A MOM-based ensemble method for robustness, subsampling and hyperparameter tuning

Joon Kwon, Guillaume Lecué, Matthieu Lerasle

Electronic Journal of Statistics
Vol. 15 (2021) 1202–1227
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1814

A MOM-based ensemble method for robustness, subsampling and hyperparameter tuning*

Joon Kwon

INRAE — AgroParisTech
16 rue Claude Bernard, 75231 Paris Cedex 05
e-mail: joon.kwon@inra.fr

Guillaume Lecué and Matthieu Lerasle

ENSAE
5 avenue Henry Le Chatelier
91120 Palaiseau
e-mail: guillaume.lecue@ensae.fr; matthieu.lerasle@ensae.fr

Abstract: Hyperparameter tuning and model selection are important steps in machine learning. Unfortunately, classical hyperparameter calibration and model selection procedures are sensitive to outliers and heavy-tailed data. In this work, we construct a selection procedure which can be seen as a robust alternative to cross-validation and is based on a median-of-means principle. Using this procedure, we also build an ensemble method which, trained with algorithms and corrupted heavy-tailed data, selects an algorithm, trains it with a large uncorrupted subsample and automatically tunes its hyperparameters. In particular, the approach can transform any procedure into a robust to outliers and to heavy-tailed data procedure while tuning automatically its hyperparameters.

The construction relies on a divide-and-conquer methodology, making this method easily scalable even on a corrupted dataset. This method is tested with the LASSO which is known to be highly sensitive to outliers.

MSC2020 subject classifications: 62F35, 60K35.
Keywords and phrases: Robustness, heavy-tailed.

Received September 2019.

Contents

*The authors gratefully acknowledge financial support from Labex ECODEC (ANR - 11-LABEX-0047).

1. Introduction

Robustness has become an important subject of interest in the machine learning community over the last few years because large datasets are very likely to be corrupted. This may happen due to hardware, storage or transmission issues, for instance, or as a result of (human) reporting errors. As can be seen, for instance, in Figure 1 in [29] and Figure 1 and 5 in [30], many learning algorithms based on empirical risk minimization (including the LASSO) may be completely mislead by a single corrupted example.

Robust alternatives to empirical risk minimizers and their penalized/regularized versions have been studied in density estimation [5] and least-squares regression [4, 36, 20, 50, 55]. Various robust descent algorithms have also been recently considered [46, 44, 45, 23, 22]. Despite these important advances, the final steps of a data-scientist routine, which are estimator selection and hyperparameter tuning [8, 11, 7] are yet to receive a proper treatment. In fact, practitioners usually have at disposal several algorithms, each of these requiring one or several parameters to be tuned. An alternative to estimator selection is aggregation (aka ensemble methods) [16, 52, 42] which outputs e.g. a linear or convex combination of the candidate estimators; classical examples include binning, boosting, bagging or stacking.

The most common procedure used to select or aggregate candidate estimators is (cross-)validation: the dataset is partitioned (several times in the case of cross-validation) into a *training sample* used to build candidate estimators and a *test sample* used to estimate their risks. The final estimator is either the candidate with lowest estimated risk, or a linear combination of the candidates with coefficients depending on the estimated risks. Even if some candidate estimators are robust, outliers from the test set may mislead the selection/aggregation

step, resulting in a poor final estimator. This raises the question of a robust selection/aggregation procedure, which is addressed in the present work.

There exist many data-driven methods to tune hyperparameters or to select an estimator from a collection of candidates. Among these, one can mention the SURE method [49], model selection [8, 11, 12, 35, 9, 40] where penalization methods are used to select among candidates built with *the same data* as those used to build the original estimators, selection, convex or linear aggregation [52, 47, 54], cross-validation [2, 3] or Lepski [33] and the Goldenschluger-Lepski [21] methods to name a few. To the best of our knowledge, all these techniques either use a classical non-robust validation principle or estimate the risk with the non-robust empirical risk. A notable exception is the estimator selection procedure of [6, 7] which is robust in general settings [6] and extremely efficient in Gaussian linear regression [7]. The main drawback is that this procedure requires robust tests in Hellinger distance that may be hard to compute for general learning problems where one does not specify statistical models with bounded complexities.

The first contribution of this paper is a general and robust estimator selection procedure with provable theoretical guarantees, which can be viewed as a robust alternative to cross-validation. Roughly stated, the procedure uses a median-of-means principle [1, 24, 43] to build robust pairwise comparisons between candidates, and the final estimator is then selected by a minmax procedure in the spirit of [4, 28] or the Goldenshluger-Lepski method [21], see Section 3 for details. The method is easily implementable. We here focus on least-squares regression and refer to [34] for other examples including density estimation and classification.

The second contribution is the definition of an ensemble method based on this selection procedure and a subsampling strategy. Two of the main ideas behind this method is that subsampling can provide robustness by avoiding outliers and that the choice of the subsample can itself be seen an a hyperparameter to be tuned. Estimator selection procedures can then be used to simultaneously select the best algorithm, an uncorrupted subsample and the best hyperparameters. Moreover, the method is computationally attractive.

Subsampling is usually used in machine learning for computational reasons: some algorithms require to break large datasets into smaller pieces [25], for instance in supervised learning [17, for classification and regression] and [37, for matrix factorization]. A natural way to divide-and-conquer corresponds to the older idea of *subagging* [15, subsample aggregating]—which is a variant of bagging [14]: one randomly chooses several small subsets of data, build an estimator from each subsample, and aggregate them into a single estimator. For instance, the bag of little bootstraps [26] builds confidence intervals in such a way. Subagging is also used for large-scale sparse regression [13].

The paper is divided as follows. Section 2 presents the general prediction setting we consider, and Section 3 introduces the robust estimator selection procedure. Theoretical guarantees for the latter are given in Theorem 3.2. The ensemble method is defined in Section 4 and applied to the LASSO in Section 5. Applications to the ERM in linear aggregation are presented in Appendix 6.

Numerical experiments are presented in Section 7. The proofs are outsourced in the appendix in Appendices A and B.

2. Setting

For positive integers $k \leqslant l$, let $[k] = \{1, 2, \dots, k\}$, $[\![k, l]\!] = \{k, k+1, \dots, l\}$, and reversed double-bar brackets mean exclusion of the corresponding integer, e.g. $]\!]k, l]\!] = \{k+1, k+2, \dots, l\}$. We call partition of a set E any family of disjoint subsets of E with union equal to E.

Let $\mathbb{X}$ be a measurable space. Let P be a probability distribution on $\mathbb{X} \times \mathbb{R}$, and let $(X, Y) \sim P$. Denote P_X the marginal distribution of X. Assume that $\mathbb{E}\left[Y^2\right] < +\infty$. Denote $L^2(P_X)$ the Hilbert space of measurable functions $f : \mathbb{X} \to \mathbb{R}$ such that $\mathbb{E}\left[f(X)^2\right] < +\infty$, the norm being denoted by $\|f\| = \sqrt{\mathbb{E}\left[f(X)^2\right]}$. For any probability measure Q on $\mathbb{X} \times \mathbb{R}$ and any measurable function $g : \mathbb{X} \times \mathbb{R} \to \mathbb{R}$, that belongs to $L^1(Q)$, let $Q[g] := \mathbb{E}_{Z \sim Q}[g(Z)]$.

Let F be a linear subspace of $L^2(P_X)$. For $f \in F$, let $\gamma(f) : \mathbb{X} \times \mathbb{R} \to \mathbb{R}$ denote the square-loss function associated with f, defined by for all $(x, y) \in \mathbb{X} \times \mathbb{R}$ by $\gamma(f)(x, y) = (y - f(x))^2$. For any function $f : \mathbb{X} \to \mathbb{R}$ in $L^2(P_X)$, let $R(f)$ denote its *risk* $R(f) := P[\gamma(f)]$ and let f^* be the *oracle*: $f^* := \arg\min_{f \in F} R(f)$. Let ℓ denote the *excess risk* with respect to f^*:

$$\ell(f) = R(f) - R(f^*) = P\left[(f - f^*)^2\right] = \|f - f^*\|^2 .$$

The second equality holds since F is a linear space. A *learning algorithm* is a measurable map G: $\bigcup_{n=1}^{+\infty} (\mathbb{X} \times \mathbb{R})^n \to F$ which takes a dataset of any (finite) size as input and outputs an estimator in F.

Assumption 1. *Let $\chi, \sigma > 0$ such that for every $f \in F$,*

$$(Pf^4)^{1/4} \leqslant \chi (Pf^2)^{1/2} \quad \textit{and} \quad P\left[(Y - f^*)^2 (f - f^*)^2\right] \leqslant \sigma^2 P(f - f^*)^2 .$$

This assumption only involves second and fourth moments. The first assumption $(Pf^4)^{1/4} \leqslant \chi (Pf^2)^{1/2}$ is satisfied for instance by linear functions $f(\cdot) = \langle \cdot, t \rangle$ for $t \in \mathbb{R}^d$ and X which is a d-dimensional vectors with independent entries with a fourth moment [41]. It therefore covers heavy-tailed cases beyond classical L_∞-boundedness or subgaussian assumptions. The second assumption $P\left[(Y - f^*)^2 (f - f^*)^2\right] \leqslant \sigma^2 P(f - f^*)^2$ holds for instance when the noise $Y - f^*(X)$ is independent of X and has a second moment—which is a very standard statistical modeling assumption when $Y = f^*(X) + \zeta$ with ζ independent of X. It also holds when $Y - f^*(X)$ has a fourth moment by using Cauchy-Schwarz.

