Electronic Journal of Statistics
Vol. 6 (2012) 1803–1837
ISSN: 1935-7524
DOI: 10.1214/12-EJS730

Oracle inequalities for cross-validation type procedures

Guillaume Lecué*

CNRS, LAMA, Université Paris-Est Marne-la-vallée, 77454 France
e-mail: guillaume.lecue@univ-mlv.fr

and

Charles Mitchell

Seminar für Statistik, ETH Zürich
e-mail: mitchell@stat.math.ethz.ch

Abstract: We prove oracle inequalities for three different types of adaptation procedures inspired by cross-validation and aggregation. These procedures are then applied to the construction of Lasso estimators and aggregation with exponential weights with data-driven regularization and temperature parameters, respectively. We also prove oracle inequalities for the cross-validation procedure itself under some convexity assumptions.

AMS 2000 subject classifications: Primary 62G99.
Keywords and phrases: Adaptation, aggregation, cross-validation, sparsity.

Received February 2012.

1. Introduction

In this paper, we construct adaptation procedures inspired by cross-validation. Adaptation procedures are of particular interest when one wants to adapt to an unknown parameter. Such a parameter can appear in statistical procedures for two reasons: either it is an unknown parameter of the model (complexity parameter, "concentration" parameter, geometric parameter, variance of the noise,...), or the construction of the procedure requires fitting a parameter that no theory is able to determine (regularization parameter, smoothing parameter, threshold,...). Thus it is very useful to have at hand some statistical procedure which can choose these unknown parameters in a data-dependent way. The construction of adaptation procedures has been one of the main topics in nonparametric statistics for the two last decades. Retracing the entire bibliography here is not possible. Nevertheless, we would like to refer the reader to some classical – and now pioneering – steps in this field like the model selection approach (cf. i.e. [3] and [25]), aggregation methods (cf. i.e. [27, 9] and [8]),

*Supported by French Agence Nationale de la Recherche ANR Grant "Prognostic" ANR-09-JCJC-0101-01.

empirical risk minimization (cf. i.e. [36, 16] and [5]) or Lepskii's adaptation method in [22, 23]. Of course many other approaches in some particular setups have been developed. But one of the most popular and universal strategy used for fitting unknown parameters or more generally to select algorithms is the Cross-Validation (CV). Cross-validation is a very important and widely applied family of model/ estimator/ parameter selection methods. The CV procedures can be traced back up the 30s with [17] where the key idea that training and testing a statistical procedure on the same data yield overoptimistic results. Among other, the CV procedure was studied for the selection of the bandwidth in kernel density estimation in [15] and [30], for the regression model in [29], in classification in [11]. Many other authors have been studying or using this method and we refer the reader to the survey of CV methods in model selection [2], the PhD thesis [10, 28] or [34] for more bibliographical references on this topic. The aim of this paper is to present and to study three procedures inspired by the CV procedure in the following general framework.

Let $(\mathcal{Z}, \mathcal{T})$ be a measurable space and $\mathcal{F}$ be a class of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$. On a very general level, our aim is to minimize a risk function $R : \mathcal{F} \to \mathbb{R}$ over its domain $\mathcal{F}$. This risk function is assumed to exist, but is unknown to us. To obtain information about it, though, we assume that it also appears as the expectation of a quantity we can sample from:

Let Z be a random variable with values in $\mathcal{Z}$ and denote its probability measure by π. Assume that there exists a "contrast" or loss function $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$ such that the risk of any $f \in \mathcal{F}$ can be written in the form

$$R(f) := \mathbb{E}\left[Q(Z, f)\right],$$

and that there exists a sequence $(Z_i)_{i\in\mathbb{N}}$ of i.i.d. random variables distributed according to π. For the purpose of statistical estimation, we have only access to a finite amounts of data from this sequence, say the first n variables $Z_1, \ldots, Z_n$.

The problem of risk minimization is a general formulation for many different kinds of statistical problems, and we shall introduce all our examples using this form. If the infimum

$$R^* := \inf_{f\in\mathcal{F}} R(f)$$

over all f in $\mathcal{F}$ is achieved by at least one function, we write f^* for some choice of such a minimizer in $\mathcal{F}$. In this paper, we will assume that $\inf_{f\in\mathcal{F}} R(f)$ is achieved – otherwise we can replace f^* by f_n^*, an element in $\mathcal{F}$ satisfying $R(f_n^*) \leq \inf_{f\in\mathcal{F}} R(f) + n^{-1}$, and still obtain the same results.

This model is best illustrated by its three key examples: regression, density estimation and classification.

Regression: Take $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, where $(\mathcal{X}, \mathcal{A})$ is a measurable space, and let $Z = (X, Y)$ be a random pair on $\mathcal{Z}$. In the regression framework, we would like to estimate the regression function $f^*(x) = \mathbb{E}\left[Y|X = x\right]$, $\forall x \in \mathcal{X}$. Take $\mathcal{F} = L^2(\mathcal{X}, \mathcal{A}, P^X)$, where P^X is the distribution of X. Consider the contrast function $Q((x, y), f) = (y - f(x))^2$ defined for any $(x, y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$. We have $R(f) = \mathbb{E}\left[Q((X, Y), f)\right] = \|f^* - f\|^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right]$, where $\zeta = Y - f^*(X)$

is usually called the noise or residual. Thus f^* is a minimizer of $R(\cdot)$ and the minimum achievable risk is $R^* = \mathbb{E}[\zeta^2]$.

Density estimation: Let $(\mathcal{Z}, \mathcal{T}, \mu)$ be a measure space, and take Z to be a random variable with values in $\mathcal{Z}$. We assume that the probability distribution π of Z is absolutely continuous with respect to μ and denote by f^* one version of its density. Consider $\mathcal{F}$ the set of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$, i.e. the set of all $\mathcal{T}$-measurable functions $f : \mathcal{Z} \to \mathbb{R}_+$ that integrate to 1. We consider the contrast function $Q(z, f) = -\log f(z)$ for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. The corresponding risk computes as $R(f) = \mathbb{E}[Q(Z, f)] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z))d\pi(z)$. Thus f^* is a minimizer of $R(\cdot)$ and the minimum achievable risk is $R^* = -\int_{\mathcal{Z}} \log(f^*(z))d\pi(z)$.

Instead of using the Kullback-Leibler loss, one can use the quadratic loss. The corresponding contrast function is $Q(z, f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z)$ for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. Using this contrast function, the risk of any $f \in \mathcal{F}$ works out as $R(f) = \mathbb{E}[Q(Z, f)] = ||f^* - f||^2_{L^2(\mu)} - \int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$. Thus f^* is a minimizer of $R(\cdot)$ and the corresponding minimal risk is $R^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

Classification framework: Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We assume that the space $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ is endowed with an unknown probability measure π, and consider a random pair $Z = (X, Y)$ which takes on values in $\mathcal{Z}$ and whose probability distribution is π. Denote by $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$, and furthermore let ϕ be a function from $\mathbb{R}$ to $\mathbb{R}$. For any $f \in \mathcal{F}$ consider the $\phi-$risk, $R(f) = \mathbb{E}[Q((X, Y), f)]$, where the contrast function is given by $Q((x, y), f) = \phi(yf(x))$ for any $(x, y) \in \mathcal{X} \times \{-1, 1\}$. In many situations, a minimizer f^* of the $\phi-$risk R over $\mathcal{F}$ (or the sign of f^*, if the latter takes on arbitrary real values) is equal to the Bayes rule $f^*_{Bayes}(x) = \text{Sign}(2\eta(x)-1), \forall x \in \mathcal{X}$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$ (cf. [38] and [4]).

We say that a statistic is a sequence of functions $\hat{f} = (\hat{f}^{(n)})_{n\in\mathbb{N}}$ such that each $\hat{f}^{(n)}$ is a map associating a function $\hat{f}^{(n)}(\cdot) := \hat{f}^{(n)}(D^{(n)})(\cdot)$ in $\mathcal{F}$ to each data set $D^{(n)} = \{Z_1, \ldots, Z_n\}$. If $\hat{f}$ is a statistic and n is an integer, the risk of its nth element $\hat{f}^{(n)}$ is defined as the $\sigma(D^{(n)})$-measurable random variable

$$R(\hat{f}^{(n)}(D^{(n)})) = \mathbb{E}[Q(Z, \hat{f}(D^{(n)}))|D^{(n)}].$$

We assume that we know how to construct some statistics $\hat{f}_\lambda$ for λ in a set of indexes Λ. The aim of this work is to construct procedures $\bar{f} := (\bar{f}^{(n)})_{n\in\mathbb{N}}$ satisfying oracle inequalities that is inequalities like, for any sample size n,

$$\mathbb{E}[R(\bar{f}^{(n)}(D^{(n)})) - R^*] \leq C \inf_{\lambda\in\Lambda} \mathbb{E}[R(\hat{f}^{(n)}_\lambda(D^{(n)})) - R^*] + r(n, \Lambda) \tag{1.1}$$

where $C \geq 1$ is a constant and $r(n, \Lambda)$ is a residue term which we would like to keep as small as possible. Controlling this residue will depend on some complexity parameter of the excess loss function class $\{Q(\cdot, \hat{f}^{(n)}_\lambda(D^{(n)})) - Q(\cdot, f^*) : \lambda \in \Lambda\}$, as well as on a margin parameter that limits the behavior of the contrast function around the risk minimizer (cf. Assumptions (A) in Section 2). Note that for any measurable function f, the function $z \in \mathcal{Z} \mapsto Q(z, f) - Q(z, f^*)$ is called the excess loss functions of f.

The paper is organized as follows. In Section 2, we introduce some adaptation procedures which are then proved to satisfy oracle inequalities in the finite case $|\Lambda| = p$. Section 3 is devoted to the study of a general non-finite case. In Section 4, we apply our adaptation procedures to the construction of Lasso estimators with a data-driven regularization parameter and aggregates with exponential weights with a data-driven temperature parameter. Finally, the main proofs are provided in Section 5.

2. Procedures and oracle inequalities

In this section we provide some oracle inequalities for several procedures selecting or aggregating estimators: first two modified versions of the cross-validation procedure, then cross-validation procedure itself, and then finally we discuss aggregation with multiple splitting.

2.1. Classical cross-validation procedures

The key feature of the CV procedure, the use of multiple splits to train and test the candidate estimator, renders it somewhat more difficult to handle in a theoretial way. Nevertheless, we shall show that a carefully crafted risk inequality opens the door to oracle inequalities for cross-validation too. In this section, we have to pay careful attention to the exact choice of the splits of our data, especially when retraining the selected model to obtain our final estimator(s).

First we shall introduce some notation. Let n be an integer, and V be a divisor of n. We split the data set $D^{(n)}$ into V disjoint subsets of equal size $n_C = n/V$, namely, for every $k = 1, \dots, V$,

$$B_k = \{Z_{(k-1)n_C+1}, \dots, Z_{kn_C}\}, \tag{2.1}$$

which shall be test sets, and their complements

$$D_k = \cup_{j=1:j\neq k}^{V} B_j, \tag{2.2}$$

the corresponding training sets. Note that D_k is a data set of size $n_V := n - n_C$.

Let $Q(Z, f)$ be a contrast function whose arguments are a data point Z and a parameter $f \in \mathcal{F}$. For a statistic $\hat{f} = (\hat{f}^{(n)})_n$, we define the *V-fold CV empirical risk* by

$$R_{n,V}(\hat{f}) = \frac{1}{V}\sum_{k=1}^{V} \frac{1}{n_C} \sum_{i=(k-1)n_C+1}^{kn_C} Q(Z_i, \hat{f}^{(n_V)}(D_k)). \tag{2.3}$$

Let p statistics $\hat{f}_1, \dots, \hat{f}_p$ be given. The *V-fold CV procedure* is the procedure $\bar{f}_{VCV} = (\bar{f}^{(n)}_{VCV})_n$ defined, for any n, by

$$\bar{f}^{(n)}_{VCV}(D^{(n)}) = \hat{f}^{(n)}_{\hat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \hat{\jmath}(D^{(n)}) \in \operatorname{Arg}\min_{j\in\{1,\dots,p\}} R_{n,V}(\hat{f}_j). \tag{2.4}$$

Perhaps the oldest, and certainly the most frequently studied, cross-validation scheme is n-fold or *leave-one-out* cross-validation. It forms the intersection be-

tween the class of V-fold cross-validation schemes and the class of *leave-m-out CV* schemes, defined by

$$\bar{f}^{(n)}_{lmo}(D^{(n)}) = \hat{f}^{(n)}_{\hat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \hat{\jmath}(D^{(n)}) \in \operatorname{Arg}\min_{j\in\{1,\dots,p\}} R_{n,-m}(\hat{f}_j), \tag{2.5}$$

where $R_{n,-m}$ is defined as

$$R_{n,-m}(\hat{f}) = \binom{n}{m}^{-1} \sum_{C\subset\{1,\dots,n\}:|C|=m} \frac{1}{m}\sum_{i\in C} Q(Z_i, \hat{f}^{(n-m)}((Z_k)_{k\in\{1,\dots,n\}\setminus C})).$$

This method does however become very computationally inadequate as soon as m is no longer 1, as there are far too many subsets of $\{1,\dots,n\}$ to average over (however, it has been pointed out in [2] that in some settings the leave-m-out procedures are tractable in practice). One possible solution for this is *balanced incomplete cross-validation*, where cross-validation is treated as a block design and the available pieces of data are all used equally often for training, and equally often for testing. Alternatively, we could use *Monte Carlo cross-validation*, where the training and testing subsets are drawn randomly – without replacement – from the available data. See [28] for a discussion of all these methods.

We can place all of these cross-validation schemes into one general framework as follows. For any subset $C \subset \{1,\dots,n\}$ of indices, write $D_{(C)}$ for $\{Z_i : i \in C\}$ and $D_{(C)'}$ for $\{Z_i : i \notin C\}$. Assume that a fixed value n_C be given (the size of test sets), and define $n_V = n - n_C$. Let $C_1,\dots,C_{N_C}$ be N_C subsets of $\{1,\dots,n\}$, each of size n_V. Now for any statistic $\hat{f}$ define the *CV risk*

$$R_{n_C}(\hat{f}) = \frac{1}{N_C}\sum_{k=1}^{N_C}\frac{1}{n_C}\sum_{i\notin C_k} Q(Z_i, \hat{f}^{(n_V)}(D_{(C_k)})), \tag{2.6}$$

and its minimizer by

$$\hat{f}^{(n)}_{CV}(D^{(n)}) = \hat{f}^{(n)}_{\hat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \hat{\jmath}(D^{(n)}) \in \operatorname{Arg}\min_{j\in\{1,\dots,p\}} R_{n_C}(\hat{f}_j). \tag{2.7}$$

2.2. The modified CV procedure and its average version

In this subsection, we introduce the selection procedures that we will be studying later. We use the notations introduced in the previous subsection.

