$, 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 $x \geq x_0 \geq 0$,

$$2\left[1 - e^{2(x_0-x)}\left(1 - \frac{g(x_0)}{2}\right)\right] \geq g(x) \geq 1 + (g(x_0) - 1)e^{x_0-x}. \tag{73}$$

Lower bound The general lower bound (68) gives

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

which can be improved. Indeed, if $g(x_0) \geq 1$, (73) shows that $g(x) \geq 1$ for every $x \geq x_0$. Since $g = H/h$ and for every $u \geq 0$,

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

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

Upper bound Using (73) 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 $2\epsilon^{-1}$,

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

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. The author also acknowledges several suggestions from the anonymous referees that greatly improved the paper.

References

[1] Marc Aerts, Gerda Claeskens, and Jeffrey D. Hart. Testing the fit of a parametric function. *J. Amer. Statist. Assoc.*, 94(447):869–879, 1999. MR1723323

[2] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970. MR0286233

[3] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. MR0483125

[4] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974. MR0343481

[5] Miguel A. Arcones and Evarist Giné. 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., pages 13–47. Wiley, New York, 1992. MR1197777

[6] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[7] Sylvain Arlot. Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression, December 2008. arXiv:0812.3141v1

[8] Sylvain Arlot. Technical appendix to "Model selection by resampling penalization", 2009. Appendix to hal-00262478.

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

[10] Sylvain Arlot, Gilles Blanchard, and Étienne Roquain. Some non-asymptotic results on resampling in high dimension, I: confidence regions. *Ann. Statist.*, 2008. To appear.

[11] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10(Feb):245–279, 2009.

[12] Jean-Yves Audibert. *Théorie Statistique de l'Apprentissage: une approche PAC-Bayésienne*. PhD thesis, Université Paris VI, June 2004.

[13] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000. MR1777129

[14] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002. MR1918295

[15] Philippe Barbe and Patrice Bertail. *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995. MR2195545

[16] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. MR1679028

[17] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[18] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005. MR2166554

[19] Peter L. Bartlett, Shahar Mendelson, and Petra Philips. Local complexities for empirical risk minimization. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 270–284. Springer, Berlin, 2004. MR2177915

[20] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. MR1848946

[21] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. MR2288064
[22] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989. MR1040644
[23] Prabir Burman. Estimation of equifrequency histograms. *Statist. Probab. Lett.*, 56(3):227–238, 2002. MR1892984
[24] Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Inst. Math. Statist., 2007. MR2483528
[25] Joseph E. Cavanaugh and Robert H. Shumway. A bootstrap variant of AIC for state-space model selection. *Statist. Sinica*, 7(2):473–496, 1997. MR1466691
[26] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. MR1843146
[27] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995. MR1379464
[28] Sam Efromovich and Mark Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996. MR1422411
[29] Bradley Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979. MR0515681
[30] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983. MR0711106
[31] Bradley Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470, 1986. MR0845884
[32] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438):548–560, 1997. MR1467848
[33] Magalie Fromont. Model selection by bootstrap penalization for classification. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 285–299. Springer, Berlin, 2004. MR2177916
[34] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3):165–207, 2007.
[35] Leonid Galtchouk and Sergey Pergamenshchikov. Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression via model selection, October 2008. arXiv:0810.1173.
[36] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.
[37] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electronic Journal of Statistics*, 2:1345–1372, 2008. MR2471290
[38] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A*

distribution-free theory of nonparametric regression. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[39] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992. MR1145237

[40] Peter Hall and Enno Mammen. On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.*, 22(4):2011–2030, 1994. MR1329180

[41] Don Hush and Clint Scovel. Concentration of the hypergeometric distribution. *Statist. Probab. Lett.*, 75(2):127–132, 2005. MR2206293

[42] Marie Hušková and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *Ann. Statist.*, 21(4):1811–1823, 1993. MR1245770

[43] Makio Ishiguro, Yosiyuki Sakamoto, and Genshiro Kitagawa. Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.*, 49(3):411–434, 1997. MR1482365

[44] C. Matthew Jones and Anatoly A. Zhigljavsky. Approximating the negative moments of the Poisson distribution. *Statist. Probab. Lett.*, 66(2):171–181, 2004. MR2029732

[45] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001. MR1842526

[46] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006. MR2329442

[47] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. MR1226450

[48] Robert A. Lew. Bounds on negative moments. *SIAM J. Appl. Math.*, 30(4):728–731, 1976. MR0501260

[49] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987. MR0902239

[50] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004. MR2089138

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

[52] Enno Mammen. *When does bootstrap work? Asymptotic results and simulations*, volume 77 of *Lecture Notes in Statistics*. Springer, 1992.

[53] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. MR1765618

[54] David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3):1611–1624, 1992. MR1186268

[55] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

[56] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999. MR1707286

[57] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993. MR1245301

[58] Marie Sauvé. Histogram selection in non Gaussian regression. *ESAIM: Probability and Statistics*, 13:70–86, 2009.

[59] Jun Shao. Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665, 1996. MR1395733

[60] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264, 1997. With comments and a rejoinder by the author. MR1466682

[61] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981. MR0614940

[62] Ritei Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica*, 7(2):375–394, 1997. MR1466687

[63] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980. MR0594650

[64] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth. MR0822050

[65] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G.A. Barnard, A.C. Atkinson, L.K. Chan, A.P. Dawid, F. Downton, J. Dickey, A.G. Baker, O. Barndorff-Nielsen, D.R. Cox, S. Giesser, D. Hinkley, R.R. Hocking, and A.S. Young, and with a reply by the authors. MR0356377

[66] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. MR1385671

[67] Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1350, 1986. With discussion and a rejoinder by the author. MR0868303

[68] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6):2450–2473, 2007. MR2382654

[69] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999. MR1742500

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