Let $N \geqslant 1$ be the size of the dataset $(X_i, Y_i)_{i \in [N]}$, which is partitioned into informative data and outliers: $[N] = \mathcal{O} \sqcup \mathcal{I}$. Informative data $(X_i, Y_i)_{i \in \mathcal{I}}$ is assumed independent and identically distributed (i.i.d.), with common distribution P. No assumption is granted on outliers $(X_i, Y_i)_{i \in \mathcal{O}}$. Of course, the partition $\mathcal{O} \sqcup \mathcal{I}$ is unknown to the learner. We call a *subsample* any nonempty subset

$B \subset [N]$ (or the corresponding data $(X_i, Y_i)_{i\in B}$), and for any measurable function $g : \mathbb{X} \times \mathbb{R} \to \mathbb{R}$, denote:

$$P_B[g] = \frac{1}{|B|} \sum_{i\in B} g(X_i, Y_i).$$

3. Minmax-MOM selection: a robust alternative to cross-validation

Let $(\hat{f}_m)_{m\in\mathcal{M}}$ be a finite collection of estimators. For each index $m \in \mathcal{M}$, we assume that there exists a learning algorithm G_m and a subsample $B_m \subset [N]$ of cardinality less than $N/4$ such that $\hat{f}_m$ is the estimator trained by algorithm G_m using subsample B_m; in other words $\hat{f}_m = G_m((X_i, Y_i)_{i\in B_m})$ (the remaining of the dataset will be used to estimate the risk of the estimator, like in cross-validation). The best choice of $m \in \mathcal{M}$ regarding our final objective satisfies

$$m_o := \arg\min_{m\in\mathcal{M}} P\left[\gamma(\hat{f}_m)\right] = \arg\min_{m\in\mathcal{M}} \max_{m'\in\mathcal{M}} P\left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\right] \quad . \tag{3.1}$$

However, the real-valued expectations $P\left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\right]$ are unknown: let us construct a robust estimator of those quantities. Let $V \in [\![1, N/8]\!]$. For each couple $(m, m') \in \mathcal{M}^2$, let $(T_v^{(m,m')})_{v\in[V]}$ be a partition into V blocks of a subset of $[N] \setminus (B_m \cup B_{m'})$, such that $\left|T_v^{(m,m')}\right| \geqslant N/4V$ for all $v \in [V]$. The estimates $\mathcal{T}(m, m')$ of $P\left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\right]$ are defined by:

$$\mathcal{T}(m, m') := \operatorname*{med}_{v\in[V]} \left\{ P_{T_v^{(m,m')}} \left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\right] \right\},$$

in other words, $\mathcal{T}(m, m')$ is the median of the V empirical means $P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\right], v \in [V]$. The selection of the final estimator is obtained by plugging these median-of-means (MOM) estimators into equation (3.1). In other words, we select $\hat{f}_{\hat{m}}$, where

$$\hat{m} := \arg\min_{m\in\mathcal{M}} \max_{m'\in\mathcal{M}} \mathcal{T}(m, m'). \tag{3.2}$$

Thanks to the median-of-means operator, the risk of the selected estimator $\hat{f}_{\hat{m}}$ is expected to be close to the risk of the best estimator $\hat{f}_{m_o}$, even for heavy-tailed and corrupted data because $\mathcal{T}(m, m')$ is a robust (to outliers) sub-Gaussian estimator of $\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})$, even for heavy-tailed data [34].

Median-of-means have been introduced in [1, 24, 43]. Median-of-means pairwise comparisons have been used to build robust estimators in [36, 27]. Minmax strategies have been used in [4, 28] for least-squares regression and in [5] for density estimation. The minmax principle has been used for (non-robust) selection of estimators in [21].

The procedure is related the one introduced in [34, Eq (7) in Section 5], an important difference though is that in this original paper, a partition of the data-set was given in advance and a single estimator was built using data in

one block. By comparison here, the initial collection can be much larger than V and the new procedure is much more flexible to select hyperparameters or turn non-robust procedures into more robust ones as discussed later in the paper. Besides, the new procedure can also help reduce computational time of some procedures as discussed in the following remark.

Remark 3.1 (Minmax-MOM selection to divide-and-conquer). It is classical to use divide-and-conquer approaches [25] to deal with large datasets: the dataset is divided in small batches, algorithms are run on each batch and the results are "aggregated". Minmax-MOM selection procedure (3.2) can perform this kind of aggregation. Denote by B_m the block of data hosted on server $m \in \mathcal{M}$. Train estimators $\hat{f}_m$ for all $m \in \mathcal{M}$. Then, for all $m, m' \in \mathcal{M}$, compute the V real numbers $P_{T_v^{(m,m')}}\big[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\big]$, $v \in [V]$ and take their median. Then compute the minmax-MOM estimator (if there are too many medians, choose m and m' at random in $\mathcal{M}$). Following the map-reduce terminology [18], the mapper is the training of the procedure itself and the V evaluations. The reducer is the computation of the $\binom{|\mathcal{M}|}{2}$ medians of differences of empirical risks and the minmax-MOM selection (3.2).

Theorem 3.2 (Robust oracle inequality)**.** *Grant Assumption 1 and assume* $V \in [\![3\,|\mathcal{O}|\,, N/8]\!]$. *Then with probability larger than* $1 - |\mathcal{M}|^2\, e^{-V/48}$, *the estimator* $\hat{f}_{\hat{m}}$, *where* $\hat{m}$ *is selected by the minmax-MOM selection procedure* (3.2) *satisfies, for all* $\varepsilon > 0$,

$$(1 - a_{\varepsilon,V})\,\ell(\hat{f}_{\hat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \min_{m \in \mathcal{M}} \ell(\hat{f}_m) + 2b_{\varepsilon,V},$$

where $f \mapsto \ell(f) = R(f) - R(f^*)$ *is the excess loss function,* $a_{\varepsilon,V} := 8\chi^2\sqrt{2V/N} + 2\sqrt{2}\varepsilon$ *and* $b_{\varepsilon,V} := (64V\sigma^2)/N\varepsilon$.

The proof of Theorem 3.2 is postponed to Section B.1. Roughly speaking, Theorem 3.2 states that, with exponentially large probability, the selected estimator (3.2) has the excess risk of the best estimator in the collection $(\hat{f}_m)_{m\in\mathcal{M}}$. Following [19], this result is called an oracle inequality. We call it *robust* as it holds under moment assumptions on the linear space F (see Assumption 1) and for a dataset that may contain outliers.

The residual term $b_{\varepsilon,V}$ is of order V/N. If $\log|\mathcal{M}| \gtrsim |\mathcal{O}|$ and $V \asymp \log|\mathcal{M}|$, the residual term is of order $\log|\mathcal{M}|/N$ and the deviation probability is of order of $1 - 1/|\mathcal{M}|^\alpha$ for $V = 48\log|\mathcal{M}|^{2+\alpha}$ and $\alpha > 0$, which is minimax optimal according to [51]. Indeed, the rate $\log|\mathcal{M}|/N$ is the price we have to pay in the 'Model Selection' aggregation setup (see Theorem 1 in [51]). This result holds in both expectation and with constant probability, even if it is only stated in expectation in [51]. Indeed, one can for instance apply Theorem 2.5 in [53]. In the last version of the book, Theorem 2.5 in [53] is available with constant deviation, and plugging this result in the proof of Theorem 1 from [51] shows that the conclusion of the theorem still holds with constant deviation. Note that the result in Theorem 1 from [51] (in expectation and the extended version to the constant deviation setup) holds for the regression model $Y = f^*(X) + \zeta$ for

a Gaussian noise ζ independent of X admitting a bounded density uniformly lower bounded by some absolute positive constant over a cube. It does not allow any type of corruption. Our result shows that one can still achieve the rate $\log|\mathcal{M}|/N$ with constant probability even though there are up to $\log|\mathcal{M}|$ outliers in the data set and the inliers are heavy-tailed. To be more precise, let us now state Theorem 1 from [51] in the 'constant deviation case': there exists a dictionary $\mathcal{M}$ such that for any aggregation method $\tilde{f}$, there exists f^* bounded by 1 in L_∞ such that with constant probability

$$\left\|\tilde{f}-f^*\right\|^2 \geqslant 2\min_{f\in\mathcal{M}}\|f-f^*\|^2 + c_0\frac{\log|\mathcal{M}|}{N}$$

where $c_0>0$ is some absolute constant. Note that in Theorem 1 from [51] the factor in front of $\min_{f\in\mathcal{M}}\|f-f^*\|^2$ is 1, however in the proof of this result this term equals zero so that one can choose any constant in front of this term, here we took it equal to 2.

The oracle inequality from Theorem 3.2 is interesting when $a_{\varepsilon,V}<1$ which holds if $\chi\lesssim\sqrt{N/V}$. The "constant" χ in Assumption 1 may therefore grow with the dimension of F as in the examples of [48] without breaking the results.