To introduce the modified CV procedure, we consider some integer V and we assume that V divides n. We consider the splits $(B_1, D_1),\dots,(B_V, D_V)$ of the data introduced in (2.1) and (2.2). We define the **modified CV procedure (mCV)** by

$$\bar{f}^{(n)}_{mCV}(D^{(n)}) = \hat{f}^{(n_V)}_{\hat{\jmath}(D^{(n)})}(D^{(n_V)}) \tag{2.8}$$

where $D^{(n_V)} = \{Z_1,\dots,Z_{n_V}\}$ and, for the V-fold CV empirical risk $R_{n,V}$ introduced in (2.3),

$$\hat{\jmath}(D^{(n)}) \in \operatorname{Arg}\min_{j\in\{1,\dots,p\}} R_{n,V}(\hat{f}_j).$$

For the average version of the mCV procedure, we don't have to split the data in the same "organized" way as in (2.1) and (2.2). We can consider the more general second partition scheme introduced in the second part of the previous subsection that we recall now for the reader convenience: Let N_C and $1 \leq n_C < n$ be two integers and set $n_V = n - n_C$. Let $C_1, \ldots, C_{N_C}$ be subsets of $\{1, \ldots, n\}$ each of size n_V. We define the **averaged version of the modified CV procedure (amCV)** by:

$$\hat{f}^{(n)}_{amCV}(D^{(n)}) = \frac{1}{N_C} \sum_{k=1}^{N_C} \hat{f}^{(n_V)}_{\hat{\jmath}(D^{(n)})}(D_{(C_k)}). \tag{2.9}$$

where, for the CV-risk R_{n_C} introduced in (2.6),

$$\hat{\jmath}(D^{(n)}) \in \operatorname{Arg} \min_{j \in \{1,\ldots,p\}} R_{n_C}(\hat{f}_j).$$

Note that the mCV procedure is a model selection procedure (taking values in the dictionary itself) whereas the amCV procedure is a model combination or aggregation procedure (taking values in the convex hull of the dictionary).

We did not consider the same partition scheme of the data for the two procedures. The one considered for the amCV is more general but to obtain oracle inequalities for the amCV we will need the convexity of the risk. Whereas for the mCV, the partition scheme is the one used for the VCV method and will only require a weak assumption on the basis statistics $\hat{f}_1, \ldots, \hat{f}_p$. For each one of our results, we will consider two different setups depending on the procedure that we want to study and the assumptions of the problem.

Note that the difference between the classical VCV procedure defined in (2.4) and our mCV procedure is that $\bar{f}^{(n)}_{mCV}$ takes its values in $\{\hat{f}_1^{(n_V)}, \ldots, \hat{f}_p^{(n_V)}\}$ whereas $\bar{f}^{(n)}_{VCV}$ takes its values in $\{\hat{f}_1^{(n)}, \ldots, \hat{f}_p^{(n)}\}$. Therefore, under some extra "regularity" or "stability" assumptions on the basis statistics $\hat{f}_1, \ldots, \hat{f}_p$ saying that for every j, $\hat{f}_j^{(n)}$ has a smaller risk as n increases (cf., for instance, the "stability" assumption in [7]) the VCV procedure should outperform our mCV procedure. Nevertheless, we will not explore this kind of regularity assumption (even though, it may be reasonable to think that oracle inequalities for the classical VCV procedure may hold under the stability assumption of [7]) and will require only weak assumptions on the estimators $\hat{f}_1, \ldots, \hat{f}_p$. Under these weak assumptions, the mCV (as well as the amCV) will, in fact, outperform the classical VCV and CV procedures, in the sense that the mCV satisfies oracle inequalities (cf. Theorem 2.4 below) in some setup where the VCV and CV procedures do not (cf. Example 2.8 below).

2.3. Assumptions

A significant part of our analysis is based on concentration properties of sums of random variables that belong to an Orlicz space. These spaces appear to

be useful for the non-bounded setup we have in mind. We say that a function $\psi : \mathbb{R}_+ \longmapsto \mathbb{R}$ is a *Young function* (cf. [35]) when it is convex, non-decreasing, $\psi(0) = 0$ and $\psi(\infty) = \infty$. Each Young function gives rise to a norm on a suitable class of random variables as follows:

Definition 2.1. For a Young function ψ and a real-valued random variable f, the *ψ-norm* of f is $\|f\|_\psi = \inf\{C > 0 : \mathbb{E}\psi(|f|/C) \leq 1\}$. The Orlicz space associated with ψ is then the space of random variables with finite ψ-norm.

For instance, when we consider $\psi_\alpha = \exp(x^\alpha) - 1$ for $\alpha \geq 1$, the ψ_α-norm measures exponential tail behavior of a random variable. Indeed, one can show that for every $u \geq 0$, $\mathbb{P}(|f| > u) \leq 2\exp(-cu^\alpha/\|f\|_{\psi_\alpha}^\alpha)$, where c is an absolute constant independent of f (see, for example, [35]). Note that the Orlicz space associated with the Young function $\psi(x) = x^p$ is the classical L_p space.

We shall use the following assumptions on the tail behavior and the "margin" (cf. [24] and [32]) of the excess loss function of an estimator $\hat{f}$.

(A) *There exist $\kappa \geq 1$ and $K_0, K_1 > 0$ such that the following holds. For any $m \in \mathbb{N}$ and any data set $D^{(m)} = \{Z_1, \ldots, Z_m\}$*

1. $\|Q(\cdot, \hat{f}^{(m)}(D^{(m)})) - Q(\cdot, f^*)\|_{L_{\psi_1}(\pi)} \leq K_0$
2. $\|Q(\cdot, \hat{f}^{(m)}(D^{(m)})) - Q(\cdot, f^*)\|_{L_2(\pi)} \leq K_1\big(R(\hat{f}^{(m)}(D^{(m)})) - R(f^*)\big)^{1/2\kappa}$.

The first point allows us to handle unbounded loss functions and unbounded estimators. This is a crucial point when one wants to consider the regression problem with unbounded noise or when one wants to aggregate unbounded estimators.

The second point is the classical "margin assumption" (cf. [24]). This means that the L_2-diameter of the set of almost oracles is controlled by their excess risks. The idea behind this assumption is for empirical risk minimization based procedures, the L_2-diameter of the set of almost minimizers of the empirical risk will be small with high probability. This leads to a smaller complexity of the set within we are looking for the oracle. Moreover, a side effect of this kind of assumption is that the concentration of the empirical risk around the risk is improved. The margin condition is linked to the convexity of the underlying loss Q. In density and regression estimation it is naturally satisfied with the best margin parameter ($\kappa = 1$), but for non-convex losses (for instance in classification), this assumption does not hold naturally (cf. [18] for a discussion on the margin assumption and for examples of such losses).

2.4. Oracle inequalities for the modified CV procedures (mCV) and its average version (amCV)

In this section, we shall not yet introduce any conditions on how a candidate statistic $\hat{f} = \big(\hat{f}^{(n)}\big)_n$ in $\{\hat{f}_1, \ldots, \hat{f}_p\}$ behaves when its training sample size changes, i.e. about the relationship of $\hat{f}^{(m)}$ and $\hat{f}^{(n)}$ for $m \neq n$. As the usual application of cross-validation involves retraining the selected model using *all*

the available data to obtain a final estimator, such assumptions are crucial for avoiding such pathological "counter-examples" as that found in Example 2.8 below. As we shall only introduce such conditions in Section 2.5, we will first study a simpler case – the case where even after the selection (or validation) step, we still only use the estimators $\hat{f}_j^{(n_V)}, j = 1, \dots, p$ over training samples of size n_V to build the final estimator. The case where we retrain on all available data will then be handled in Section 2.5 in some very specific setting.

We will require some simple (fixed sample size) properties on the estimators $\hat{f}_1, \dots, \hat{f}_p$ to obtain an oracle inequality for the modified CV procedure.

Definition 2.2. We say that a statistic $\hat{f} = (\hat{f}^{(n)})_n$ is *exchangeable* when for any integer n, for any permutation $\phi : \{1, \dots, n\} \longmapsto \{1, \dots, n\}$ for any $\pi^{\otimes n}$-almost vector $(z_1, \dots, z_n) \in \mathcal{Z}^n$, we have $\hat{f}^{(n)}(z_1, \dots, z_n) = \hat{f}^{(n)}(z_{\phi(1)}, \dots, z_{\phi(n)})$.

Remark that most of the statistics in the batch setup (the setup of this paper) satisfy this property. On the other side, statistics coming from the on-line setup are likely to be un-exchangeable.

The following lemma shows that in the two setups considered in this work, supremum bounds on the "shifted" empirical process for the "trained" estimates $\hat{f}_j^{(n_V)}(D^{(n_V)})$ are sufficient for deriving oracle inequalities for the corresponding amCV and mCV procedures:

Lemma 2.3. *We have two different setups, depending on the procedure that we want to study. Assume that one of the two following conditions holds:*

1. *The risk function $f \longmapsto R(f)$ is convex, and our estimator $\bar{f}^{(n)} = \hat{f}_{amCV}^{(n)}$ is the averaged version of the modified CV procedure (cf. (2.9)), with N_C arbitrary deterministic splits of n pieces of data into n_V pieces of training and n_C pieces of test data.*
2. *The statistics $\hat{f}_1, \dots, \hat{f}_p$ are exchangeable and our estimator $\bar{f}^{(n)} = \bar{f}_{mCV}^{(n)}$ is the modified CV procedure defined in Equation (2.8) using the splits of the data defined in Equation (2.1) and (2.2).*

Then for any constant $a \geq 0$, the following inequality holds:

$$\begin{aligned}
&\mathbb{E}_{D^{(n)}}\left(R(\bar{f}^{(n)}(D^{(n)})) - R(f^*)\right) \\
\leq &(1+a)\min_{j=1,\dots,p}\left[\mathbb{E}_{D^{(n_V)}}R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)\right] \\
&+ \mathbb{E}_{D^{(n)}}\max_{j=1,\dots,p}\left[(P - (1+a)P_{n_C})\left(Q(\cdot, \hat{f}_j^{(n_V)}(D^{(n_V)})) - Q(\cdot, f^*)\right)\right],
\end{aligned}$$

where $P_{n_C} = n_C^{-1}\sum_{i=n_V+1}^{n}\delta_{Z_i}$ is the empirical probability measure on $\{Z_{n_V+1}, \dots, Z_n\}$.

Now combining Lemma 2.3 and the maximal inequality of Lemma 5.3 below for the shifted empirical process appearing in Lemma 2.3, we are in a position to obtain the following oracle inequality for the amCV and the mCV procedures.

Theorem 2.4. *Let $\hat{f}_1, \dots, \hat{f}_p$ be p statistics satisfying Assumption (A). We have two different setups depending on the procedure that we want to study. Assume that one of the two conditions holds:*

1. *The risk function $f \longmapsto R(f)$ is convex and our estimator is the amCV procedure $\bar{f}^{(n)} = \hat{f}^{(n)}_{amCV}$ introduced in (2.9).*
2. *The statistics $\hat{f}_1, \dots, \hat{f}_p$ are exchangeable and our procedure is the modified CV procedure $\bar{f}^{(n)} = \bar{f}^{(n)}_{mCV}$ introduced in (2.8).*

Then for any $a > 0$, there exists a constant $c = c(a, \kappa)$ such that

$$\begin{aligned}\mathbb{E}_{D^{(n)}}\Big(R(\bar{f}^{(n)}(D^{(n)})) - R(f^*)\Big) \\ \leq (1+a)\min_{j=1,\dots,p}\Big[\mathbb{E}_{D^{(n_V)}}R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)\Big] \\ + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \vee \Big(\frac{\log n_C \log p}{n_C}\Big).\end{aligned}$$

Before stating some similar results for the cross-validation method, we would like to make two remarks on Theorem 2.4.

Remark 2.5. Note that Theorem 2.4 (and Theorem 3.5 below in the continuous case) provides oracle inequalities which holds in expectation. We believe that the techniques used in the present work to obtain these results do not allow to obtain similar deviation results (i.e. oracle inequalities that hold with high probability).

Remark 2.6. The usual question when considering resampling procedures is about the choice of V or here of (n_V, n_C) – the size of the training and learning/validation/test samples. It follows from Theorem 2.4 that an "ideal" or "oracle" choice of (n_V, n_C) may follow from optimizing the right-hand side of the oracle inequality of Theorem 2.4. That is by equalizing the "bias term"

$$(1+a)\min_{j=1,\dots,p}\Big[\mathbb{E}_{D^{(n_V)}}R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)\Big]$$

and the "variance term"

$$c(a,\kappa)\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \vee \Big(\frac{\log n_C \log p}{n_C}\Big).$$

Of course this method is meaningful only if the bound in the oracle inequality of Theorem 2.4 is good enough (that is, if this bound describes in a "good enough" way the behavior of $\mathbb{E}\big(R(\bar{f}^{(n)}(D^{(n)})) - R(f^*)\big)$ in terms of n_V and n_C). Moreover, this approach is only "ideal" in the sense that we don't know the value of the excess risks $\mathbb{E}\big(R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)\big), j = 1, \dots, p$, thus such a method cannot be performed from the data only. Somehow, to overcome this problem, we have to perform this bias/variance terms equilibrium in an empirical way in the same spirit as Lepskii's method for adaptation to be able to derive some optimal choice for (n_V, n_C).

2.5. Oracle inequalities for cross-validation itself

In Part 1 of Theorem 2.4, we make the assumption that the risk $R(\cdot)$ is convex – for which e.g. the conditional convexity of the contrast function $Q(z,f)$, for all z, would suffice, and thereafter in Part 2 we assume that our candidate statistics are exchangeable. To derive a result for a CV estimator retrained on the full data $D^{(n)}$ (instead of the only data $D^{(n_V)}$ like in (2.8) and (2.9)), we shall combine and strengthen these two assumptions.

Regard the mCV procedure $\bar{f}^{(n)}_{mCV}(D^{(n)}) = \hat{f}^{(n_V)}_{\hat{j}(D^{(n)})}(D^{(n_V)})$, whose final estimator is retrained on the first n_V pieces of data. For symmetry reasons, Part 2 of Theorem 2.4 remains true for any $k = 1, \dots, V$, if we replace $\bar{f}^{(n)}_{mCV}(D^{(n)})$ by $\bar{f}^{(n)}_{mCV,k}(D^{(n)}) = \hat{f}^{(n_V)}_{\hat{j}(D^{(n)})}(D_k)$ using the training set D_k from the $k-$th split.