The result can also be compared with more standard results in model selection. These results usually compare the excess risk of the selected estimator $\hat{f}_{\hat{m}}$ with the infimum $\inf_{m\in\mathcal{M}}\ell(\hat{f}_m)$, where the estimators $\hat{f}_m$ are built using all the dataset. By comparison here, the base estimators are only trained using part of the data so the performance of the oracle is likely to be less precise when the dataset is clean. In the model selection literature see for example [39], the leading constant C in front of $\inf_{m\in\mathcal{M}}\ell(\hat{f}_m)$ is absolute and satisfies $C>1$. Here, the leading constant $1+3a_{\varepsilon,V}$ can be as close to 1 as desired, provided that χ^4V/N and ε are asymptotically negligible. To the best of our knowledge, "optimal" oracle inequalities with $C=1$ that works for any type of weak learners have only been obtained when the estimators are built with an independent dataset, using fresh (clean) data to make the selection, as this is the case, for example in [51]. An important difference with our construction here is therefore that the estimators can be considered as fixed points by conditioning on the training data and the problem reduces to a selection procedure among fixed points. Here, the estimators can be built with any data and any data can be used for selection, which makes this simplification trick impossible. The situation here can somehow be more easily compared with results obtained for cross-validation as in [2, 3, 38] for example, where typically all data are used both for training and selection. In the literature on the theory of cross-validation that we are aware of, apart in the very specific setting of L_2 density estimation, the oracle inequalities compare the excess risk of the selected estimator with the infimum of the original estimators trained on a strict subset of data with a leading constant that is at best asymptotically 1, as we do in Theorem 3.2.

4. An ensemble method to induce robustness, subsampling and hyperparameters tuning

In this section, we define an *ensemble method* which takes one or several (non necessarily robust) algorithms as input and outputs an estimator. The method is robust to the presence of outliers, has subsampling capabilities, and automatically tune hyperparameters. The main ideas behind the construction are: using subsampling as a way of achieving robustness (by avoiding outliers), viewing the choice of the subsample as a hyperparameter to be tuned, and using the robust selection procedure from Section 3 to select the final estimator. Performance guaranties are established in Corollary 4.1.

4.1. Definition of the method

Let $(G_\lambda)_{\lambda\in\Lambda}$ be a finite collection of learning algorithms which outputs estimators in F. The collection may in fact correspond to a single algorithm with several combinations of hyperparameters values, or even several different algorithms with several combinations of hyperparameters values.

We now construct the set $\mathcal{B}$ of subsamples to be considered by the method. Assume $N \geqslant 8$. Let $K_{\min}$ and $K_{\max}$ integers such that $3 \leqslant K_{\min} \leqslant K_{\max} \leqslant \log_2 N$ (these parameters will specify the subsamples cardinality range). For each $K \in [\![K_{\min},\ K_{\max}]\!]$, consider a partition $(B_k^{(K)})_{k\in[2^K]}$ of $[N]$ such that for all $k \in [2^K]$, $\lfloor N/2^K \rfloor \leqslant \left|B_k^{(K)}\right|$. We call $(B_k^{(K)})_{k\in[2^K]}$ *the* 2^K*-partition*. Let

$$\mathcal{B} = \bigcup_{K=K_{\min}}^{K_{\max}} \bigcup_{k\in[2^K]} \left\{B_k^{(K)}\right\} \quad \text{and} \quad \mathcal{M} = \Lambda \times \mathcal{B}. \tag{4.1}$$

From $K_{\min} \geqslant 3$ we can easily deduce that each subsample in $\mathcal{B}$ has cardinality less than $N/4$. Then, as in Section 3, for each $m = (\lambda, B)$, let $\hat{f}_m$ be the estimator trained by algorithm G_λ using subsample B, in other words: $\hat{f}_m = G_\lambda((X_i, Y_i)_{i\in B})$. Let $V \in [\![3, N/8]\!]$. For each couple $(m, m') \in \mathcal{M}^2$, let $(T_v^{(m,m')})_{v\in[V]}$ be a partition of a subset of $[N] \setminus (B_m \cup B_{m'})$ such that $\left|T_v^{(m,m')}\right| \geqslant N/4V$ for all $v \in [V]$. Then, using the minmax-MOM selection procedure (3.2) from Section 3, we select from collection $(\hat{f}_m)_{m\in\mathcal{M}}$ the final estimator $\hat{f}_{\hat{m}}$.

4.2. Theoretical guarantees

The following result states that if risk bounds hold for algorithms $(G_\lambda)_{\lambda\in\Lambda}$ in a context with no-outlier, then $\hat{f}_{\widehat{m}}$ essentially satisfies the best of those risk bounds, even in the presence of outliers.

Corollary 4.1. *Let $\mathcal{M}$ be defined by (4.1). Grant Assumption 1. Let $\rho : \Lambda \times \mathbb{N}^* \to \mathbb{R}_+ \cup \{+\infty\}$ be a non-increasing function in its second variable and $\nu : \Lambda \to$*

$\mathbb{R}_+^*$. *Denote* $\nu_{max} := \lceil \max_{\lambda\in\Lambda} \nu(\lambda) \rceil$. *Assume that* $N \geqslant \nu_{max} \max(8V, 2^{K_{min}+1})$ *and* $V \in [\![3\,|\mathcal{O}|\,, 2^{K_{max}-1}]\!]$. *Assume that, for all* $\lambda \in \Lambda$ *and* $B \subset \mathcal{I}$ *such that* $|B| \geqslant \nu(\lambda)$, *it holds that* $\ell(\hat{f}_{\lambda,B}) \leqslant \rho(\lambda, |B|)$ *with probability larger than* $1-\exp(-1/48)$. *Then for all* $\varepsilon > 0$, *the estimator* $\hat{f}_{\widehat{m}}$ *defined in* (3.2) *satisfies*

$$(1 - a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \min_{\lambda\in\Lambda} \rho\left(\lambda, \left\lfloor \frac{N}{\max(4V, 2^{K_{min}})} \right\rfloor\right) + 2b_{\varepsilon,V} \tag{4.2}$$

with probability larger than

$$1 - (|\Lambda|^2 N^2 + 1)e^{-V/48}. \tag{4.3}$$

Corollary 4.1 is proved in Section B.2. Let us stress some important aspects.

Estimators $\hat{f}_{\lambda,B}$ for $(\lambda, B) \in \mathcal{M}$ are assumed to satisfy an excess risk bound with rates $\rho(\lambda, |B|)$ only with constant probability (the constant $1-\exp(-1/48)$ chosen here has nothing special), when B is large enough and only contains informative data. For example, this condition is met by ERM when informative data satisfy moment assumptions, see Propositions 5.2 and 6.1. With these arguably weak requirement, the above statement claims that the ensemble method achieves the best bound among $\rho(\lambda, \lfloor N/\max(4V, 2^{K_{\min}}) \rfloor)$, $\lambda \in \Lambda$ with exponentially large probability (4.3).

The upper bound $\rho(\lambda, |B|)$ on the excess risk of $\hat{f}_{\lambda,B}$ depends on λ and the size $|B|$ of the subsample. It improves with the sample size by the monotonicity assumption on ρ.

Finally, the function $\lambda \mapsto \nu(\lambda)$ is introduced to handle situations where the risk bound holds only when the sample size is larger than $\nu(\lambda)$.

4.3. An efficient partition scheme of the dataset

The partitions $(B_k^{(K)})_{k\in[2^K]}$ ($K \in [\![K_{\min}, K_{\max}]\!]$) and $(T_v^{(m,m')})_{v\in[V]}$ (for $(m, m') \in \mathcal{M}^2$) can be constructed in many different ways. This section presents a specific choice for those partitions which yields a computational advantage by significantly reducing the number of empirical risks $P_{T_v^{(m,m')}}[\gamma(\hat{f}_m)]$ to be computed (by making many of them redundant). This complexity reduction makes the computations from Section 7 possible in a reasonable amount of time.

The minmax-MOM selection procedure (3.2) requires, for all $(m, m') \in \mathcal{M}^2$ and $v \in [V]$, the computation of $P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m)\right]$ and $P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_{m'})\right]$. Since the partition $(T_v^{(m,m')})_{v\in[V]}$ may be different for each couple $(m, m') \in \mathcal{M}^2$, this requires, in the worst case, the computation of $V\,|\mathcal{M}|^2$ empirical risks. By comparison, the construction presented here will only require the computation of $8V|\mathcal{M}|/3$ empirical risks.

For $K \in [\![3,\ \lfloor \log_2 N \rfloor]\!]$ and $k \in [2^K]$, define $B_k^{(K)} := \left[\!\left[\left\lfloor \frac{(k-1)N}{2^K} \right\rfloor,\ \left\lfloor \frac{kN}{2^K} \right\rfloor \right]\!\right]$. For each $K \in [\![3,\ \lfloor \log_2 N \rfloor]\!]$, $(B_k^{(K)})_{k\in[2^K]}$ is a partition of $[N]$ such that, for each $k \in [2^K]$, $\lfloor N/2^K \rfloor \leqslant \left|B_k^{(K)}\right| \leqslant N/4$, as required. Moreover, the following key property holds.

Lemma 4.2. *Let* $3 \leqslant K' \leqslant K \leqslant \lfloor \log_2 N \rfloor$.

(i) For all $k \in [2^K]$, $B_k^{(K)} \subset B^{(K')}_{\lfloor (k-1)2^{K'-K} \rfloor + 1}$.

(ii) For all $k' \in [2^{K'}]$, $(B_k^{(K)})_{k \in [\![(k'-1)2^{K-K'},\ k'2^{K-K'}]\!]}$ *is a partition of* $B_{k'}^{(K')}$.

Let

$$K_0 := \lceil \log_2(V/3) \rceil + 2. \tag{4.4}$$

For all $K_1, K_2 \in [\![3,\ \lfloor \log_2 N \rfloor]\!]$ and $k_1 \in \left[2^{K_1}\right]$, $k_2 \in \left[2^{K_2}\right]$, let $\mathcal{K}_0(K_1, k_2, K_2, k_2)$ be the set of indices from the 2^{K_0}-partition which have empty intersection with both $B_{k_1}^{(K_1)}$ and $B_{k_1}^{(K_2)}$:

$$\mathcal{K}_0(K_1, k_1, K_2, k_2) := \left\{ k \in \left[2^{K_0}\right] \,\middle|\, B_k^{(K_0)} \cap (B_{k_1}^{(K_1)} \cup B_{k_2}^{(K_2)}) = \varnothing \right\}.$$

Lemma 4.3. *For all* $3 \leqslant K_1, K_2 \leqslant \lfloor \log_2 N \rfloor$ *and* $k_1 \in \left[2^{K_1}\right]$, $k_2 \in \left[2^{K_2}\right]$, *we have* $|\mathcal{K}_0(K_1, k_1, K_2, k_2)| \geqslant V$.

Let $(m, m') \in \mathcal{M}^2$ and K_1, k_1, K_2, k_2 be such that $B_m = B_{k_1}^{(K_1)}$ and $B_{m'} = B_{k_2}^{(K_2)}$. Then, the collection of sets $(B_k^{(K_0)})_{k \in \mathcal{K}_0(K_1,k_1,K_2,k_2)}$ is a sub-collection of the 2^{K_0}-partition, whose sets have empty intersection with both B_m and $B_{m'}$, and which, according to Lemma 4.3, contains at least V sets. We can thus define $(T_v^{(m,m')})_{v \in [V]}$ as a sub-collection of size exactly V. Consequently, $(T_v^{(m,m')})_{v \in [V]}$ is indeed a partition of a subset of $[N] \setminus (B_m \cup B_{m'})$. Moreover, we have the following lower bound on the cardinality of its sets, which is required (see Section 3).

Lemma 4.4. *For all* $(m, m') \in \mathcal{M}^2$ *and* $v \in [V]$, $\left| T_v^{(m,m')} \right| \geqslant N/(4V)$.

Consequently, to compute the minmax-MOM selection procedure in the context of the ensemble method defined in Section 4.1, the empirical risk of each estimator $\hat{f}_m$ has to be computed on the 2^{K_0}-partition only, which thanks to (4.4) means the computation of at most $8V|\mathcal{M}|/3$ empirical risks, as advertised.

5. Application to fine-tuning the regularization parameter of the LASSO

This section applies the ensemble method from Section 4 with the LASSO as input procedure. Consider here $\mathbb{X} = \mathbb{R}^d$, and denote $\hat{\beta}_{\lambda,B}$ the LASSO estimator trained with regularization parameter λ and subsample B:

$$\hat{\beta}_{\lambda,B} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{|B|} \sum_{i \in B} (Y_i - \langle \beta, X_i \rangle)^2 + \lambda \|\beta\|_1 \right\}. \tag{5.1}$$

Statistical guarantees for the LASSO, which we recall below, have been obtained in Theorem 1.4 in [32] with a regularization parameter independent of s. However, as mentionned in the proof of Theorem 1.4 in [32] if one knows an upper

bound s on the sparsity of the signal then one can take $\lambda \asymp \|\zeta\|_{L_q} \sqrt{\log(ed/s)/N}$ instead of the classical choice $\|\zeta\|_{L_q} \sqrt{\log(ed)/N}$. In that case, one can get better results valid under the following assumption.

Assumption 2. *Let $(X,Y) \sim P$. For all $t \in \mathbb{R}^d, \mathbb{E}\langle X,t\rangle^2 = \|t\|_2^2$ and there exists $L > 0$ such that, for all $p \geqslant 1$ and $t \in \mathbb{R}^d, (\mathbb{E}|\langle X,t\rangle|^p)^{1/p} \leqslant L\|t\|_2$. Moreover, there exists $q > 2$ such that, for $\beta^* \in \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}(Y - \langle X,\beta\rangle)^2$, $\zeta := Y - \langle X,\beta^*\rangle \in L_q$.*

Under Assumption 2 and if one has some a priori knowledge on the sparsity of β^* then it is possible to apply a remark in the proof of Theorem 1.4 in [32] to get the following constant probability result for the LASSO with a sparsity dependent regularization parameter.

Proposition 5.1. *[Proof of Theorem 1.4 in [32]] Grant Assumption 2. Assume that β^* is s_0-sparse for some $s_0 \in [d]$. Let $B \subset \mathcal{I}$ be such that $|B| \geqslant s_0 \log(ed/s_0)$. Then, there exist absolute constants c_0 and c_1 such that the LASSO with regularization parameter $\lambda = c_0 \|\zeta\|_{L_q} \sqrt{\log(ed/s_0)\,|B|^{-1}}$ satisfies, with probability at least $1 - \exp(-1/48)$,*

$$\ell(\hat{\beta}_{\lambda,B}) \leqslant c_1 \|\zeta\|_{L_q}^2 \frac{s_0 \log(ed/s_0)}{|B|} \quad . \tag{5.2}$$

Proposition 5.1 is an (exact) oracle inequality with optimal residual term [10]. It is satisfied by the LASSO with a constant probability when trained on a set of informative data and for an optimal choice of regularization parameter $\lambda \sim \|\zeta\|_{L_q} \sqrt{\log(ed/s_0)/N}$.

Proposition 5.1 is interesting because it shows that the optimal minimax rate $s_0 \log(ed/s_0)/|B|$ can be achieved by the LASSO even in the heavy-tailed noise setup of Assumption 2. However, it has three drawbacks: 1) the choice of the regularization parameter requires the knowledge of the sparsity s_0 (without this choice, the LASSO achieves the rate $s_0 \log(ed)/|B|$ – there is therefore an adaptation to s_0 problem for the LASSO as raised and solved in [10] using a Lepski's method); 2) the result holds only with constant probability (it is because the choice of λ in Theorem 1.4 from [32] depends on the deviation parameter δ in a multiplicative way, so that we have to take it equal to a constant to reccover the minimax rate); 3) Finally, the LASSO has to be trained with uncorrupted data; a single outlier completely breaks down its statistical properties—see Figure 1 in [28]. We now show how to use the Minmax MOM selection procedure to overcome these weak points.

Let us now combine Corollary 4.1 and Proposition 5.1 to apply the ensemble method to this example. Let $V, K_{\min}, K_{\max}$ satisfy the assumptions from Section 4 and Corollary 4.1. Denote by s^* the largest integer s such that $N/\max(8V, 2^{K_{\min}+1}) \geqslant s\log(ed/s)$, and assume $1 \leqslant \|\beta^*\|_0 \leqslant s^*$ (where $\|\cdot\|_0$ denotes the number of nonzero coefficients). Consider the set of subsamples $\mathcal{B}$ defined as in Section 4.1, the set $\Lambda := \left\{c_0 \|\zeta\|_{L_q} \sqrt{\log(ed/s)}\right\}_{s\in[s^*]}$, and

$\mathcal{M} = \Lambda \times \mathcal{B}$. For $m = (\lambda, B) \in \mathcal{M}$, consider the corresponding estimator $\hat{f}_m := \hat{\beta}_{\lambda/\sqrt{|B|},B}$.

Corollary 5.2. *Grant Assumptions 1 and 2. Let $\hat{m}$ be the output of the ensemble method from Section 4.1. Then, with probability at least $1 - ((s^*)^2 N^2 + 1)\exp(-V/48)$, for all $\varepsilon > 0$,*

$$(1 - a_{\varepsilon,V})\ell(\hat{\beta}_{\hat{\lambda},\hat{B}}) \leqslant (1 + 3a_{\varepsilon,V})c_1 \|\zeta\|_{L_q}^2 \frac{\|\beta^*\|_0 \log(ed\|\beta^*\|_0^{-1})}{\lfloor N/\max(4V, 2^{K_{min}})\rfloor} + 2b_{\varepsilon,V}.$$

While Proposition 5.1 shows statistical guarantee with constant probability for the estimators $\hat{\beta}_{\lambda,B}$ trained on uncorrupted data, Corollary 5.2 shows that the ensemble method improves the constant probability into an exponential probability, allows $|\mathcal{O}|$ outliers as long as $V \geqslant 3|\mathcal{O}|$ and selects the best hyperparameter λ. The proof is given in Appendix B.6. A similar application to the ERM in linear aggregation is given in Appendix 6.

6. Application to ERM and linear aggregation

This section applies Corollary 4.1 by considering non-robust linear aggregation as input algorithms. Let $(F_\lambda)_{\lambda\in\Lambda}$ be a finite collection of subspaces of F, typically spanned by previous estimators. For each $\lambda \in \Lambda$, denote by d_λ the dimension of F_λ and by f_λ^* an oracle in F_λ, meaning $f_\lambda^* := \arg\min_{f\in F_\lambda} R(f)$. Denote $\hat{f}_{\lambda,B}$ the empirical risk minimizer (ERM) on F_λ trained with subsample B:

$$\hat{f}_{\lambda,B} := \arg\min_{f\in F_\lambda} \frac{1}{|B|}\sum_{i\in B}(Y_i - f(X_i))^2. \tag{6.1}$$

The performance of ERM in linear aggregation like $\hat{f}_{\lambda,B}$ under a L_4/L_2 assumption such as Assumption 1 have been obtained in [31].