Now assume that $\mathcal{Z} = \mathbb{R}$ and the statistics $\hat{f}_1, \dots, \hat{f}_p$ can all be written as functionals on the cumulative distribution function of the data, i.e. that there exist functionals $G_1, \dots, G_p$ such that

$$\hat{f}^{(m)}_j(D^{(m)}) = G_j(F_{D^{(m)}}),\ j = 1, \dots, p, m \in \mathbb{N}, \tag{2.10}$$

where $F_{D^{(m)}}(z) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{Z_i \le z\}, \forall z \in \mathbb{R}$. (This assumption automatically implies the exchangeability of the statistics. In particular, all M-estimators, such as the mean or median, have such a functional form.) Obviously $F_{D^{(n)}} = V^{-1}\sum_{k=1}^{V} F_{D_k}$. Thus if the risk $R(\cdot)$ is convex, and all the compositions $R \circ G_j$ too, then we can combine the upper bounds for the estimators $\bar{f}^{(n)}_{mCV,k}(D^{(n)})$ obtained in Part 2 of Theorem 2.4 to derive a bound for the VCV procedure (2.4) as follows:

$$\begin{aligned} R\Big(\bar{f}^{(n)}_{VCV}(D^{(n)})\Big) &= R\left(G_{\hat{j}(D^{(n)})}\left(F_{D^{(n)}}\right)\right) = R\left(G_{\hat{j}(D^{(n)})}\left(\frac{1}{V}\sum_{k=1}^{V} F_{D_k}\right)\right) \\ &\le \frac{1}{V}\sum_{k=1}^{V} R\left(G_{\hat{j}(D^{(n)})}\left(F_{D_k}\right)\right) = \frac{1}{V}\sum_{k=1}^{V} R\left(\bar{f}^{(n)}_{mCV,k}(D^{(n)})\right), \end{aligned}$$

and thus it easily follows from Part 2 of Theorem 2.4 the result:

Theorem 2.7. *Let $\hat{f}_1, \dots, \hat{f}_p$ be p statistics that can be written as functionals $G_1, \dots, G_p$ as in (2.10) and which satisfy Assumption (A), and assume that all the compositions $R \circ G_1, \dots, R \circ G_p$ are convex, as also is the risk function $R(\cdot)$. Then for the V-fold cross-validation procedure, we have the oracle inequality*

$$\begin{aligned} &\mathbb{E}_{D^{(n)}}\Big(R(\bar{f}^{(n)}_{VCV}(D^{(n)})) - R(f^*)\Big) \\ &\le (1+a)\min_{j=1,\dots,p}\Big[\mathbb{E}_{D^{(n_V)}} R(\hat{f}^{(n_V)}_j(D^{(n_V)})) - R(f^*)\Big] \\ &+ c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \vee \Big(\frac{\log n_C \log p}{n_C}\Big) \end{aligned}$$

where c is a constant depending only on a and κ.

Note. The "functional convexity condition" on the $R \circ G_j$ is a strong one, but can be relaxed – it suffices for it to hold up to a summand that converges to zero no slower than the residual term in Theorem 2.4, and versions of it averaged over the training data may also suffice. In most practical cases, the only straightforward way of showing the convexity of the $R \circ G_j$ (with high certainty) is by simulation. In the standard example of least-squares regression with underlying Gaussian linear model, for instance, $R \circ G_j$ is convex for the fixed-design setup, regardless of other parameters, but for the random-design setup we need additional conditions such as a reasonable signal-to-noise ratio or large enough sample size (indicating that such a convexity condition does in fact hold *up to* a quickly-decaying extra summand). Simulations of a straightforward sparse Lasso example with 100-dimensional Gaussian covariates and Gaussian noise have shown that the neccessary functional convexity condition for 10-fold cross-validation holds from a sample size of $n = 40$ and a signal-to-noise ratio of 2.0 upwards, for a range of penalty tuning parameters. However, discussing this issue at length is beyond the scope of this paper.

The reason why we need extra assumptions such as the functional form of the candidate statistics is that the computation of the index $\hat{\jmath}(D^{(n)})$ only involves the performances of the estimators for n_V observations ($R_{n_C}(\hat{f})$ depends only on $\hat{f}^{(n_V)}$). Without extra assumptions, it is thus easy to contrive counter-examples for which $\hat{f}^{(n_V)}$ performs well and $\hat{f}^{(n)}$ performs badly:

Example 2.8. Fix an integer V and a sample size $n > 1$ that is a multiple of V. We will construct a set $\mathcal{F}_n = \{\hat{f}_1, \hat{f}_2\}$ of two estimators (which are functionals of the training data) for which V-fold cross-validation does *not* satisfy the oracle inequality from Theorem 2.7.

We consider the classification problem with 0−1 loss $Q(Z, f) = Q((X, Y), f) = \mathbb{1}_{f(X) \neq Y}$. Assume that $Y \equiv 1$ a.s. and X is uniformly distributed on $[0, 1]$. The Bayes rule is thus given by $f^*(x) = \mathbb{P}(Y = 1 | X = x) = 1, \forall x \in [0, 1]$. We define statistics $\hat{f}_1 = (\hat{f}_1^{(n)})_n$ and $\hat{f}_2 = (\hat{f}_2^{(n)})_n$ by

$$\hat{f}_1^{(p)} \equiv \begin{cases} 0 & \text{if} \quad 1 \leq p \leq n-1 \\ 1 & \text{if} \quad p \geq n \end{cases} \quad \text{and} \quad \hat{f}_2^{(p)} \equiv \begin{cases} 1 & \text{if} \quad 1 \leq p \leq n-1 \\ 0 & \text{if} \quad p \geq n \end{cases} .$$

It is easy to see that $\hat{\jmath}(D^{(n)}) = \arg\min_{j \in \{1,2\}} R_{n,V}(\hat{f}_j)$ is always equal to 2. Thus the V-fold CV procedure is $\bar{f}_{VCV}^{(n)}(D^{(n)}) = \hat{f}_{\hat{\jmath}(D^{(n)})}^{(n)}(D^{(n)}) = \hat{f}_2^{(n)}(D^{(n)})$. Set $\mathcal{F}_n = \{\hat{f}_1, \hat{f}_2\}$. For any $1 \leq p \leq n$, it is easy to check that

$$\min_{\hat{f} \in \mathcal{F}_n} \mathbb{E}_{D^{(n)}}[R(\hat{f}^{(p)}(D^{(p)})) - R^*] = 0 \text{ and } \mathbb{E}_{D^{(n)}}[R(f_{VCV}^{(n)}(D^{(n)})) - R^*] = 1.$$

As we can do this for arbitrarily high sample sizes n, V-fold cross-validation is not even risk-consistent at this level of generality – and certainly does not satisfy any meaningful oracle inequality.

2.6. Aggregation with multiple splits

Let a dictionary $\mathcal{F} = \{f_1, \dots, f_p\}$ be given and assume that $\tilde{f} = (\tilde{f}^{(n)})_n$ is an aggregation method satisfying the following oracle inequality under a ψ_1 assumption on the excess loss functions and a margin assumption with margin parameter $\kappa \geq 1$ (like in Assumption (A)):

$$\mathbb{E}[R(\tilde{f}^{(n)}(D^{(n)})) - R^*] \leq K_{agg} \min_{f \in \mathcal{F}} [R(f) - R^*] + c\Big(\frac{\log p}{n}\Big)^{\frac{\kappa}{2\kappa-1}} \tag{2.11}$$

where $K_{agg} \geq 1$ is the leading constant. For instance, both the empirical risk minimization algorithm

$$\tilde{f}^{(n)}_{ERM}(D^{(n)}) \in \operatorname{Arg\,min}_{f \in \mathcal{F}} R_n(f), \text{ where } R_n(f) = \frac{1}{n}\sum_{i=1}^{n} Q(Z_i, f)$$

and the aggregate with exponential weights and temperature parameter $T > 0$,

$$\tilde{f}^{(n)}_{AEW}(D^{(n)}) = \sum_{j=1}^{p} w_j^{(n)} f_j, \text{ where } w_j^{(n)} = \frac{\exp\big(-nR_n(f_j)/T\big)}{\sum_{k=1}^{p} \exp\big(-nR_n(f_k)/T\big)}, \tag{2.12}$$

satisfy an oracle inequality of the form (2.11) (cf. [19]).

Let $\hat{f}_1, \dots, \hat{f}_p$ be p statistics. Assume that a fixed value n_C be given (the size of test sets), and define $n_V = n - n_C$. Let $C_1, \dots, C_{N_C}$ be subsets of $\{1, \dots, n\}$, each of size n_V. For any $1 \leq k \leq N_C$, we consider $\tilde{f}_k^{(n)}(D^{(n)})$ an aggregation procedure where the weights have been constructed on the data set $D_{(C_k)'}$ and for the dictionary $\mathcal{F}_k = \{\hat{f}_1^{(n_V)}(D_{(C_k)}), \dots, \hat{f}_p^{(n_V)}(D_{(C_k)})\}$, for instance, when the ERM aggregation procedure is chosen for the basic aggregation procedure,

$$\tilde{f}_k^{(n)}(D^{(n)}) \in \operatorname{Arg\,min}_{f \in \mathcal{F}_k} \frac{1}{n_C} \sum_{i \notin C_k} Q(Z_i, f).$$

Then we average all these aggregates over the N_C different splits of $D^{(n)}$, namely: $(D_{(C_k)}, D_{(C_k)'})_{1 \leq k \leq N_C}$. We define the *aggregate with multiple splits* $\bar{f}_{Agg} := (\bar{f}^{(n)}_{Agg})_n$ by

$$\bar{f}^{(n)}_{Agg}(D^{(n)}) = \frac{1}{N_C} \sum_{k=1}^{N_C} \tilde{f}_k^{(n)}(D^{(n)}) \ . \tag{2.13}$$

Theorem 2.9. *Let $\hat{f}_1, \dots, \hat{f}_p$ be p statistics satisfying Assumption (A). Assume that the risk function $f \longmapsto R(f)$ is convex. Consider an aggregation procedure satisfying (2.11). The aggregate with multiple splits (defined in (2.13)) associated with this aggregation procedure and the p statistics $\hat{f}_1, \dots, \hat{f}_p$ satisfies the inequality*

$$\mathbb{E}_{D^{(n)}}\Big(R(\bar{f}^{(n)}_{Agg}(D^{(n)})) - R(f^*)\Big) \leq K_{agg} \min_{j=1,\dots,p} \Big[\mathbb{E}_{D^{(n_V)}} R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)\Big] + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \ .$$

Proof. By the convexity of the risk, we have

$$\begin{aligned}
\mathbb{E}[R(\bar{f}_{Agg}^{(n)}(D^{(n)})) - R(f^*)] &= \mathbb{E}\Big[R\Big(\frac{1}{N_C}\sum_{k=1}^{N_C}\tilde{f}_k^{(n)}(D^{(n)})\Big) - R(f^*)\Big] \\
&\leq \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}\Big[R(\tilde{f}_k^{(n)}(D^{(n)})) - R(f^*)\Big] \\
&\leq \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}_{C_k}\mathbb{E}_{C'_k}\Big[R\Big(\tilde{f}_k^{(n)}(D^{(n)})\Big) - R\left(f^*\right)\Big] \\
&\leq \frac{1}{N_C}\sum_{k=1}^{N_C}\mathbb{E}_{C_k}\Big[K_{agg}\min_{j=1,\dots,p}[R(\hat{f}_j^{(n_V)}(D_{(C_k)})) - R(f^*)] + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}}\Big] \\
&\leq K_{agg}\min_{1\leq j\leq p}\mathbb{E}_{D^{(n_V)}}(R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*)) + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}}.
\end{aligned}$$

□

Note that when we chose an optimal aggregation procedure (cf. the progressive mixture of [9] or [37]) for the basic aggregation procedure, we can take $K_{agg} = 1$. Such oracle inequalities with leading constant $K_{agg} = 1$ cannot be achieved by "selection procedure" that is by procedures taking their values in the dictionary itself and note in its convex hull (cf. [18]).

3. Continuous case

We consider Λ a set of indexes and $F = \{\hat{f}_\lambda : \lambda \in \Lambda\}$ a set of statistics indexed by Λ. In the previous part of this paper, we have explored the case $\Lambda = \{1, \dots, p\}$. In this section, we need not assume Λ to be finite.

We consider the notation introduced in Section 2, and define the continuous version of the modified CV procedure by

$$\bar{f}_{mCV}^{(n)}(D^{(n)}) = \hat{f}_{\hat{\lambda}(D^{(n)})}^{(n_V)}(D^{(n_V)}) \text{ where } \hat{\lambda}(D^{(n)}) \in \operatorname{Arg\,min}_{\lambda\in\Lambda} R_{n,V}(\hat{f}_\lambda) \tag{3.1}$$

and the continuous version of the averaged version of the modified CV procedure by

$$\hat{f}_{amCV}^{(n)}(D^{(n)}) = \frac{1}{N_C}\sum_{k=1}^{N_C}\hat{f}_{\hat{\lambda}(D^{(n)})}^{(n_V)}(D_{(C_k)}) \text{ where } \hat{\lambda}(D^{(n)}) \in \operatorname{Arg\,min}_{\lambda\in\Lambda} R_{n_C}(\hat{f}_\lambda). \tag{3.2}$$

Remark that we assume that the infima of $\lambda \longmapsto R_{n,V}(\hat{f}_\lambda)$ and $\lambda \longmapsto R_{n_C}(\hat{f}_\lambda)$ are achieved. We also called these two infima by the same name but there will be no ambiguity since we will use them in two clearly separated setups.

Following the line of Lemma 2.3, it is easy to obtain the following result.