Proposition 6.1 (Theorem 1.3 in [31])**.** *Let $\lambda \in \Lambda$. Assume that there exists $\chi_\lambda > 0$ such that for all $f \in F_\lambda$, $(Pf^4)^{1/4} \leqslant \chi_\lambda (Pf^2)^{1/2}$. Denote $\zeta_\lambda := Y - f_\lambda^*(X)$ and assume that $(P\zeta_\lambda^4)^{1/4} \leqslant \sigma_\lambda$. Let $B \subset \mathcal{I}$ be such that $|B| \geqslant (1600\chi_\lambda^4)^2 d_\lambda$. Then, for every $x > 0$, with probability larger than $1 - \exp(-|B|/(64\chi_\lambda^8)) - 1/x$, the ERM $\hat{f}_{\lambda,B}$ defined in (6.1) satisfies*

$$\ell(\hat{f}_{\lambda,B}) \leqslant \ell(f_\lambda^*) + (256)^2 \chi_\lambda^{12} \frac{\sigma_\lambda^2 d_\lambda x}{|B|}.$$

In Proposition 6.1, the (exact) oracle inequality satisfied by $\hat{f}_{\lambda,B}$ guarantees an optimal residual term of order $\sigma_\lambda^2 d_\lambda/N$ only when the deviation parameter x is constant. This may seem weak, but it cannot be improved in general—see Proposition 1.5 in [31]: ERM is not robust to "stochastic outliers" in general.

We can now combine these algorithms with our ensemble method. Let $\mathcal{M} = \Lambda \times \mathcal{B}$ be as in (4.1) and V, $K_{\min}$ and $K_{\max}$ satisfy the assumptions from Section 4 and Corollary 4.1. Consider the output estimator $\hat{f}_{\widehat{m}}$ from (3.2). The following result combines Corollary 4.1 and Proposition 6.1.

Corollary 6.2. *Grant Assumption 1 on F and assume that for all $\lambda \in \Lambda$ and all $f \in F_\lambda$, $(Pf^4)^{1/4} \leqslant \chi_\lambda (Pf^2)^{1/2}$ and $(P\zeta_\lambda^4)^{1/4} \leqslant \sigma_\lambda$ for $\zeta_\lambda := Y - f_\lambda^*(X)$. Assume also that $N \geqslant \max_{\lambda\in\Lambda}(1600\chi_\lambda^4)^2 d_\lambda \max(8V, 2^{K_{min}+1})$. Then, with probability at least $1 - (|\Lambda|^2 N^2 + 1)\exp(-V/48)$, for all $\varepsilon > 0$,*

$$(1 - a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \times \min_{\lambda\in\Lambda} \left\{ \ell(f_\lambda^*) + 2\exp(1/48)(256)^2 \chi_\lambda^{12} \frac{\sigma_\lambda^2 d_\lambda}{\lfloor N/\max(4V, 2^{K_{min}})\rfloor} \right\} + 2b_{\varepsilon,V}.$$

Proof. The proof follows from Proposition 6.1 and Corollary 4.1. Let us check the assumption and the features of both results. For $x = 2\exp(1/48)$ and when $|B| \geqslant (1600\chi_\lambda^4)^2 d_\lambda$ we have $1 - \exp(-|B|/(64\chi_\lambda^8)) - 1/x \geqslant 1 - \exp(-1/48)$ therefore, $\hat{f}_{\lambda,B}$ satisfies an (exact) oracle inequality with probability larger than $1 - \exp(-1/48)$ when $|B| \geqslant \nu(\lambda) := (1600\chi_\lambda^4)^2 d_\lambda$ with a residual term given by

$$\rho(\lambda, |B|) = \ell(f_\lambda^*) + 2(256)^2 \exp(1/48)\chi_\lambda^{12} \frac{\sigma_\lambda^2 d_\lambda}{|B|}.$$

Therefore, all the condition of Corollary 4.1 are satisfied and the result follows from a direct application of the latter result. ∎

7. Numerical experiments with the LASSO

7.1. Presentation

In this section, the ensemble method from Section 4 is implemented and fed with the LASSO algorithm, as in Section 5. Numerical experiments are performed with various amount and *types* of outliers in order to investigate their effects on the output estimator $\hat{f}_{\hat{m}}$ and the corresponding parameter $(\lambda_{\hat{m}}, B_{\hat{m}})$.

We consider a framework with 2000 features, i.e. $\mathbb{X} = \mathbb{R}^{2000}$ and let $\beta_0 \in \mathbb{R}^{2000}$ which we assume 20-sparse. The datasets are of size $N = 1000$ and we consider the following numbers of outliers $|\mathcal{O}| = 0, 4, 8 \ldots, 150$. We construct two types of outliers ($\mathcal{O} = \mathcal{O}_1 \sqcup \mathcal{O}_2$), both of which are present in equal amount ($|\mathcal{O}_1| = |\mathcal{O}_2|$). The first type, which we call *hard outliers* are defined to simulate corruption due, for instance, to hardware issues: $X_i = (1, \ldots, 1) \in \mathbb{R}^{2000}$, $Y_i = 10000$, $i \in \mathcal{O}_1$, and second type, which we call *heavy-tail outliers* are constructed as $X_i \sim \mathcal{N}(0, I_{2000})$, $Y_i = \langle X_i, \beta_0\rangle + \zeta_i$, $i \in \mathcal{O}_2$, where the variables $(X_i, Y_i)_{i\in\mathcal{O}_2}$ are i.i.d., ζ_i is a noise independent of X_i and distributed according to Student's t-distribution with 2 degrees of freedom. Informative data is drawn according to $X_i \sim \mathcal{N}(0, I_{2000})$, $Y_i = \langle X_i, \beta_0\rangle + \zeta_i$, $i \in \mathcal{I}$, where variables $(X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d., and ζ_i ($i \in \mathcal{I}$) is a standard Gaussian noise independent of X_i. On the one hand, a *hard outlier*, if contained in the training sample of an estimator, is likely to significantly deteriorate its performance. On the other hand, *heavy-tail outliers* only differ from informative data in the distribution of the noise, and should not deteriorate too much the performance of affected estimators.

Nevertheless, we expect the informative data to be preferred over the type 2 outliers in the selected subsample $B_{\hat{m}}$ (this is indeed the case in Figure 1c).

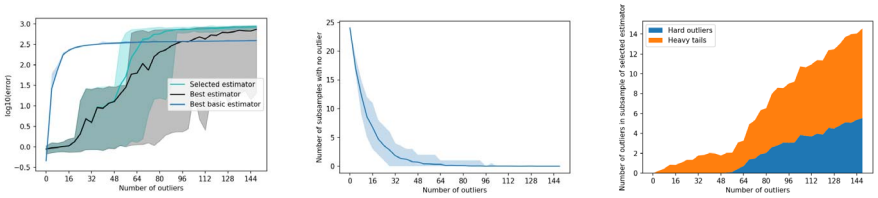
We consider $\Lambda = \left\{ e^k \,\middle|\, k \in \frac{1}{2}[\![-2, 4]\!] \right\}$ as the grid of values for the regularization parameter of the LASSO. We implement the ensemble method from Section 4 with parameters $V = 40$, $K_{\min} = 3$ and $K_{\max} = 4$. The set $\mathcal{B}$ of subsamples is constructed as in Section 4.3 and we set $\mathcal{M} = \Lambda \times \mathcal{B}$. For each $m = (\lambda, B) \in \mathcal{M}$, we train the LASSO estimator $\hat{\beta}_m$ with hyperparameter λ and subsample B (see (5.1)). We then compute the output estimator $\hat{\beta}_{\widehat{m}}$, which uses partitions $(T_v^{(m,m')})_{v\in[V]}$ (for $m, m' \in \mathcal{M}$) constructed as in Section 4.3. Let us denote $\hat{\beta}_{\tilde{m}}$ the best oracle estimator among $(\hat{\beta}_m)_{m\in\mathcal{M}}$, in other words, let $\tilde{m} := \arg\min_{m\in\mathcal{M}} R(\hat{\beta}_m)$ where $\beta \mapsto R(\beta) = \|\beta - \beta_0\|_2^2$ is the true risk function which is not known—so that $\tilde{m}$ cannot be computed using only the data. For comparison, we also compute the LASSO estimators $\hat{\beta}_{\lambda,[N]}$ trained with the whole dataset, which we will call *basic estimators*, and let $\hat{\beta}_{\tilde{\lambda},[N]}$ be the best among those, so that $\tilde{\lambda} := \arg\min_{\lambda\in\Lambda} R(\hat{\beta}_{\lambda,[N]})$.

7.2. On the choices of V and K_{max}

The choices of V and $K_{\max}$ have an impact on both the performance the output estimator and the computation time. The higher is V, the higher is the number of outliers that the MOM-selection procedure (3.2) can handle, and as a matter of fact, Theorem 3.2 requires $V \geqslant 3|\mathcal{O}|$. However, higher values of V increase computation time and deteriorates the statistical guarantee (through the values of $a_{\varepsilon,V}$ and $b_{\varepsilon,V}$ from the statement of Theorem 3.2). Here we choose $V = 40$, so we can expect the minmax-MOM selection procedure to perform well at least up until we get as many as $\lfloor 40/3 \rfloor = 13$ outliers. The number of considered subsamples is increasing with $K_{\max}$. High values of $K_{\max}$ increase computation time. Moreover, we don't want to go for the maximum value $K_{\max} = \lceil \log_2 N \rceil$, which would imply the training of estimators with subsamples of size 2, which is irrelevant. We therefore want a low value of $K_{\max}$, but we would like to have at least one subsample which contains no outlier. This is necessarily the case when $2^{K_{\max}} > |\mathcal{O}|$. Since the choice of $V = 40$ allows to hope for a good selection performance up to $|\mathcal{O}| = 13$ outliers, we choose $K_{\max} = 4$ which indeed satisfies $2^{K_{\max}} > |\mathcal{O}|$.

7.3. Results and discussion

The plots presented in Figure 1 are averaged over 100 experiments. Figure 1a shows estimation error rates against the number $|\mathcal{O}|$ of outliers in the dataset and a 95% confidence interval 1) of the output estimator $\hat{\beta}_{\widehat{m}}$ of the ensemble method, 2) of $\hat{\beta}_{\tilde{m}}$, the best estimator among $(\hat{\beta}_m)_{m\in\mathcal{M}}$, and 3) of the best *basic* estimator $\hat{\beta}_{\tilde{\lambda},[N]}$. As soon as the dataset contains outliers, basic estimators $(\hat{\beta}_{\lambda,[N]})_{\lambda\in\Lambda}$ have larger errors than $\hat{\beta}_{\tilde{m}}$ (the best estimator computed on a subsample). For $|\mathcal{O}| \leqslant 48$ the ensemble method procedure has the same error as the best estimator among $(\hat{\beta}_m)_{m\in\mathcal{M}}$.



(a) Error of estimators $\hat{\beta}_{\widehat{m}}$, $\hat{\beta}_{\widetilde{m}}$ and $\hat{\beta}_{\widehat{\lambda},[N]}$

(b) Number of subsamples with no outlier

(c) Number of outliers in $B_{\widehat{m}}$

FIG 1. *ensemble method run with* $N = 1000$, $V = 40$, $K_{max} = 4$, *and averaged over 200 experiments.*

For a given value of parameter V, the minmax-MOM selection procedure (3.2) is expected to fail at some point when the number of outliers increases, but it seems here to resist to a much higher number of outliers than predicted by the theory. Theorem 3.2 holds for $|\mathcal{O}| \leqslant V/3$, that is $|\mathcal{O}| \leqslant 13$ here. It seems here that the minmax-MOM selection procedure performs satisfactorily for $|\mathcal{O}| \leqslant 48$ and even selects a reasonably good estimator for $|\mathcal{O}| \leqslant 56$.

Figure 1c shows the number of each type of outliers in the selected subsample $B_{\widehat{m}}$. The method manages to rule out hard outliers when $|\mathcal{O}| \leqslant 48$, and the output estimator $\hat{\beta}_{\widehat{m}}$ has in these cases minimal risk, as the best estimator $\hat{\beta}_{\widetilde{m}}$. Figure 1b also shows that almost all subsample contain outliers when $|\mathcal{O}| \geqslant 48$. Besides, the selected subsample $B_{\widehat{m}}$ contains heavy-tail outliers even for small values of $|\mathcal{O}|$. As heavy-tail outliers and informative data define the same oracle, these heavy-tail outliers are actually informative for the learning task and the minmax-MOM selection procedure use this extra information automatically in an optimal way. In particular, the ensemble method distinguishes between non-informative hard outliers and possibly informative heavy-tailed outliers.

Overall, our method shows very strong robustness to the presence of outliers and outputs an estimator with the best possible performance among the given class of estimators.

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996). MR1688610

[2] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010. MR2602303

[3] S. Arlot and M. Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, 17:Paper No. 208, 50, 2016. MR3595142

[4] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011. MR2906886

[5] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017. MR3595933
[6] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011. MR2834722
[7] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1092–1119, 2014. MR3224300
[8] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. MR1679028
[9] Pierre C. Bellec. Optimal bounds for aggregation of affine estimators. *Ann. Statist.*, 46(1):30–59, 2018. MR3766945
[10] Pierre C Bellec, Guillaume Lecué, Alexandre B Tsybakov, et al. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018. MR3852663
[11] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997. MR1462939
[12] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. MR1848946
[13] Jelena Bradic. Randomized maximum-contrast selection: subagging for large-scale regression. *Electronic Journal of Statistics*, 10(1):121–170, 2016. MR3466179
[14] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.
[15] Peter Bühlmann. Bagging, subagging and bragging for improving some prediction algorithms. In *Recent advances and trends in nonparametric statistics*, pages 19–34. Elsevier B. V., Amsterdam, 2003. MR2498230
[16] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920
[17] Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014. MR3308656
[18] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
[19] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. MR1311089
[20] J. Fan, Q. Li, and Y. Wang. Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions. *Journal of Royal Statistical Society B*, 79:247–265, 2017. MR3597972
[21] A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 11 2008. MR2543590
[22] Matthew J Holland. Classification using margin pursuit. *arXiv preprint 1810.04863*, 2018.

[23] Matthew J Holland. Robust descent using smoothed multiplicative noise. *arXiv preprint 1810.06207*, 2018.
[24] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986. MR0855970
[25] Michael I. Jordan. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013. MR3102908
[26] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014. 1112.5016. MR3248677
[27] G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud, 2017. MR4013866
[28] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. Technical report, CNRS, ENSAE, Paris-sud, 2017.
[29] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint 1711.10306*, 2017. MR4102681
[30] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via mom minimization. *arXiv preprint 1808.03106*, 2018. MR4137195
[31] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016. MR3474824
[32] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical report, CNRS, ENSAE and Technion, I.I.T., 2016. MR3782379
[33] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991. MR1147167
[34] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint 1112.3914*, 2011.
[35] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. MR2242356
[36] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS*, 2016. MR4055993
[37] Lester Mackey, Ameet Talwalkar, and Michael I. Jordan. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 16:913–960, 2015. MR3361307
[38] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *arXiv preprint 1909.04890*, 2019.
[39] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007. MR2319879
[40] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006. MR2291502

[41] Shahar Mendelson. Learning without concentration. In *Proceedings of the 27th annual conference on Learning Theory COLT14*, pages pp 25–39. 2014. MR3367000

[42] Arkadii Nemirovski. *Lectures on probability theory and statistics*, volume 1738 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998, Edited by Pierre Bernard. MR1775638

[43] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. MR0702836

[44] Roberto I Oliveira and Philip Thompson. Sample average approximation with heavier tails i: non-asymptotic bounds with weak assumptions and stochastic constraints. *arXiv preprint 1705.00822*, 2017.

[45] Roberto I Oliveira and Philip Thompson. Sample average approximation with heavier tails ii: localization in stochastic convex optimization and persistence results for the lasso. *arXiv preprint 1711.04734*, 2017.

[46] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint 1802.06485*, 2018. MR4112778

[47] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007. MR2356821

[48] Adrien Saumard et al. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018. MR3757527

[49] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981. MR0630098

[50] Q. Sun, W.-X. Zhou, and J. Fan. Adaptive huber regression: Optimality and phase transition. *Preprint available in 1706.06991*, 2017.

[51] Alexandre B. Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.

[52] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. MR2051002

[53] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359

[54] A. B. Yuditskiĭ, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96, 2005. MR2198228

[55] W.-X. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *To appear in Ann. Statist*, 2017. MR3845005

Appendix A: Lemmas and proofs

Let $\mathcal{V}^{(m,m')} := \left\{ v \in [V] \,\middle|\, T_v^{(m,m')} \subset \mathcal{I} \right\}$ denote the set of indices of blocks from the partition $(T_v^{(m,m')} : v \in [V])$ containing only informative data. In particular, we have

$$\left| \mathcal{V}^{(m,m')} \right| \geqslant V - |\mathcal{O}| \,. \tag{A.1}$$

Lemma A.1. *Let $m, m' \in \mathcal{M}$ and $v \in \mathcal{V}^{(m,m')}$. The conditional variance of random variable $P_{T_v^{(m,m')}} \left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}) \right]$ given random variables $(X_i, Y_i)_{i \in B_m \cup B_{m'}}$ is bounded from above as:*

$$\operatorname{Var}\left(P_{T_v^{(m,m')}} \left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}) \right] \,\middle|\, (X_i, Y_i)_{i \in B_m \cup B_{m'}} \right) \leqslant C_{m,m'},$$

where

$$C_{m,m'} := \frac{V}{N} \left(16\chi^4 \left(\ell(\hat{f}_m)^2 + \ell(\hat{f}_{m'})^2 \right) + 64\sigma^2 \left(\ell(\hat{f}_m) + \ell(\hat{f}_{m'}) \right) \right).$$

Proof. By assumption, random variables $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent. In particular, random variables $(X_i, Y_i)_{i \in T_v^{(m,m')}}$ are independent conditionally to $(X_i, Y_i)_{i \in B_m \cup B_{m'}}$ since $v \in \mathcal{V}^{(m,m')}$. Using the shorthand notation $\operatorname{Var}_{m,m'}(\,\cdot\,) := \operatorname{Var}\left(\cdot \,\middle|\, (X_i, Y_i)_{i \in B_m \cup B_{m'}} \right)$, we have

$$\begin{aligned}
&\operatorname{Var}_{m,m'} \left(P_{T_v^{(m,m')}} \left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}) \right] \right) \\
&= \operatorname{Var}_{m,m'} \left(\frac{1}{\left| T_v^{(m,m')} \right|} \sum_{i \in T_v^{(m,m')}} (\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}))(X_i, Y_i) \right) \\
&= \frac{1}{\left| T_v^{(m,m')} \right|^2} \sum_{i \in T_v^{(m,m')}} \operatorname{Var}_{m,m'} \left((\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}))(X_i, Y_i) \right).
\end{aligned}$$