Lemma 3.1. *We have two different setups, depending on the procedure that we want to study:*

1. *If the risk function $f \longmapsto R(f)$ is convex, then the averaged version of the modified CV (cf. (3.2)) with N_C arbitrary deterministic splits of n pieces of data into n_V pieces of training and n_C pieces of test data satisfies the following oracle inequality with $\bar{f}^{(n)} = \hat{f}^{(n)}_{amCV}$;*
2. *If the statistics $\hat{f}_\lambda, \lambda \in \Lambda$ are exchangeable, then the modified V-fold CV procedure defined in Equation (3.1) for the splits of the data defined in Equation (2.1) and (2.2) with $1 \leq V \leq n$ satisfies the following oracle inequality with $\bar{f}^{(n)} = \bar{f}^{(n)}_{mCV}$;*

for any constant $a \geq 0$, we have the following inequality

$$\mathbb{E}_{D^{(n)}}\Big(R(\bar{f}^{(n)}(D^{(n)})) - R(f^*)\Big) \leq (1+a)\inf_{\lambda\in\Lambda}\Big[\mathbb{E}_{D^{(n_V)}}R(\hat{f}^{(n_V)}_\lambda(D^{(n_V)})) - R(f^*)\Big] + \mathbb{E}_{D^{(n)}}\sup_{\lambda\in\Lambda}\Big[(P-(1+a)P_{n_C})\Big(Q(\cdot,\hat{f}^{(n_V)}_\lambda(D^{(n_V)})) - Q(\cdot,f^*)\Big)\Big],$$

where $P_{n_C} = (1/n_C)\sum_{i=n_V+1}^{n}\delta_{Z_i}$.

To control the expectation of the supremum of the "shifted" empirical process appearing in Lemma 3.1, we need some results from empirical process theory (the proof is provided in Section 5).

Lemma 3.2. *Let $a > 0$ and $\mathcal{Q} := \{Q_\lambda : \lambda \in \Lambda\}$ be a set of measurable functions defined on $(\mathcal{Z}, \mathcal{T})$. Let $Z, Z_1, \ldots, Z_m$ be i.i.d. random variables with values in $(\mathcal{Z}, \mathcal{T})$ such that $\forall Q \in \mathcal{Q}, \mathbb{E}Q(Z) \geq 0$. Suppose that there exists some constants $c, L, \epsilon_{min} > 0$ such that for all $\epsilon \geq \epsilon_{min}$ and all $u \geq 1$, with probability greater than $1 - L\exp(-cu)$*

$$\sup_{Q\in\mathcal{Q}:PQ\leq\epsilon}((P-P_m)Q)_+ \leq \frac{uJ(\epsilon)}{\sqrt{m}}, \tag{3.3}$$

where J is a strictly increasing function such that J^{-1} is strictly convex. Let ψ be the convex conjugate of J^{-1} defined by $\psi(u) = \sup_{v>0}(uv - J^{-1}(v)), \forall u > 0$. Assume that for some $r \geq 1$, $x > 0 \longmapsto \psi(x)/x^r$ decreases and define for $q > 1$ and $u \geq 1$,

$$\epsilon_q(u) = \psi\Big(\frac{2q^{r+1}(1+a)u}{a\sqrt{m}}\Big) \vee \epsilon_{min}.$$

Then, there exists a constant L_1 (depending only on L) such that for every $u \geq 1$, with probability greater than $1 - L_1\exp(-cu)$

$$\sup_{Q\in\mathcal{Q}}\Big((P-(1+a)P_m)Q\Big)_+ \leq \frac{a\epsilon_q(u/q)}{q}.$$

Moreover, assume that ψ increases such that $\psi(\infty) = \infty$, then there exists a constant c_1 depending only on L and c such that

$$\mathbb{E}\sup_{Q\in\mathcal{Q}}\Big((P-(1+a)P_m)Q\Big)_+ \leq \frac{ac_1\epsilon_q(1/q)}{q}.$$

The function $\epsilon \longmapsto \sup_{Q\in\mathcal{Q}:PQ\leq\epsilon}(P-P_m)Q$, appearing in Equation (3.3), is a classical measure of the complexity of the set of functions $\mathcal{Q}$ (cf. for instance [33, 5, 16] and references therein). A common way to upper bound this function with high probability is to use some metric complexity measure like the Dudley entropy integral (cf. [12] or, for instance, [35]) coming out of the chaining argument. In this paper, we use the γ function of Talagrand (cf. [31]) as a metric complexity measure of $\mathcal{Q}$. We recall here the definition.

Let (T,d) be a metric space. An *admissible sequence* of T is a collection $\{T_s : s\in\mathbb{N}\}$ of subsets of T, such that $|T_0|=1$ and $|T_s|\leq 2^{2^s}, \forall s\geq 1$.

Definition 3.3 ([31])**.** For a metric space (T,d) and $\alpha\geq 0$ define

$$\gamma_\alpha(T,d)=\inf_{(T_s)}\sup_{t\in T}\sum_{s=0}^{\infty}2^{s/\alpha}d(t,T_s),$$

where the infimum is taken over all admissible sequences (T_s) of T.

The generic chaining mechanism can be used to show (cf. theorem 1.2.7 in [31]) that if $\{X_t : t\in T\}$ (where T is a set provided with two distances d_1 and d_2) is such that $\mathbb{E}X_t=0$ and

$$\mathbb{P}\Big(|X_s-X_t|\geq u\Big)\leq 2\exp\Big(-\min\Big(\frac{u^2}{d_2(s,t)^2},\frac{u}{d_1(s,t)}\Big)\Big), \forall s,t\in T, u>0$$

then, there exists some absolute constant $L,c>0$ such that for all $u\geq 1$,

$$\sup_{s,t\in T}|X_s-X_t|\leq uL(\gamma_1(T,d_1)+\gamma_2(T,d_2)) \tag{3.4}$$

with probability at least $1-L\exp(-cu)$.

Note that one potential (yet, usually suboptimal) choice for the sets T_s that constitutes an admissible sequence (T_s) in Definition 3.3 are ϵ_s-covers of T, where each ϵ_s is such that the entropy number $N(T,\epsilon_s,d)$ is less than 2^{2^s}. Then, for this choice, an easy computation (cf. [31]) shows that

$$\gamma_\alpha(T,d)\leq c\int_0^{\mathrm{Diam}(T,d)}(\log N(T,\epsilon,d))^{1/\alpha}d\epsilon. \tag{3.5}$$

Lemma 3.4. *Let $\mathcal{Q}:=\{Q_\lambda : \lambda\in\Lambda\}$ be a set of measurable functions defined on $(\mathcal{Z},\mathcal{T})$. Let $Z,Z_1,\ldots,Z_m$ be i.i.d. random variables with values in $(\mathcal{Z},\mathcal{T})$. Grant that there exists $C_1>0$ and an increasing function $G(\cdot)$ such that*

$$\forall Q\in\mathcal{Q}, \|Q(Z)\|_{L_2}\leq G(\mathbb{E}Q(Z)) \text{ and } \sup_{q\in\mathcal{Q}}\|Q(Z)\|_{L_{\psi_1}}\leq C_1.$$

Then, there exists some absolute constant $L,c>0$ such that for all $\epsilon>0$ and for all $u\geq 1$, with probability at least $1-L\exp(-cu)$,

$$\sup_{Q\in\mathcal{Q}:PQ\leq\epsilon}((P-P_m)Q)_+\leq uL\Big(\frac{(\log m)\gamma_1(\mathcal{Q}_\epsilon^{L_2},\|\cdot\|_{\psi_1})}{m}+\frac{\gamma_2(\mathcal{Q}_\epsilon^{L_2},\|\cdot\|_{L_2})}{\sqrt{m}}\Big),$$

where $\mathcal{Q}_\epsilon^{L_2}=\{Q\in\mathcal{Q} : \|Q(Z)\|_{L_2}\leq G(\epsilon)\}$.

The proof of Lemma 3.4 is provided in Section 5. Now, combining Lemma 3.2 and Lemma 3.4, we obtain a continuous version of Theorem 2.4.

Theorem 3.5. *Let Λ a set of indexes and $F = \{\hat{f}_\lambda : \lambda \in \Lambda\}$ a set of statistics indexed by Λ. Fix $n_V \leq n$ the size of the validation sample and define the set of excess loss functions associated with F by $\mathcal{Q} = \{Q(\cdot, \hat{f}_\lambda^{(n_V)}(D^{(n_V)})) - Q(\cdot, f^*) : \lambda \in \Lambda\}$. We assume that the tail behavior of the statistics in F and the complexity of F satisfy the following assumptions:*

Any statistic $\hat{f}$ in F satisfies (A) and there exist ϵ_{min} and a strictly increasing function J such that J^{-1} is strictly convex, the convex conjugate ψ of J^{-1} increases, $\psi(\infty) = \infty$ and there exists $r \geq 1$ such that $x \mapsto \psi(x)/x^r$ decreases and

$$J(\epsilon) \geq \gamma_2(\mathcal{Q}_\epsilon^{L_2}, \|\cdot\|_{L_2}) + \frac{(\log n_C)\gamma_1(\mathcal{Q}_\epsilon^{L_2}, \|\cdot\|_{\psi_1})}{\sqrt{n_C}}, \forall \epsilon > \epsilon_{min}$$

where $\mathcal{Q}_\epsilon^{L_2} = \{Q \in \mathcal{Q} : \|Q(Z)\|_{L_2} \leq \epsilon^{1/2\kappa}\}$.

We consider two different setups depending on the procedure we want to study. Assume that one of the two condition holds:

1. *The risk function $f \longmapsto R(f)$ is convex and our procedure is the amCV procedure $\bar{f}^{(n)} = \hat{f}_{amCV}^{(n)}$ defined in (3.2).*
2. *The statistics $\hat{f}_1, \ldots, \hat{f}_p$ are exchangeable and our procedure is the mCV procedure $\bar{f}^{(n)} = \bar{f}_{mCV}^{(n)}$ introduced in (3.1).*

Then, for every $a > 0$ and $q > 1$, the following inequality holds

$$\mathbb{E}_{D^{(n)}}\left(R(\bar{f}^{(n)}(D^{(n)})) - R(f)\right) \leq (1+a)\inf_{\lambda\in\Lambda}\left[\mathbb{E}_{D^{(n_V)}}R(\hat{f}_\lambda^{(n_V)}(D^{(n_V)})) - R(f^*)\right] + \frac{ac\epsilon_q(1/q)}{q},$$

where we set $\epsilon_q(u) = \psi\big(\frac{2q^{r+1}(1+a)u}{a\sqrt{n_C}}\big) \vee \epsilon_{min}, \forall u > 0$.

Note that Theorem 3.5 generalizes Theorem 2.4 to a continuous family of estimators. Indeed, it is easy to verify that, in the finite case $|\Lambda| = p$, we obtain the same result as in Theorem 2.4. For instance, under the assumptions of Theorem 2.4 by using Equation 3.5, we have, for any $\epsilon > 0$,

$$\frac{(\log n_C)\gamma_1(\mathcal{Q}_\epsilon^{L_2}, \|\cdot\|_{\psi_1})}{\sqrt{n_C}} + \gamma_2(\mathcal{Q}_\epsilon^{L_2}, \|\cdot\|_{L_2}) \leq \frac{K_0(\log n_C)\log p}{\sqrt{n_C}} + \epsilon^{1/2\kappa}\sqrt{\log p} := J(\epsilon);$$

thus, the convex conjugate of J^{-1} is

$$\psi(v) = \frac{K_0(\log n_C)\log p}{\sqrt{n_C}}v + c_\kappa\big(v\sqrt{\log p}\big)^{\frac{2\kappa}{2\kappa-1}}, \forall v > 0.$$

Thus, $\epsilon_q(1/q)$ is, up to some constant depending only on K_0 and κ, of the same order as the residue of the oracle inequality of Theorem 2.4. Furthermore, the

same reasoning used for Theorem 2.7 can also be applied here in sufficiently convex setups where the full data set is used for retraining. Nevertheless, from a technical point of view, there is a major difference between the finite and the continuous cases. In the finite case, it is only a side effect of the margin assumption (cf. second point of Assumption (A)) that is actually used, namely a better concentration of the empirical risk to the actual risk. Whereas in the continuous case, all the strength of the margin assumption is used: a reduction of the L_2 diameter of the set of potential almost oracle. This control on the diameter can be easily seen in the Dudley's entropy integral, where this diameter appears in the upper bound of integration.

4. Applications

In this section, we will be interested in two procedures which initially are non-adaptive to one unknown parameter of the model or to one parameter for which we have no canonical choice: First, the Lasso procedure where theoretical results have been obtained under the assumption that the variance of the noise is known (we will provide a procedure with a data dependent regularization parameter). Second, aggregation with exponential weights, which depends on a temperature parameter. We could just as well have applied this adaptation procedure to other problems, like the choice of the regularization parameter for penalized empirical risk minimization, or the choice of the threshold constant in wavelet methods.

4.1. Adaptive choice of the regularization parameter for the Lasso

We consider the linear regression model $Y = \langle X, \beta^* \rangle + \sigma\epsilon$, where $Y \in \mathbb{R}$ is a random vector, $X \in \mathbb{R}^p$ is a random vector and $\epsilon \in \mathbb{R}$ is a random variable (the noise) independent of X such that $\mathbb{E}\epsilon = 0$ and $\mathbb{E}\epsilon^2 = 1$. We have n i.i.d. observations in this model, and the total dataset consists of $\mathbf{Y} = (Y_1, \dots, Y_n)^t$ and $\mathbf{X} = (X_1, \dots, X_n)^t$. We consider the function $\Phi : \mathbb{R}^p \times \mathbb{R}^+ \longmapsto \mathbb{R}$ defined by

$$\Phi(\beta, \lambda) = |\mathbf{Y} - \mathbf{X}\beta|_n^2 + \lambda|\beta|_1.$$

Given a regularization parameter λ, the Lasso estimator $\hat{f}_\lambda$ is defined by

$$\hat{f}_\lambda^{(n)}(\cdot, D^{(n)}) = \langle \cdot, \hat{\beta}(\lambda)^{(n)}(D^{(n)}) \rangle \text{ where } \hat{\beta}(\lambda)^{(n)}(D^{(n)}) \in \operatorname{Arg}\min_{\beta \in \mathbb{R}^p} \Phi(\beta, \lambda).$$

We consider the regularization parameter λ to be normalized so as to lie in $[0, 1]$. Such a normalization is possible, since for $\lambda_{max} := 2 \max_i |\langle X_i, Y \rangle|$, the zero vector is a minimizer of $\Phi(\beta, \lambda_{max})$; that is, the Lasso penalty is always able to shrink the coefficient estimate for β down to zero.

Usually the choice of the regularization parameter λ is a problem. A possible solution is to use the adaptation procedures that were introduced in the previous sections. One can think of using the adaptation procedures introduced in Section 3 (in the continuous case) for the dictionary $\{\hat{f}_\lambda : \lambda \in [0, 1]\}$. The

continuous case will be illustrated in the next subsection, therefore, we focus our attention on the finite case in this subsection. We should consider two different types of dictionary:

1. the dictionary of estimators is a finite set $\{\hat{f}_\lambda : \lambda \in \mathcal{G}\}$ where $\mathcal{G}$ is a finite grid of $[0,1]$. This means somehow that the regularization parameters are given based upon some a priori knowledge.
2. the dictionary is a set $\{\hat{f}_1, \dots, \hat{f}_N\}$ of estimators that may have been constructed using the LARS algorithm or any other solution path procedures or other estimators in the regression model. In particular, this type of dictionary allows for data-dependent regularization parameters (like in the LARS algorithm). Note that in this setup, the number N of elements may also be random (like in the LARS case). Here, we assume that N is fixed. In the "LARS case", we know that the number N of solutions is always less than n therefore, in this case, one can "complete" the dictionary up to n by repeating the final LARS estimator - for instance - and therefore one can take $N = n$.

It appears that oracle inequalities follow directly from Theorem 2.4 and Theorem 2.9 for the second type of dictionary. For instance, the LARS algorithm provides a family of regularization parameters $0 = \lambda^{(0)} < \lambda^{(1)} < \ldots < \lambda^{(N)}$, (where $N \leq n$), and the corresponding Lasso estimators $\hat{f}_{\lambda^{(j)}}, j = 1, \dots, N$. Thus we believe that using the LARS algorithm combined with the mCV, amCV or Agg procedures for the dictionary $\{\hat{f}_{\lambda^{(0)}}, \dots, \hat{f}_{\lambda^{(N)}}\}$ will prove to be efficient in practice.

More theoretical details on the type of assumption and results can be provided in the first case. We now turn to the study of this first case. We construct the mCV procedure (cf. (2.8)) in this setup. Consider the family of splits $(B_1, D_1), \dots, (B_V, D_V)$ of $D^{(n)}$ defined in (2.1) and (2.2) for some $1 \leq V \leq n$ dividing n. For any Lasso estimator $\hat{f}_\lambda$ the r-V-fold CV empirical risk, for $r > 0$, is defined by

$$R^{(r)}_{n,V}(\hat{f}_\lambda) = \frac{1}{V}\sum_{k=1}^{V} \frac{1}{n_C} \sum_{i=(k-1)n_C+1}^{kn_C} |Y_i - \langle X_i, \hat{\beta}(\lambda)^{(n_V)}(D_k)\rangle|^r.$$

The mCV procedure is defined in this context by

$$\begin{aligned}\bar{f}^{(n)}_{mCV}(\cdot, D^{(n)}) &= \hat{f}^{(n_V)}_{\hat{\lambda}_r(D^{(n)})}(\cdot, D^{(n_V)}) = \langle \cdot, \hat{\beta}(\hat{\lambda}_r(D^{(n)}))^{(n_V)}(D^{(n_V)})\rangle \\ &:= \langle \cdot, \bar{\beta}^{(n)}_{mCV}(D^{(n)})\rangle\end{aligned}$$

where

$$\hat{\lambda}_r(D^{(n)}) \in \operatorname{Arg}\min_{\lambda\in\mathcal{G}} R^{(r)}_{n,V}(\hat{f}_\lambda).$$

Now, we construct the amCV (cf. (2.9)) and the *Agg* (cf. (2.13)) procedures using the subsets $C_1, \dots, C_{N_C}$ of $\{1, \dots, n\}$ each of size n_V: the mCV is defined,

in this context, by

$$\hat{f}_{amCV}(D^{(n)})(\cdot, D^{(n)}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \hat{f}^{(n_V)}_{\hat{\lambda}_r(D^{(n)})}(\cdot, D_{(C_k)})$$
$$= \big\langle \cdot, \frac{1}{N_C}\sum_{k=1}^{N_C} \hat{\beta}(\hat{\lambda}_r(D^{(n)}))^{(n_V)}(D_{(C_k)})\big\rangle := \big\langle \cdot, \hat{\beta}^{(n)}_{amCV}(D^{(n)})\big\rangle,$$

where $\hat{\lambda}_r(D^{(n)}) \in \operatorname{Arg\,min}_{\lambda\in\mathcal{G}} R^{(r)}_{n_C}(\hat{f}_\lambda)$ and $R^{(r)}_{n_C}$ is the r-CV risk. Finally the Agg procedure (with respect to the aggregate with exponential weights as a based aggregation procedure) is defined by

$$\tilde{f}^{(n)}_{Agg}(\cdot, D^{(n)}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \tilde{f}^{(n)}_k(\cdot, D^{(n)}) = \big\langle \cdot, \frac{1}{N_C}\sum_{k=1}^{N_C}\sum_{\lambda\in\mathcal{G}} w^{(n_C)}_\lambda(C'_k)\hat{\beta}(\lambda)^{(n_V)}(D_{(C_k)})\big\rangle$$
$$:= \big\langle \cdot, \tilde{\beta}^{(n)}_{Agg}(D^{(n)})\big\rangle$$

where

$$\tilde{f}^{(n)}_k(D^{(n)}) = \sum_{\lambda\in\mathcal{G}} w^{(n_C)}_\lambda(C'_k)\hat{f}^{(n_V)}_\lambda(D_{(C_k)})$$

and

$$w^{(n_C)}_\lambda(C'_k) := \frac{\exp\big(-T^{-1}\sum_{i\notin C_k}|Y_i - \langle X_i, \hat{\beta}(\lambda)^{(n_V)}(C_k)\rangle|^r\big)}{\sum_{\mu\in\mathcal{G}}\exp\big(-T^{-1}\sum_{i\notin C_k}|Y_i - \langle X_i, \hat{\beta}(\mu)^{(n_V)}(C_k)\rangle|^r\big)}$$

Note that for values of r close to 0, the Lasso vector $\bar{\beta}^{(n)}_{mCV}$ constructed with a data-driven choice of the regularization parameter $\hat{\lambda}_r(D^{(n)})$ is likely to enjoy some model selection (or sign consistency) properties. Nevertheless, from a theoretical point of view, we will obtain results only for the prediction problem with respect to the L_2-risk.

We would like to apply Theorem 2.4 and Theorem 2.9 to the three procedures that we have introduced here. To this end, we have to check assumption (A) for the elements of the dictionary $F := \{\hat{f}_\lambda : \lambda \in \mathcal{G}\}$ and so the design vector X has to enjoy some properties.

Definition 4.1. Let X be a random vector of $\mathbb{R}^p$ and denote by μ its probability distribution. We say that X is *log-concave* when for all nonempty measurable sets $A, B \subset \mathbb{R}^p$ and every $\alpha \in [0,1]$, $\mu(\alpha A + (1-\alpha)B) \geq \mu(A)^\alpha \mu(B)^{1-\alpha}$. We say that X is a ψ_2 *vector* when $\|X\|_{\psi_2} := \sup_{x\in\mathcal{S}^{p-1}} \big\|\langle X, x\rangle\big\|_{\psi_2} < \infty$.

Many natural measures are log-concave. Among the examples are measures that have a log-concave density, the volume measure of a convex body, and many others. A well known fact on a log-concave random vector X of $\mathbb{R}^p$ follows from Borell's inequality (cf. [26]): for every $x \in \mathbb{R}^p$, $\big\|\langle X, x\rangle\big\|_{\psi_1} \leq c\big\|\langle X, x\rangle\big\|_{L_1}$ where c is an absolute constant. In particular, the moments of linear functionals satisfy, for all $p \geq 1$, $\big\|\langle X, x\rangle\big\|_{L_p} \leq cp\big\|\langle X, x\rangle\big\|_{L_1}$.

In the following we assume that X is a ψ_2, log-concave vector and the noise ϵ is ψ_2.

Let $m \in \mathbb{N}$, $\lambda \in \mathcal{G}$, $\beta := \hat{\beta}(\lambda)^{(m)}(D^{(m)})$ be fixed for the moment, and let $\mathcal{L}_\beta(X,Y) = (Y - \langle X, \beta\rangle)^2 - (Y - \langle X, \beta^*\rangle)^2$ be the corresponding excess loss function. We need to bound the ψ_1-norm of $\mathcal{L}_\beta$ and to check the margin condition. For the second task, we use the log-concavity of X to obtain

$$\begin{aligned}\mathbb{E}\mathcal{L}_\beta(X,Y)^2 \leq& 2\mathbb{E}\langle X, \beta - \beta^*\rangle^4 + 8\sigma^2\mathbb{E}\langle X, \beta - \beta^*\rangle^2 \\ \leq& (c + 8\sigma^2)\mathbb{E}\langle X, \beta - \beta^*\rangle^2 = (c + 8\sigma^2)\mathbb{E}\mathcal{L}_\beta.\end{aligned}$$

This proves that the dictionary F satisfies the margin assumption with parameter $\kappa = 1$. For the first task, we use the fact that X is a ψ_2-vector to get

$$\begin{aligned}\|\mathcal{L}_\beta(X,Y)\|_{\psi_1} &= \left\|\langle X, \beta - \beta^*\rangle^2 + 2\sigma\epsilon\langle X, \beta^* - \beta\rangle\right\|_{\psi_1} \\ &\leq (1 + 2\sigma)\left\|\langle X, \beta - \beta^*\rangle\right\|_{\psi_2}^2 + 2\sigma\|\epsilon\|_{\psi_2}^2 \\ &\leq (1 + 2\sigma)\|X\|_{\psi_2}^2\|\beta - \beta^*\|_2^2 + 2\sigma\|\epsilon\|_{\psi_2}^2.\end{aligned}$$

Now for the construction of the dictionary, we threshold all the Lasso vectors $\hat{\beta}(\lambda), \lambda \in \mathcal{G}$ in such a way that the ℓ_2-norm of these vectors is smaller than a constant K_0'. Then the dictionary F satisfies Assumption (A) (with $K_0 := K_0' + \|\beta^*\|_2$). Note that under assumption (A), the aggregate with exponential weights satisfies the oracle inequality (2.11) (up to a $\log n$ factor when $\kappa = 1$, cf. [19]). Thus, we are now in position to apply Theorem 2.4 and Theorem 2.9.

Let $\hat{\beta}$ be either $\bar{\beta}_{mCV}^{(n)}(D^{(n)})$, $\hat{\beta}_{amCV}^{(n)}(D^{(n)})$ or $\tilde{\beta}_{Agg}^{(n)}(D^{(n)})$, we have

$$\begin{aligned}\mathbb{E}[(Y - \langle X, \hat{\beta}\rangle)^2] \leq& (1 + a)\min_{\lambda\in\mathcal{G}}\mathbb{E}[(Y - \langle X, \tau(\hat{\beta}(\lambda)^{(n_V)}(D^{(n_V)}))\rangle)^2] \\ &+ c\frac{\log|\mathcal{G}|\log(n_C)}{n_C},\end{aligned} \tag{4.1}$$

where τ is a thresholded function such that $\forall\beta \in \mathbb{R}^p, \|\tau(\beta)\|_2 \leq K_0'$.

This proves that the adaptation procedures provided in Section 2 optimize the prediction task of the Lasso thanks to a data-driven choice of the regularization parameter. Note that a classical theoretical choice of the regularization parameter λ is such that $\lambda \gtrsim \sigma\sqrt{(\log p)/n}$ (cf. [6]) – even though in many real-world applications such a choice is in general too conservative. By taking a grid $\mathcal{G}$ with a thin enough step (for instance of the order of $1/n$), such theoretical result can be used in (4.1) to prove that the procedures $\hat{\beta}$ have good prediction properties. But the real advantage of $\hat{\beta}$ is that in cases where taking λ such that $\lambda \gtrsim \sigma\sqrt{(\log p)/n}$ is not a good choice then, thanks to the adaptation properties of $\hat{\beta}$, a better choice of λ will be made "automatically" (that is in a data-dependent way).

4.2. Adaptive choice of the temperature parameter for aggregation with exponential weights

In the aggregation setup, one is given a set of data $D^{(n)}$ and a finite set F_0 of M functions $f_1, \dots, f_M$. The problem is to construct a an estimator which has a risk as close as possible to the risk of the *oracle*, the best element in F_0. A common aggregation procedure is the aggregation procedure with exponential weights (AEW for short) defined in Equation (2.12); this procedure is defined up to a free parameter which is called the temperature parameter. In this subsection, we use the adaptive procedures introduced in the previous section to choose the temperature parameter.

Let $(B_1, D_1), \dots, (B_V, D_V)$ be the family of splits of $D^{(n)}$ defined in (2.1) for some $1 \le V \le n$. For any AEW procedure $\tilde{f}(T)$ (where $T \ge 0$ is the temperature parameter) the V-fold-CV empirical risk is defined by

$$R_{n,V}(\tilde{f}(T)) = \frac{1}{V}\sum_{k=1}^{V}\frac{1}{n_C}\sum_{i=(k-1)n_C+1}^{kn_C} Q(Z_i, \tilde{f}(T)^{(n_V)}(D_k))$$

We consider the following data-driven temperature and the mCV procedure

$$\hat{T}(D^{(n)}) \in \operatorname{Arg}\min_{T\in\mathcal{G}} R_{n,V}(\tilde{f}(T)); \qquad \bar{f}^{(n)}(D^{(n)}) = \tilde{f}(\hat{T}(D^{(n)}))^{(n_V)}(D^{(n_V)}),$$

where $\mathcal{G}$ is a subset of $(0, +\infty)$.

We want to apply Theorem 3.5 to the procedure $\bar{f}^{(n)}(D^{(n)})$. We consider the bounded regression model $Y = f^*(X) + \sigma\epsilon$ with respect to the quadratic loss function $Q((x,y),f) = (y - f(x))^2$. We consider a finite dictionary F_0 (constructed with a previous sample that we supposed fixed). We assume that

$$\|\epsilon\|_\infty, \|f^*\|_\infty, \max_{f\in F_0}\|f\|_\infty = K < \infty.$$

For every $T > 0$, we construct the aggregate with exponential weights $\tilde{f}(T)$ associated with the dictionary F_0 (cf. (2.12)). Fix $A > 0$ and construct the infinite dictionary $F := \{\tilde{f}(T) : T \ge A\}$. It is easy to check that the elements of the dictionary F satisfy Assumption (A) with margin parameter $\kappa = 1$. Moreover, for every pair $T_1, T_2 > 0$ of temperature parameters, for any n, and any data set $D^{(n)}$, we have

$$\left|\tilde{f}^{(n)}(T_1)(D^{(n)})(\cdot) - \tilde{f}^{(n)}(T_2)(D^{(n)})(\cdot)\right| \le c_b d(T_1, T_2)$$

where $d(T_1, T_2) := |T_1^{-1} - T_2^{-1}|$ and thereby for $\mathcal{Q} := \{Q(\cdot, \tilde{f}(T)^{(n_V)}(D^{(n_V)})) - Q(\cdot, f^*) : T \ge A\}$, the complexity function J of Theorem 3.5 can be taken equal to

$$J(\epsilon) := c_{A,b}\left(\sqrt{\epsilon}\log\left(\frac{1}{A\sqrt{\epsilon}} + 1\right) + \frac{\log n_C}{\sqrt{n_C}}\right)$$

Thus, the exchangeability of the AEW being obvious, Theorem 3.5 yields the following oracle inequality

$$\mathbb{E}(\bar{f}^{(n)}(D^{(n)})(X)-Y)^2 \le (1+a)\min_{T\ge A}\mathbb{E}(\tilde{f}(T)^{(n_V)}(D^{(n_V)})(X)-Y)^2 + c_{a,A,b}\frac{\log^2(n_C)}{n_C}.$$

Thus the procedure $\bar{f}^{(n)}(D^{(n)})$ is optimal amongst all the AEW procedures with temperature parameter $T \ge A$ for a given A.