Fix $i \in T_v^{(m,m')} \subset \mathcal{I}$, and let us bound from above each variance terms from the latter expression:

$$\begin{aligned}
&\operatorname{Var}_{m,m'} \left((\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}))(X_i, Y_i) \right) \\
&\leqslant \mathbb{E} \left[\left((\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}))(X_i, Y_i) \right)^2 \,\middle|\, (X_{i'}, Y_{i'})_{i' \in B_m \cup B_{m'}} \right] \\
&= P \left[\left(\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}) \right)^2 \right] \\
&= P \left[(\gamma(\hat{f}_m) - \gamma(f^*) + \gamma(f^*) - \gamma(\hat{f}_{m'}))^2 \right] \\
&\leqslant 2\, P \left[(\gamma(\hat{f}_m) - \gamma(f^*))^2 \right] + 2\, P \left[(\gamma(\hat{f}_{m'}) - \gamma(f^*))^2 \right],
\end{aligned}$$

where we used the basic inequality $(x+y)^2 \leqslant 2(x^2+y^2)$ in the last inequality. Let us bound from above the first term. The second term is handled similarly. We use a quadratic/multiplier decomposition of the excess loss:

$$P\left[(\gamma(\hat{f}_m)-\gamma(f^*))^2\right] = P\left[\left(\left(\hat{f}_m - f^*\right)^2 - 2(Y-f^*)(\hat{f}_m - f^*)\right)^2\right]$$
$$\leqslant 2P\left[(\hat{f}_m - f^*)^4\right] + 8P\left[(Y-f^*)^2(\hat{f}_m - f^*)^2\right].$$

By Assumption 1, it follows that

$$P\left[(\hat{f}_m - f^*)^4\right] \leqslant \chi^4 \left(P\left[(\hat{f}_m - f^*)^2\right]\right)^2 = \chi^4\ \ell(\hat{f}_m)^2.$$

Likewise, Assumption 1 yields

$$P\left[(Y-f^*)^2(\hat{f}_m - f^*)^2\right] \leqslant \sigma^2 P\left[(\hat{f}_m - f^*)^2\right] = \sigma^2 \ell(\hat{f}_m).$$

The result follows from combining these pieces and using $\left|T_v^{(m,m')}\right| \geqslant N/4V$. ■

Lemma A.2. *With probability higher than* $1-|\mathcal{M}|^2\, e^{-(V-|\mathcal{O}|)/32}$, *for all* $m, m' \in \mathcal{M}$,

$$\ell(\hat{f}_m) - \ell(\hat{f}_{m'}) - \sqrt{8\,C_{m,m'}} \leqslant \mathcal{T}(m,m') \leqslant \ell(\hat{f}_m) - \ell(\hat{f}_{m'}) + \sqrt{8\,C_{m,m'}}$$

where $\mathcal{T}(m,m') := \mathrm{med}_{v\in[V]}\left\{P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\right]\right\}$.

Proof. Fix $m, m' \in \mathcal{M}$ and $v \in \mathcal{V}^{(m,m')}$. Conditionally to $(X_i,Y_i)_{i\in B_m \cup B_{m'}}$, it follows from Chebychev's inequality and Lemma A.1 that, with probability higher than $1-1/8$,

$$\left|P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\right] - (\ell(\hat{f}_m)-\ell(\hat{f}_{m'}))\right|$$
$$\leqslant \sqrt{8\,\mathrm{Var}\left(P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\right]\,\middle|\,(X_i,Y_i)_{i\in B_m\cup B_{m'}}\right)}$$
$$\leqslant \sqrt{8\,C_{m,m'}}.$$

As the probability estimate does not depend on $(X_i,Y_i)_{i\in B_m\cup B_{m'}}$, the above also holds unconditionally and, with probability larger than $1-1/8$,

$$\ell(\hat{f}_m) - \ell(\hat{f}_{m'}) - \sqrt{8\,C_{m,m'}} \leqslant P_{T_v^{(m,m')}}\left[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\right]$$
$$\leqslant \ell(\hat{f}_m) - \ell(\hat{f}_{m'}) + \sqrt{8\,C_{m,m'}}. \tag{A.2}$$

Denote by $\Omega_v^{(m,m')}$ the event defined by (A.2) and see that $\mathbb{P}\left[\Omega_v^{(m,m')}\right] \geqslant 1-1/8$. Apply now Hoeffding's inequality to random variables $\mathbb{1}_{\Omega_v^{(m,m')}}, v \in \mathcal{V}^{(m,m')}$

which are independent conditionally to $(X_i, Y_i)_{B_m \cup B_{m'}}$: on an event $\Omega^{(m,m')}$ of probability larger than $1 - e^{-2\left|\mathcal{V}^{(m,m')}\right|(1/8)^2} \geqslant 1 - e^{-(V-|\mathcal{O}|)/32}$, see (A.1),

$$\begin{aligned}\frac{1}{\left|\mathcal{V}^{(m,m')}\right|} \sum_{v \in \mathcal{V}^{(m,m')}} \mathbb{1}_{\Omega_v^{(m,m')}} &\geqslant \mathbb{E}\left[\frac{1}{\left|\mathcal{V}^{(m,m')}\right|} \sum_{v \in \mathcal{V}^{(m,m')}} \mathbb{1}_{\Omega_v^{(m,m')}}\right] - \frac{1}{8} \\ &= \frac{1}{\left|\mathcal{V}^{(m,m')}\right|} \sum_{v \in \mathcal{V}^{(m,m')}} \mathbb{P}\left[\Omega_v^{(m,m')}\right] - \frac{1}{8} \geqslant \frac{3}{4}.\end{aligned}$$

Then, on $\Omega^{(m,m')}$, using (A.1) and the assumption $V \geqslant 3|\mathcal{O}|$,

$$\sum_{v \in \mathcal{V}^{(m,m')}} \mathbb{1}_{\Omega_v^{(m,m')}} \geqslant \frac{3}{4}\left|\mathcal{V}^{(m,m')}\right| \geqslant \frac{3}{4}\left(V - |\mathcal{O}|\right) \geqslant \frac{V}{2}.$$

In other words, inequalities (A.2) hold for more than half of the indices $v \in [V]$. Therefore, on event $\Omega^{(m,m')}$, the same inequality holds for the median over $v \in [V]$:

$$\ell(\hat{f}_m) - \ell(\hat{f}_{m'}) - \sqrt{8C_{m,m'}} \leqslant \mathcal{T}(m,m') \leqslant \ell(\hat{f}_m) - \ell(\hat{f}_{m'}) + \sqrt{8\,C_{m,m'}}.$$

By a union bound, the above holds for all $m, m' \in \mathcal{M}$ with probability at least $1 - |\mathcal{M}|^2\, e^{-(V-|\mathcal{O}|)/32}$. ∎

Lemma A.3. *For all $m, m' \in \mathcal{M}$, $\varepsilon' > 0$ and $b > 0$,*

$$\sqrt{8\,C_{m,m'}} \leqslant \sqrt{\frac{8V}{N}}\left((4\chi^2 + \varepsilon')(\ell(\hat{f}_m) + \ell(\hat{f}_{m'})) + \frac{16\sigma^2}{\varepsilon'}\right).$$

Proof. By definition of $C_{m,m'}$ and the inequalities $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ and $2xy \leqslant x^2/\varepsilon' + \varepsilon' y^2$,

$$\begin{aligned}\sqrt{8C_{m,m'}} &= \sqrt{\frac{8V}{N}\left(16\chi^4\left(\ell(\hat{f}_m)^2 + \ell(\hat{f}_{m'})^2\right) + 64\sigma^2\left(\ell(\hat{f}_m) + \ell(\hat{f}_{m'})\right)\right)} \\ &\leqslant \sqrt{\frac{8V}{N}}\left(4\chi^2\sqrt{\ell(\hat{f}_m)^2 + \ell(\hat{f}_{m'})^2} + 8\sigma\sqrt{\ell(\hat{f}_m) + \ell(\hat{f}_{m'})}\right), \\ &\leqslant \sqrt{\frac{8V}{N}}\left(4\chi^2(\ell(\hat{f}_m) + \ell(\hat{f}_{m'})) + \frac{16\sigma^2}{\varepsilon'} + \varepsilon'(\ell(\hat{f}_m) + \ell(\hat{f}_{m'}))\right) \\ &= \sqrt{\frac{8V}{N}}\left((4\chi^2 + \varepsilon')(\ell(\hat{f}_m) + \ell(\hat{f}_{m'})) + \frac{16\sigma^2}{\varepsilon'}\right).\end{aligned}$$

∎

Lemma A.4. *With probability at least $1 - |\mathcal{M}|^2\, e^{-(V-|\mathcal{O}|)/32}$, for all $m, m' \in \mathcal{M}$ and $\varepsilon > 0$:*

$$\begin{aligned}&(1 - a_{\varepsilon,V})\ell(\hat{f}_m) - (1 + a_{\varepsilon,V})\ell(\hat{f}_{m'}) - b_{\varepsilon,V} \\ &\leqslant \mathcal{T}(m,m') \leqslant (1 + a_{\varepsilon,V})\ell(\hat{f}_m) - (1 - a_{\varepsilon,V})\ell(\hat{f}_{m'}) + b_{\varepsilon,V}.\end{aligned}$$