5. Proofs

5.1. Preliminaries from empirical process

We start with the following lemma which is a ψ_1 version of Bernstein's inequality (see, for instance, [35], Chapter 2.2).

Lemma 5.1. *Let $Y, Y_1, \dots, Y_m$ be i.i.d mean zero random variables with $\|Y\|_{\psi_1} < \infty$. Then for any $u > 0$,*

$$\mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^m Y_i > 3u\|Y\|_{\psi_1}\Big) \le \exp\big(-m(u^2 \wedge u)\big) \ .$$

Nevertheless, it appears that Lemma 5.1 does not suit the analysis we have in mind. Indeed, most of the models worked out here satisfy a margin condition; that is, a relation of the form $\mathbb{E}Y^2 \le K(\mathbb{E}Y)^{1/(2\kappa)}$. In Lemma 5.1, the sub-Gaussian and the sub-exponential (or Poisson) behavior of $\frac{1}{m}\sum_{i=1}^m Y_i$ are given with respect to the ψ_1-norm of Y without reference to the term $\mathbb{E}Y^2$. According to the Central Limit Theorem, we would expect sub-Gaussian behavior of the sum $\frac{1}{m}\sum_{i=1}^m Y_i$ with respect to the L^2-norm of Y. That is the objective of the following result. The price that one pays for replacing the ψ_1-norm by the L_2-norm under sub-Gaussian behavior is, in general, an extra factor $\log m$ in the sub-exponential behavior.

Proposition 5.2. *There exists an absolute constant $c > 0$ such that the following holds. Let $Y, Y_1, \dots, Y_m$ be i.i.d mean zero random variables such that $\|\max_{i=1,\dots,m} Y_i\|_{\psi_1} < \infty$. Then for any $u > 0$,*

$$\mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^m Y_i > u\Big) \le 2\exp\left(-cm\Big(\frac{u^2}{\mathbb{E}Y^2}\Big) \wedge \Big(\frac{u}{\|\max_{i=1,\dots,m} Y_i\|_{\psi_1}}\Big)\right).$$

Proof. We follow the line of [1]. Let $\rho > 0$ be the truncation level to be chosen later. For every $i = 1, \dots, m$ we defined $X_i := Y_i \mathbb{1}_{|Y_i|\le\rho}$ and $Z_i := Y_i - X_i$. For every $u > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^m Y_i \ge u\Big) \le \mathbb{P}\Big(\sum_{i=1}^m X_i - \mathbb{E}X_i \ge u/2\Big) + \mathbb{P}\Big(\sum_{i=1}^m Z_i - \mathbb{E}Z_i \ge u/2\Big). \quad (5.1)$$

To bound the first term of (5.1), we use the classical Bernstein inequality for bounded variables together with the inequality $\mathbb{V}(X_i) \leq \mathbb{E}X_i^2 \leq \mathbb{E}Y_i^2$ for all $i = 1, \ldots, n$

$$\mathbb{P}\Big(\sum_{i=1}^{m} X_i - \mathbb{E}X_i \geq u/2\Big) \leq \exp\Big(-\frac{cu^2}{m\mathbb{E}Y^2 + \rho u}\Big). \tag{5.2}$$

Now take $\rho := 8\mathbb{E}\max_{1\leq i\leq m}|Y_i|$. For the second term of (5.1), we note that, by Chebyshev's inequality,

$$\mathbb{P}\Big(\max_{1\leq k\leq m}\Big|\sum_{i=1}^{k} Z_i\Big| > 0\Big) \leq \mathbb{P}\Big(\max_{i=1,\ldots,m}|Z_i| > 0\Big) = \mathbb{P}\Big(\max_{i=1,\ldots,m}|Y_i| > \rho\Big) \leq 1/8.$$

Thus, by Hoffman-Jorgensen's inequality (cf. Proposition 6.8 in [21]), we have

$$\mathbb{E}\max_{k=1,\ldots,m}\Big|\sum_{i=1}^{k} Z_i\Big| \leq 8\mathbb{E}\max_{i=1,\ldots,m}|Z_i|.$$

Consequently, since $\mathbb{E}|X| \leq K\|X\|_{\psi_1}$ for any random variable X,

$$\mathbb{E}\Big|\sum_{i=1}^{m} Z_i - \mathbb{E}Z_i\Big| \leq 2\mathbb{E}\Big|\sum_{i=1}^{m} Z_i\Big| \leq 16\mathbb{E}\max_{i=1,\ldots,m}|Z_i| \leq K_0\Big\|\max_{i=1,\ldots,m}|Z_i|\Big\|_{\psi_1}. \tag{5.3}$$

Now, we use Theorem 6.21 of [21] to obtain

$$\Big\|\sum_{i=1}^{m} Z_i - \mathbb{E}Z_i\Big\|_{\psi_1} \leq K_1\left(\Big\|\sum_{i=1}^{m} Z_i - \mathbb{E}Z_i\Big\|_{L_1} + \Big\|\max_{i=1,\ldots,m}|Z_i - \mathbb{E}Z_i|\Big\|_{\psi_1}\right).$$

Combining the last result and Equation (5.3), we get

$$\Big\|\sum_{i=1}^{m} Z_i - \mathbb{E}Z_i\Big\|_{\psi_1} \leq K_2\Big\|\max_{i=1,\ldots,m} Z_i\Big\|_{\psi_1} \leq K_2\Big\|\max_{i=1,\ldots,m} Y_i\Big\|_{\psi_1}$$

In particular, we have

$$\mathbb{P}\Big(\sum_{i=1}^{m} Z_i - \mathbb{E}Z_i \geq u/2\Big) \leq \exp\Big(-\frac{cu}{\|\max_{i=1,\ldots,m} Y_i\|_{\psi_1}}\Big). \tag{5.4}$$

We obtain the result by using the last inequality together with Equation (5.2) in Equation (5.1) and noting that $\rho \leq K_3\|\max_i|Y_i|\|_{\psi_1}$. □

To obtain oracle inequalities for the mCV and amCV procedures, we need to control the suprema of some empirical processes. The next lemma is precisely such a bound for a (shifted) empirical process, and its conditions are formulated in terms of a general risk bound and a margin condition.

Lemma 5.3. *Let $\mathcal{Q} := \{Q_1, \dots, Q_p\}$ be a set of $p \geq 1$ measurable functions defined on $(\mathcal{Z}, \mathcal{T})$. Let $Z, Z_1, \dots, Z_m$ be i.i.d. random variables with values in $(\mathcal{Z}, \mathcal{T})$ such that $\forall Q \in \mathcal{Q}, \mathbb{E}Q(Z) \geq 0$. Assume the existence of constants $C_0 > 0$ and $\kappa \geq 1$ such that $\forall Q \in \mathcal{Q}, \|Q(Z)\|_{L_2} \leq C_0(\mathbb{E}Q(Z))^{1/2\kappa}$ and set $b_m := \max_{Q\in\mathcal{Q}} \|\max_{i=1,\dots,m} Q(Z_i)\|_{\psi_1}$. For any given shift parameter $a > 0$ there exists a constant $c = c(a, \kappa)$ such that*

$$\mathbb{E}\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \leq c\left(\frac{\log(ep)}{m}\right)^{\frac{\kappa}{2\kappa-1}} \vee \Big(\frac{b_m \log(ep)}{m}\Big).$$

Proof. For any $\delta > 0$, an union bound yields

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
&\qquad \leq \sum_{Q\in\mathcal{Q}} \mathbb{P}\Big[\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \geq \delta\Big] \\
&\qquad \leq \sum_{Q\in\mathcal{Q}} \mathbb{P}\Big[\mathbb{E}Q(Z) - \frac{1}{m}\sum_{i=1}^{m} Q(Z_i) \geq \frac{\delta + a\mathbb{E}Q(Z)}{1+a}\Big].
\end{aligned}$$

Now, we apply Proposition 5.2 to the random variables $Q(Z), Q(Z_1), \dots, Q(Z_m)$ and combine this with the margin-type condition $\|Q(Z)\|_{L_2} \leq C_0(\mathbb{E}Q(Z))^{1/2\kappa}$ to get the inequality

$$\begin{aligned}
&\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \\
&\qquad \leq \sum_{Q\in\mathcal{Q}} 2\exp\Big[-C_2 m\left(\Big(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}}\Big)^2 \wedge \Big(\frac{\delta + a\mathbb{E}Q(Z)}{b_m}\Big)\right)\Big],
\end{aligned}$$

where we use the constant $C_2 := \big(3C_0(1+a) \vee 9C_0^2(1+a)^2\big)^{-1}$. By comparing $\mathbb{E}Q(Z)$ and δ, it is easy to see that

$$C_2\left(\Big(\frac{\delta + a\mathbb{E}Q(Z)}{(\mathbb{E}Q(Z))^{1/(2\kappa)}}\Big)^2 \wedge \Big(\frac{\delta + a\mathbb{E}Q(Z)}{b_m}\Big)\right) \geq C_3\delta^{2-1/\kappa} \wedge (\delta/b_m)$$

when $C_3 := C_2 \cdot \big(a \wedge C_1^{-1}\big)^{1/(2\kappa)}$. Thus, for any $\delta > 0$,

$$\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z) - (1+a)\frac{1}{m}\sum_{i=1}^{m} Q(Z_i)\Big) \geq \delta\Big] \leq 2p\exp\big(-C_3 m\big(\delta^{2-1/\kappa} \wedge (\delta/b_m)\big)\big).$$

Now we use the fact that $\int_A^\infty \exp(-Bt^\alpha)dt \leq (\alpha B A^{\alpha-1})^{-1}\exp(-BA^\alpha)$ for any $\alpha \geq 1$ and $A, B > 0$ to get, for any $u > 0$ and $v > 0$,

$$\begin{aligned}
&\mathbb{E}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z)-(1+a)\frac{1}{m}\sum_{i=1}^{m}Q(Z_i)\Big)\Big]\\
&\quad\le\int_0^\infty\mathbb{P}\Big[\max_{Q\in\mathcal{Q}}\Big(\mathbb{E}Q(Z)-(1+a)\frac{1}{m}\sum_{i=1}^{m}Q(Z_i)\Big)\ge\delta\Big]d\delta\\
&\quad\le u+2p\int_u^{b_m^{-\kappa/(\kappa-1)}}\exp(-C_3m\delta^{2-1/\kappa})d\delta+v+2p\int_{b_m^{-\kappa/(\kappa-1)}+v}^{\infty}\exp(-C_3m\delta)d\delta\\
&\quad\le u+2p\frac{\exp(-C_3mu^{2-1/\kappa})}{C_3mu^{1-1/\kappa}}+v+2p\frac{\exp(-C_3m(v/b_m))}{C_3m/b_m}\,. \qquad (5.5)
\end{aligned}$$

We denote by $\mu(p)$ the unique solution of $\mu = p\exp(-\mu)$. For this quantity, we have the inequality $\mu(p)\le\log(ep)$. Take u such that $C_3mu^{2-1/\kappa}=\mu(p)$; then

$$u+2p\frac{\exp(-C_3mu^{2-1/\kappa})}{C_3mu^{1-1/\kappa}}\le 3\left(\frac{\mu(p)}{C_3m}\right)^{\frac{\kappa}{2\kappa-1}}\le 3\left(\frac{\log(ep)}{C_3m}\right)^{\frac{\kappa}{2\kappa-1}}\,.$$

Now take v such that $C_3mv=b_m\mu(p)$ to obtain

$$v+2p\frac{\exp(-C_3mv/b_m)}{C_3m/b_m}\le\frac{3b_m\mu(p)}{C_3m}\le\frac{3b_m\log(ep)}{C_3m}\,.$$

Then by plugging these values of u and v in Equation (5.5), we obtain the result. □

Note that one of the main advantages of the set of assumptions of Lemma 5.3 is that we are allowed to use unbounded random variables. And, in the bounded case $\max_{Q\in\mathcal{Q}}\|Q(Z)\|_\infty\le b_0$, we recover the classical Bernstein inequality since $\max_{Q\in\mathcal{Q}}\|\max_i Q(Z_i)\|_{\psi_1}\le b_0$. But, if we only have a ψ_1 control of the type $\max_{Q\in\mathcal{Q}}\|Q(Z)\|_{\psi_1}\le b_0$, then by using the following classical result due to Pisier (cf. for instance [35])

$$\Big\|\max_{1\le i\le m}Q(Z_i)\Big\|_{\psi_1}\le c(\log m)\max_{1\le i\le m}\|Q(Z_i)\|_{\psi_1}\,,\forall Q\in\mathcal{Q}$$

we obtain $\max_{Q\in\mathcal{Q}}\|\max_i Q(Z_i)\|_{\psi_1}\le c(\log m)b_0$. Thus, by using this approach a (maybe extra) $\log m$ term may appear in the upper bound of the shift process in Lemma 5.3 when the margin parameter κ equals to 1. If $\kappa>1$, then we obtain the same upper bound for both L_∞ and L_{ψ_1} control.

5.2. Proof of Lemma 2.3

We first prove the result for $\bar{f}^{(n)}=\hat{f}^{(n)}_{amCV}$. By definition of $\hat{\jmath}(D^{(n)})$, we have, for any $j\in\{1,\dots,p\}$,

$$\tilde{R}_{n_C}(\hat{f}_{amCV}):=\frac{1}{N_C}\sum_{k=1}^{N_C}\frac{1}{n_C}\sum_{i\notin C_k}Q(Z_i,\hat{f}^{(n_V)}_{\hat{\jmath}(D^{(n)})}(D_{(C_k)}))\le R_{n_C}(\hat{f}_j). \qquad (5.6)$$

Using inequality (5.6), we have the following basic inequality for all j and any set of data $D^{(n)}$,

$$\begin{aligned} & R(\hat{f}^{(n)}_{amCV}(D^{(n)})) - R(f^*) \\ = \; & (1+a)(\tilde{R}_{n_C}(\hat{f}_{amCV}) - R_{n_C}(f^*)) + (R(\hat{f}^{(n)}_{amCV}(D^{(n)})) - R(f^*)) \\ & -(1+a)(\tilde{R}_{n_C}(\hat{f}_{amCV}) - R_{n_C}(f^*)) \\ \leq \; & (1+a)(R_{n_C}(\hat{f}_j) - R_{n_C}(f^*)) \qquad (5.7) \\ & + \left(R(\hat{f}^{(n)}_{amCV}(D^{(n)})) - R(f^*) - (1+a)(\tilde{R}_{n_C}(\hat{f}_{amCV}) - R_{n_C}(f^*))\right). \end{aligned}$$

Since the Z_i's are i.i.d., it follows that the expectation of the first term in (5.7) is such that for every j,

$$\begin{aligned} & \mathbb{E}_{D^{(n)}} R_{n_C}(\hat{f}_j) - \mathbb{E}_{D^{(n)}} R_{n_C}(f^*) \\ = \; & \frac{1}{N_C}\sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i \notin C_k} \left(\mathbb{E}_{D^{(n)}} Q(Z_i, \hat{f}_j^{(n_V)}(D_{(C_k)})) - \mathbb{E}_{D^{(n)}} Q(Z_i, f^*)\right) \\ = \; & \frac{1}{N_C}\sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i \notin C_k} \left(\mathbb{E}_{D_{(C_k)}} R(\hat{f}_j^{(n_V)}(D_{(C_k)})) - R(f^*)\right) \\ = \; & \mathbb{E}_{D^{(n_V)}} R(\hat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*), \end{aligned}$$

and, by using the convexity of the risk, the expectation of the second term in (5.7) is such that

$$\begin{aligned} & \mathbb{E}_{D^{(n)}}\left[R(\hat{f}^{(n)}_{amCV}(D^{(n)})) - R(f^*) - (1+a)(\tilde{R}_{n_C}(\hat{f}^{(n)}_{amCV}) - R_{n_C}(f^*))\right] \\ \leq \; & \frac{1}{N_C}\sum_{k=1}^{N_C} \mathbb{E}_{D^{(n)}}\left[PQ(\cdot, \hat{f}^{(n_V)}_{\hat{j}(D^{(n)})}(D_{(C_k)})) - PQ(\cdot, f^*)\right. \\ & \left. - \frac{1+a}{n_C}\sum_{i \notin C_k}\left(Q(Z_i, \hat{f}^{(n_V)}_{\hat{j}(D^{(n)})}(D_{(C_k)})) - Q(Z_i, f^*)\right)\right] \\ \leq \; & \frac{1}{N_C}\sum_{k=1}^{N_C} \mathbb{E}_{D^{(n)}} \max_{j=1,\ldots,p}\left[PQ(\cdot, \hat{f}_j^{(n_V)}(D^{(n_V)})) - PQ(\cdot, f^*)\right. \\ & \left. - \frac{1+a}{n_C}\sum_{i \notin C_k}\left(Q(Z_i, \hat{f}_j^{(n_V)}(D_{(C_k)})) - Q(Z_i, f^*)\right)\right] \\ = \; & \frac{1}{N_C}\sum_{k=1}^{N_C} \mathbb{E}_{D^{(n)}} \max_{j=1,\ldots,p}\left[PQ(\cdot, \hat{f}_j^{(n_V)}(D^{(n_V)})) - PQ(\cdot, f^*)\right. \\ & \left. - \frac{1+a}{n_C}\sum_{i=n_V+1}^{n}\left(Q(Z_i, \hat{f}_j^{(n_V)}(D^{(n_V)})) - Q(Z_i, f^*)\right)\right] \\ = \; & \mathbb{E}_{D^{(n)}} \max_{j=1,\ldots,p}\left[PQ(\cdot, \hat{f}_j^{(n_V)}(D^{(n_V)})) - PQ(\cdot, f^*)\right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{1+a}{n_C} \sum_{i=n_V+1}^{n} \Big(Q(Z_i, \hat{f}_j^{(n_V)}(D^{(n_V)})) - (Z_i, f^*) \Big) \Bigg] \\
&= \mathbb{E}_{D^{(n)}} \max_{j=1,\dots,p} \Big[(P - (1+a)P_{n_C}) \Big(Q(\cdot, \hat{f}_j^{(n_V)}(D^{(n_V)})) - Q(\cdot, f^*) \Big) \Big],
\end{aligned}
$$

which now gives the desired result.

We can follow the same lines to obtain the oracle inequality for $\bar{f}^{(n)} = \bar{f}_{mCV}^{(n)}$. But instead of using the convexity of the risk in the second line of the last calculus we use the exchangeability and the "organized" partition scheme of the data provided by (2.1) and (2.2). Indeed, for this partition scheme, $\hat{\jmath}$ satisfies some exchangeability properties under particular permutations of the data: For any $k = 1, \dots, V$, we introduce the permutation $\phi_k(j) = j + kn_C[n]$ (where $[n]$ means modulo n). By using the exchangeability of the statistics $\hat{f}_1, \dots, \hat{f}_p$, it is easy to see that for any $k = 1, \dots, V$ and $j = 1, \dots, p$ and for $\phi_k(B_p) := \{\phi_k((p-1)n_C + 1), \dots, \phi_k(pn_C)\}$ and $\phi_k(D_p) := \{Z_i : i \notin \phi_k(B_p)\}$,

$$
\frac{1}{V} \sum_{p=1}^{V} \frac{1}{n_C} \sum_{i \in \phi_k(B_p)} Q(Z_i, \hat{f}_j^{(n_V)}(\phi_k(D_p))) = R_{n,V}(\hat{f}_j),
$$

and thus that $\hat{\jmath}(Z_{\phi_k(1)}, \dots, Z_{\phi_k(n)}) = \hat{\jmath}(D^{(n)})$. Moreover, for each $k = 1, \dots, V$, $\phi_k(D^{(n_V)}) = \phi_k(D_{V-1}) = D_k$, so we have

$$
\begin{aligned}
&\mathbb{E}_{D^{(n)}} R(\hat{f}_{mCV}^{(n)}(D^{(n)})) = \mathbb{E}_{D^{(n)}} R\big(\hat{f}_{\hat{\jmath}(D^{(n)})}^{(n_V)}(D^{(n_V)})\big) \\
&= \frac{1}{V} \sum_{k=1}^{V} \mathbb{E}_{D^{(n)}} R\big(\hat{f}_{\hat{\jmath}(Z_{\phi_k(1)}, \dots, Z_{\phi_k(n)})}^{(n_V)}(\phi_k(D^{(n_V)}))\big) = \frac{1}{V} \sum_{k=1}^{V} \mathbb{E}_{D^{(n)}} R\big(\hat{f}_{\hat{\jmath}(D^{(n)})}^{(n_V)}(D_k)\big).
\end{aligned}
$$

5.3. Proof of Lemma 3.2

For any $q > 1$ and $u \geq 1$, the following family of inequalities holds simultaneously with probability greater than $1 - L \sum_{j=0}^{\infty} \exp(-cq^j u) \geq 1 - L_1 \exp(-cu)$:

$$
\sup_{Q \in \mathcal{Q}: PQ \leq q^j \epsilon_q(q^j u)} \big((P - P_m)Q\big)_+ \leq \frac{q^j u J\big(q^j \epsilon_q(q^j u)\big)}{\sqrt{m}} \text{ for every } j \geq 0.
$$

Thus, with probability greater than $1 - L_1 \exp(-cu)$,

$$
\begin{aligned}
&\sup_{Q \in \mathcal{Q}} \Big(PQ - (1+a)P_m Q \Big)_+ = \sup_{Q \in \mathcal{Q}} \Big((1+a)(PQ - P_m Q) - aPQ \Big)_+ \\
&\leq \sum_{j=1}^{\infty} \sup_{\substack{Q \in \mathcal{Q}: \\ q^{j-1}\epsilon_q(q^{j-1}u) < PQ \leq q^j \epsilon_q(q^j u)}} \Big((1+a)(PQ - P_m Q) - aPQ \Big)_+ + \frac{(1+a)uJ(\epsilon_q(u))}{\sqrt{m}} \\
&\leq \sum_{j=1}^{\infty} \sup_{\substack{Q \in \mathcal{Q}: \\ PQ \leq q^j \epsilon_q(q^j u)}} \Big((1+a)(PQ - P_m Q) - aq^{j-1}\epsilon_q(q^{j-1}u) \Big)_+ + \frac{(1+a)uJ(\epsilon_q(u))}{\sqrt{m}}
\end{aligned}
$$

$$\leq \sum_{j=1}^{\infty}\Big((1+a)\sup_{\substack{Q\in\mathcal{Q}:\\ PQ\leq q^j\epsilon_q(q^ju)}}(PQ-P_mQ)-aq^{j-1}\epsilon_q(q^{j-1}u)\Big)_+ + \frac{(1+a)uJ(\epsilon_q(u))}{\sqrt{m}}$$
$$\leq \sum_{j=1}^{\infty}\Big(\frac{(1+a)q^juJ(q^j\epsilon_q(q^ju))}{\sqrt{m}}-aq^{j-1}\epsilon_q(q^{j-1}u)\Big)_+ + \frac{(1+a)uJ(\epsilon_q(u))}{\sqrt{m}}$$
$$\leq \frac{(1+a)uJ(\epsilon_q(u))}{\sqrt{m}} \leq \frac{a\epsilon_q(u/q)}{q},$$

where we use that $x>0\longmapsto \psi(x)/x^r$ decreases and so for every $j\geq 0, u\geq 1$, $\epsilon_q(q^ju)\leq q^r\epsilon_q(q^{j-1}u)$ which, using the convex duality property of ψ, implies

$$\frac{(1+a)q^juJ(q^j\epsilon_q(q^ju))}{\sqrt{m}} = \frac{a}{2q^{r+1}}\frac{2q^{j+r+1}(1+a)u}{a\sqrt{m}}J(q^j\epsilon_q(q^ju))$$
$$\leq \frac{a}{2q^{r+1}}\Big(q^j\epsilon_q(q^ju)+\psi\Big(\frac{2q^{j+r+1}(1+a)u}{a\sqrt{m}}\Big)\Big)$$
$$= \frac{a(q^j+1)}{2q^{r+1}}\epsilon_q(q^ju)\leq aq^{j-1}\epsilon_q(q^{j-1}u).$$

Now, for the "moreover" part, we use that $x>0\longmapsto\psi(x)/x^r$ decreases to get $\epsilon_q(u/q)=o(\exp(cu))$ when u tends to infinity; $\epsilon_q(u/q)\leq u^r\epsilon_q(1/q), \forall u\geq 1$ and

$$\mathbb{E}\sup_{Q\in\mathcal{Q}}\Big((P-(1+a)P_m)Q\Big)_+ = \int_0^{\infty}\mathbb{P}\Big[\sup_{Q\in\mathcal{Q}}\Big((P-(1+a)P_m)Q\Big)_+\geq t\Big]dt$$
$$\leq \frac{a\epsilon_q(1/q)}{q}+\int_1^{\infty}L_1\exp(-cu)\frac{a\epsilon_q'(u/q)}{q^2}du$$
$$= \frac{a\epsilon_q(1/q)}{q}+\frac{aL_1}{q}e^{-c}\epsilon_q(1/q)+\frac{aL_1c}{q}\int_1^{\infty}\exp(-cu)\epsilon_q(u/q)du$$
$$\leq \frac{a\epsilon_q(1/q)}{q}+\frac{aL_1}{q}e^{-c}\epsilon_q(1/q)+\frac{aL_1c}{q}\int_1^{\infty}\exp(-cu)u^r\epsilon_q(1/q)du \leq \frac{ac_1\epsilon_q(1/q)}{q}.$$

5.4. *Proof of Lemma 3.4*

Let $\epsilon>0$ and take $Q_0\in\mathcal{Q}$ such that $PQ_0\leq\epsilon$. We have

$$\sup_{Q\in\mathcal{Q}:PQ\leq\epsilon}\big((P-P_m)Q\big)_+ \leq \sup_{Q\in\mathcal{Q}:PQ\leq\epsilon}|Z(Q)-Z(Q_0)|+|Z(Q_0)|$$

where $Z(Q)=(P-P_m)Q, \forall Q\in\mathcal{Q}$.

For every $Q_1,Q_2\in\mathcal{Q}_\epsilon^{L_2}$, Proposition 5.2 and Pisier's inequality yield for any $u\geq 1$

$$\mathbb{P}\Big[|Z(Q_1)-Z(Q_2)|\geq u\Big]\leq 2\exp\Big(-\min\Big(\frac{u^2}{d_2^2(Q_1,Q_2)},\frac{u}{d_1(Q_1,Q_2)}\Big)\Big)$$

where

$$d_2^2(Q_1,Q_2)=\frac{\|Q_1-Q_2\|_{L_2(\pi)}^2}{cm} \text{ and } d_1(Q_1,Q_2)=\frac{\|Q_1-Q_2\|_{L_{\psi_1}(\pi)}(\log m)}{cm}.$$

Since $\{Q\in\mathcal{Q}:PQ\le\epsilon\}\subset\mathcal{Q}_\epsilon^{L_2}$, Equation (3.4) provides two absolute constant $L,c>0$ such that for every $u\ge 1$ with probability greater than $1-L\exp(-cu)$

$$\sup_{Q\in\mathcal{Q}:PQ\le\epsilon}|Z(Q)-Z(Q_0)|\le uL\Big(\frac{(\log m)\gamma_1(\mathcal{Q}_\epsilon^{L_2},\|\cdot\|_{\psi_1})}{m}+\frac{\gamma_2(\mathcal{Q}_\epsilon^{L_2},\|\cdot\|_{L_2})}{\sqrt{m}}\Big).$$

Then, Proposition 5.2 applied to the single element $Q_0\in\mathcal{Q}$ such that $PQ_0\le\epsilon$ provides, for any $u\ge 1$, with probability greater than $1-L\exp(-cu)$,

$$|Z(Q_0)|\le\frac{u\log m}{m}\|Q_0\|_{\psi_1(\pi)}+\sqrt{\frac{u}{m}}\|Q_0\|_{L_2(\pi)}.$$

Combining the two last results with the fact that $\operatorname{diam}(T,d)\le c\gamma_\alpha(T,d)$ for every metric space (T,d) provides the result.

6. Simulation study

In this section, we compare the experimental performances of the procedures under examination. For this comparison, we shall apply these procedures to a high-dimensional model aggregating elastic net estimators, which extends the application in Section 4.1.

6.1. Model

For a given size d, let $(X^{(1)},\dots,X^{(d)})$ be a random vector uniformly distributed on the unit hypercube $[0,1]^d$, and let ε be Gaussian noise with variance σ^2. For a fixed vector $\beta\in\mathbb{R}^{p+1}$, we define

$$Y=\beta^TX+\varepsilon\ .$$

Generalizing the ℓ_1-penalty used by the Lasso, we take a weighted average of ℓ_1- and ℓ_2-penalties, the latter stemming from Ridge regression. This gives us the elastic net estimator

$$\hat{f}_{\lambda,\alpha}^{(\text{Elastic net})}:=\arg\min_\beta\left\|Y-\beta^TX\right\|_{2,n}^2+\lambda(\alpha\left|\beta\right|_1+(1-\alpha)\left|\beta\right|_2^2)\ ,$$

which has two tuning parameters, $\lambda\in[0,\lambda_{max}]$ (as in Section 4.1) and $\alpha\in[0,1]$ (cf. [39, 13]). Taking grids Λ and G for λ and α, respectively, we obtain a family of estimators for aggregation:

$$\mathcal{F}:=\left\{\hat{f}_{\lambda,\alpha}^{(\text{Elastic net})}|\lambda\in\Lambda,\alpha\in G\right\}\ .$$

For the aggregation step itself, we use L^r-loss, where r does not always have to equal 2.

6.2. Methods and results

The family of estimators thus described we aggregated in three different ways:

- Model selection by cross-validation: the aggregate estimator is

$$\bar{f}^{(n)}_{VCV}(D^{(n)}) = \hat{f}^{(n)}_{\hat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \hat{\jmath}(D^{(n)}) \in \operatorname{Arg}\min_{j\in\{1,\dots,p\}} R_{n,V}(\hat{f}_j),$$

  the definition already given in (2.4). The weights assigned to each model are either 0 or 1.
- Aggregation with exponential weights using the cross-validation score: here the aggregate estimator is

$$\bar{f}^{(n)}_{AEW}(D^{(n)}) = \sum_{j=1}^{p} w_j \hat{f}^{(n)}_j(D^{(n)})$$

  and

$$w_j = \frac{\exp\big(-nR_{n,V}(\hat{f}_j)/T\big)}{\sum_{k=1}^{p}\exp\big(-nR_{n,V}(\hat{f}_k)/T\big)}.$$

- Split-wise aggregation with exponential weights, averaging the estimator over all splits: the aggregate estimator in this situation is

$$\bar{f}^{(n)}_{SW}(D^{(n)}) = \sum_{j=1}^{p}\Big(\frac{1}{V}\sum_{k=1}^{V} w^{(k)}_j\Big)\hat{f}^{(n)}_j(D^{(n)})$$

  and

$$w^{(k)}_j = \frac{\exp\big(-nR^{(k)}_{n,V}(\hat{f}_j)/T\big)}{\sum_{k=1}^{p}\exp\big(-nR^{(k)}_{n,V}(\hat{f}_k)/T\big)}.$$

Computation We simulated the performance for the elastic net estimator using covariates in $\mathbb{R}^{200}$ and the distribution of (X, Y) given above. The total sample size for which we looked at the performance of the estimator was 100, and the cross-validation procedure we used was 10-fold cross-validation. Regardless of the sparsity of a model, we scaled the true coefficients so that $|\beta|_2 = 1$ and chose the error variance so that the signal-to-noise ratio was 5. The loss order for the aggregation step we took to be $r = 2$ for the first simulation runs.

We used the `glmnet` procedure (cf. [13]) to compute the elastic net estimators. This procedure standardizes covariates at the beginning, ensuring that $\sum_{i=1}^{n} X_i^{(j)} = 0$ and $(\sum_{i=1}^{n}(X_i^{(j)})^2)/N = 1$ for all $j = \{1,\dots,d\}$. The penalty scaling parameter λ we chose to take on values in a logarithmically equispaced grid $\Lambda := \{1, 0.05^{0.1}, 0.05^{0.2}, \dots, 0.05\}\cdot\lambda_{max}$, and the penalty mixing parameter α to take on the equispaced grid of values $G := \{0, 0.1, \dots, 1\}$.

Fixing the aggregation temperature parameter to $T = 5$ (seen to be a reasonable value from various simulation attempts), we took varying degrees of sparsity for the true model. The class of models we are interested in aggregating

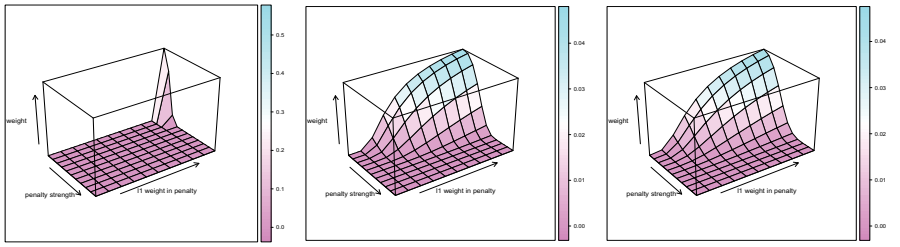

FIG 1. *Weights yielded by cross-validation (left), cross-validatory aggregation with exponential weights (middle) and split-wise aggregation (right), with the true model consisting of covariates 1, 2 and 3* ($\Delta = \sqrt{3}$).

contains both "sparse" estimators, which perform variable selection (such as the Lasso), and "non-sparse" estimators, which do not (such as Ridge regression). An important indicator of sparsity is the ratio

$$\Delta(\beta) = \frac{|\beta|_1}{|\beta|_2} \in [1, \sqrt{p}] \ .$$

In terms of this indicator, $\Delta(\beta) = 1$ corresponds to a sparse model (which should elicit comparatively good performance from the Lasso). $\Delta(\beta) = \sqrt{p}$ corresponds to a well-spread model (which is nice for ridge regression). To investigate the degree to which our estimators are sensitive to sparsity, we performed a series of simulations, in which the true model for the k-th plot is given by the coefficient vector $\beta^{(k)}$ with $\beta_j^{(k)} = 1\{j \leq k\}$. The corresponding sparsity ratios are $\Delta(\beta_j) = \sqrt{j}$.

Results When the true model is one including just the first 3 covariates ($\Delta = \sqrt{3}$), we obtained – on average over 100 repetitions – the weights displayed in Figure 1. Here, as in the other figures displaying weights, we show individual plots for aggregation with exponential weights using the cross-validated score on the left, and split-wise aggregation on the right.

Other sparse models up to $\Delta = \sqrt{20}$ provide similar weights: ones concentrated on the ℓ_1-penalized models. For $\Delta = 10$, by contrast (100 out of 200 variables contained in the true model), all 3 procedures concentrate more on the ℓ_2-penalized procedures (see Figure 2), as was to be expected.

Using other loss functions for the model selection/aggregation step The simulations that we just described were also carried out for various different loss functions in the cross-validation step (using covariates 1 to 3 in the true model). Instead of squared loss ($r = 2$), we tried the loss functions given by $r = 3/2$, $r = 1$ (absolute loss), $r = 1/2$ and $r = 1/5$ (the latter two of course being non-convex). However, these loss functions did not yield the desired increase in selection sharpness for our example.

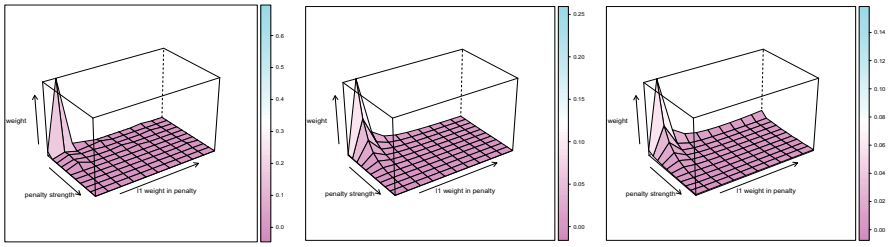

FIG 2. *Weights yielded by cross-validation (left), cross-validatory aggregation with exponential weights (middle) and split-wise aggregation (right), with the true model including covariates 1-100* ($\Delta = 10$).

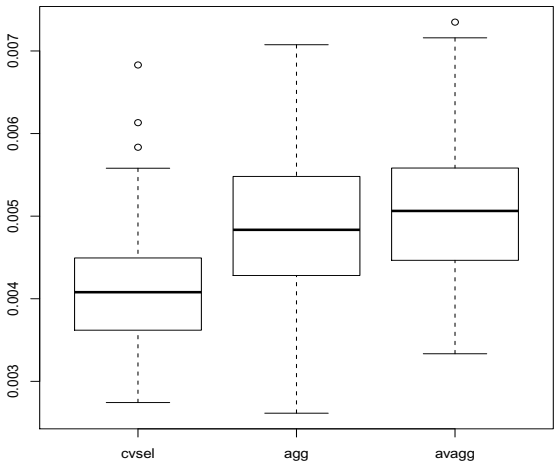

FIG 3. *L_2-estimator risk of cross-validation (left), cross-validatory aggregation with exponential weights (middle) and split-wise aggregation (right), with the true model containing covariates 1-3.*

Estimator risk In the simulations we performed, the estimator risk (as computed for each of 100 samples) is already roughly normal in distribution. For the sparse model of Figure 1, cross-validation based model selection has a marginally lower estimator risk compared to the two aggregation procedures using exponential weights (Figure 3). For the non-sparse model already seen in Figure 2, the three procedures have more similar estimator risks (Figure 4).

Acknowledgements

The authors would like to acknowledge anonymous referees for some technical comments.

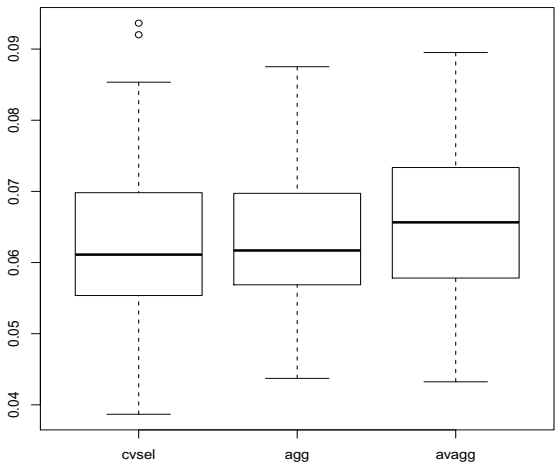

FIG 4. *L_2-estimator risk of cross-validation (left), cross-validatory aggregation with exponential weights (middle) and split-wise aggregation (right), with the true model including covariates 1-100.*

References

[1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13**, no. 34, 1000–1034. MR2424985

[2] ARLOT, S. & CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, 40–79. MR2602303

[3] BARRON, A., BIRGÉ, L. & MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413. MR1679028

[4] BARTLETT, P. L. & JORDAN, M. I. AND MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101**, 138–156. MR2268032

[5] BARTLETT, P. L. & MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135**, 311–334. MR2240689

[6] BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for high-dimensional data* in *Springer Series in Statistics. Methods, theory and applications.* Springer, Heidelberg. MR2807761

[7] BOUSQUET, O. & ELISSEEFF, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2(3)**, 499–526. MR1929416

[8] CATONI, O. (2007). *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*, *Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56.* Institute of Mathematical Statistics, Beachwood, OH. MR2483528

[9] CATONI, O. (2004). *Statistical learning theory and stochastic optimization*, vol. 1851 of *Lecture Notes in Mathematics.* Springer, Berlin. MR2163920

[10] Cornec, M. (2009). *Probability bounds for the cross-validation estimate in the context of the statistical learning theory and statistical models applied to economics and finance*. PhD Thesis, CREST - Centre de Recherche en économie et statistique.
[11] Devroye, L. & Wagner, T. (1979). Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory* **25(5)**, 601–604. MR0545015
[12] Dudley, R. M. (1999). Uniform central limit theorems. *Cambridge University Press*. MR1720712
[13] Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** (1).
[14] Gaïffas, S. & Lecué, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.* **1**, 538–573. MR2369025
[15] Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11(4)**, 1156–1174. MR0720261
[16] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34**, 2593–2656. MR2329442
[17] Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, **22**, 45 – 55.
[18] Lecué, G. (2007). Suboptimality of penalized empirical risk minimization in classification. In *20th Annual Conference On Learning Theory, COLT07* (eds. N.H. Bshouty & C. Gentile), 142–156. Springer, Berlin. MR2397584
[19] Lecué, G. (2007). Optimal rate of aggregation in classification under low noise assumption. *Bernoulli* **13(4)**, 1000–1022. MR2364224
[20] Lecué, G. & Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145**, 591–613. MR2529440
[21] Ledoux, M. & Talagrand, M. (1991). *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Berlin.
[22] Lepskii, O. (1990). A problem of adaptive estimation in Gaussian white noise. *(Russian) Teor. Veroyatnost. i Primenen.* **35(3)** 459–470 *translation in Theory Probab. Appl.* **35(3)** 454–466 (1991). 62M05 (62G20) MR1091202
[23] Lepskii, O. (1992). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *(Russian) Teor. Veroyatnost. i Primenen.* **36(4)** (1991) 645–659; *translation in Theory Probab. Appl.* **36(4)** 682–697 (1992). MR1147167
[24] Mammen, E. & Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808–1829. MR1765618
[25] Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture notes in mathematics*. Springer, Berlin. MR2319879
[26] Milman, V. D. & Schechtman, G. (1986). *Asymptotic theory of finite-dimensional normed spaces*, vol. 1200 of *Lecture Notes in Mathematics*. Springer, Berlin. MR0856576

[27] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, vol. 1738 of *Lecture Notes in Mathematics*, 85–277. Springer, Berlin. MR1775640

[28] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486–494. MR1224373

[29] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.*, Ser. B **36**, 111–147. MR0356377

[30] Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12(4)**, 1285–1297. MR0760688

[31] Talagrand, M. (2005). *The generic chaining.* Springer Monographs in Mathematics. Springer, Berlin. MR2133757

[32] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135–166. MR2051002

[33] van de Geer, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge. MR1739079

[34] van der Vaart, A. W., Dudoit, S. & van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statist. Decisions* **24**, 351–371. MR2305112

[35] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer, New York. MR1385671

[36] Vapnik, V. (1982) Estimation of dependences based on empirical data. *Translated from the Russian by Samuel Kotz. Springer Series in Statistics. Springer-Verlag, New York-Berlin.* MR0672244

[37] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75–87. MR1762904

[38] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32**, 56–85. MR2051001

[39] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67**, 301–320. MR2137327