Proof. The result follows from Lemmas A.2 and A.3 for $\varepsilon' = \sqrt{N/V}\varepsilon$, together with the definition of $a_{\varepsilon,V}$ and $b_{\varepsilon,V}$. ■

Appendix B: Proofs of the main results

B.1. Proof of Theorem 3.2

Assume that $a_{\varepsilon,V} < 1$, otherwise the result is trivial. Denote

$$m_o := \arg\min_{m\in\mathcal{M}} \ell(\hat{f}_m),$$

so

$$(1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) = (1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) - (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}) + (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}). \quad \text{(B.1)}$$

Let Ω be the event defined by Lemma A.4. Since $V \geqslant 3|\mathcal{O}|$, by Lemma A.4

$$\mathbb{P}(\Omega) \geqslant 1 - |\mathcal{M}|^2 e^{-(V-|\mathcal{O}|)/32} \geqslant 1 - |\mathcal{M}|^2 e^{-V/48} \quad .$$

It follows from Lemma A.4 and (B.1) that, on Ω,

$$(1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant \max_{m\in\mathcal{M}} \mathcal{T}(\hat{m},m) + b_{\varepsilon,V} + (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}). \quad \text{(B.2)}$$

Then, by definition of $\hat{m}$ and using Lemma A.4, on Ω,

$$\begin{aligned}
&\max_{m\in\mathcal{M}} \mathcal{T}(\hat{m},m) \\
&= \min_{m'\in\mathcal{M}} \max_{m\in\mathcal{M}} \mathcal{T}(m',m) \leqslant \max_{m\in\mathcal{M}} \mathcal{T}(m_o,m) \\
&\leqslant \max_{m\in\mathcal{M}} \left\{ (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}) - (1-a_{\varepsilon,V})\ell(\hat{f}_m) + b_{\varepsilon,V} \right\} \\
&= (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}) - (1-a_{\varepsilon,V})\ell(\hat{f}_{m_o}) + b_{\varepsilon,V} = 2a_{\varepsilon,V}\ell(\hat{f}_{m_o}) + b_{\varepsilon,V},
\end{aligned}$$

where we used $1 - a_{\varepsilon,V} \geqslant 0$ and the definition of m_o. Plugging this into (B.2) yields the result.

B.2. Proof of Corollary 4.1

Let

$$K_0 := \max\left(\lceil \log_2(2V) \rceil, K_{\min}\right).$$

It follows from the assumption $V \leqslant 2^{K_{\max}-1}$ that $K_0 \in [\![K_{\min}, K_{\max}]\!]$. Besides, it follows from the above definition that:

$$2^{K_0} \leqslant \max(4V, 2^{K_{\min}}). \quad \text{(B.3)}$$

Let also

$$\lambda_0 := \arg\min_{\lambda\in\Lambda} \rho(\lambda, \lfloor N/2^{K_0}\rfloor) \qquad \text{and} \qquad \rho_0 := \rho(\lambda_0, \lfloor N/2^{K_0}\rfloor).$$
$$\mathcal{K}_0 := \left\{k \in [2^{K_0}] \,\middle|\, B_k^{(K_0)} \subset \mathcal{I}\right\},$$

which is nonempty because $|\mathcal{K}_0| \geqslant 2^{K_0} - |\mathcal{O}| \geqslant 2V - |\mathcal{O}| \geqslant 2V - V/3 \geqslant V$, where the first inequality follows from the definition of $\mathcal{K}_0$, the second inequality from the definition of K_0 and the third inequality from the assumption $V \geqslant 3\,|\mathcal{O}|$.

Consider the events

$$\Omega_1 = \left\{(1 - a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \min_{m\in\mathcal{M}} \ell(\hat{f}_m) + 2b_{\varepsilon,V}\right\}$$
$$\Omega_2 = \left\{\exists k \in \mathcal{K}_0,\ \ell\left(\hat{f}_{\lambda_0, B_k^{(K_0)}}\right) \leqslant \rho_0\right\}.$$

From now on, assume that $\Omega_1 \cap \Omega_2$ hold and the aim is to establish an upper bound on $\min_{m\in\mathcal{M}} \ell(\hat{f}_m)$. Write

$$\min_{m\in\mathcal{M}} \ell(\hat{f}_m) = \min_{\substack{\lambda\in\Lambda\\ B\in\mathcal{B}}} \ell(\hat{f}_{\lambda,B}) \leqslant \min_{k\in\mathcal{K}_0} \ell\left(\hat{f}_{\lambda_0, B_k^{(K_0)}}\right) \leqslant \rho_0 = \min_{\lambda\in\Lambda} \rho(\lambda, \lfloor N/2^{K_0}\rfloor)$$
$$\leqslant \min_{\lambda\in\Lambda} \rho\left(\lambda, \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor\right).$$

Here the second inequality comes from the definition of Ω_2 and the last inequality from (B.3) combined with ρ being nonincreasing in its second variable. Combining the above with the definition of Ω_1 yields the desired inequality:

$$(1 - a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \min_{\lambda\in\Lambda} \rho\left(\lambda, \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor\right) + 2b_{\varepsilon,V}.$$

To conclude the proof, let us bound from below the probability of $\Omega_1 \cap \Omega_2$. By Theorem 3.2,

$$\mathbb{P}[\Omega_1 \cap \Omega_2] = 1 - \mathbb{P}[\Omega_1^c \cup \Omega_2^c] \geqslant 1 - \mathbb{P}[\Omega_1^c] - \mathbb{P}[\Omega_2^c]$$
$$\geqslant 1 - |\mathcal{M}|^2\, e^{-V/48} - \mathbb{P}[\Omega_2^c] \geqslant 1 - |\Lambda|^2\, N^2 e^{-V/48} - \mathbb{P}[\Omega_2^c],$$

Recall that $\lfloor N/2^{K_0}\rfloor \leqslant \left|B_k^{(K_0)}\right|$ for all $k \in \left[2^{K_0}\right]$, that ρ is non-increasing in its second variable and that $B_k^{(K_0)}, k \in \mathcal{K}_0$ are disjoint, so

$$\mathbb{P}[\Omega_2^c] = \mathbb{P}\left[\forall k \in \mathcal{K}_0,\ \ell\left(\hat{f}_{\lambda_0, B_k^{(K_0)}}\right) > \rho_0\right] = \prod_{k\in\mathcal{K}_0} \mathbb{P}\left[\ell\left(\hat{f}_{\lambda_0, B_k^{(K_0)}}\right) > \rho_0\right]$$
$$\leqslant \prod_{k\in\mathcal{K}_0} \mathbb{P}\left[\ell\left(\hat{f}_{\lambda_0, B_k^{(K_0)}}\right) > \rho\left(\lambda_0, \left|B_k^{(K_0)}\right|\right)\right]$$
$$\leqslant \exp(-|\mathcal{K}_0|/48) \leqslant \exp(-V/48)\ .$$

The third inequality follows from the excess risk bound on $\hat{f}_{\lambda_0, B_k^{(K_0)}}$, which holds as soon as $\left|B_k^{(K_0)}\right| \geqslant \nu(\lambda_0)$; this is indeed case because using (B.3):

$$\left|B_k^{(K_0)}\right| \geqslant \left\lfloor \frac{N}{2^{K_0}} \right\rfloor \geqslant \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor = \frac{N}{2\max(2V, 2^{K_{\min}})} \geqslant \nu_{\max} \geqslant \nu(\lambda_0),$$

where the penultimate inequality follows from assumption $N \geqslant \nu_{\max} \max(8V, 2^{K_{\min}+1})$. Therefore, $\mathbb{P}[\Omega_1 \cap \Omega_2] \geqslant 1 - (|\Lambda|^2 N^2 + 1)\exp(-V/48)$.

B.3. Proof of Lemma 4.2

Start with *(i)*. Let $k \in [2^K]$ and $i \in B_k^{(K)}$, which by definition of $B_k^{(K)}$ means:

$$\left\lfloor \frac{(k-1)N}{2^K} \right\rfloor < i \leqslant \left\lfloor \frac{kN}{2^K} \right\rfloor.$$

We can bound from below as follows: let $k' := \lfloor (k-1)(2^{K'-K}) \rfloor + 1$,

$$\left\lfloor \frac{(k-1)N}{2^K} \right\rfloor = \left\lfloor \frac{(k-1)(2^{K'-K})N}{2^{K'}} \right\rfloor \geqslant \left\lfloor \frac{\lfloor (k-1)(2^{K'-K}) \rfloor N}{2^K} \right\rfloor = \left\lfloor \frac{(k'-1)N}{2^K} \right\rfloor.$$

Similarly, the upper bound is obtained as follows:

$$\begin{aligned} \left\lfloor \frac{kN}{2^K} \right\rfloor &= \left\lfloor \frac{((k-1)2^{K'-K} + 2^{K'-K})N}{2^{K'}} \right\rfloor \leqslant \left\lfloor \frac{((k-1)2^{K'-K} + 2)N}{2^{K'}} \right\rfloor \\ &\leqslant \left\lfloor \frac{(\lfloor (k-1)2^{K'-K} \rfloor + 1)N}{2^{K'}} \right\rfloor = \left\lfloor \frac{k'N}{2^{K'}} \right\rfloor. \end{aligned}$$

Therefore,

$$\left\lfloor \frac{(k'-1)N}{2^{K'}} \right\rfloor < i \leqslant \left\lfloor \frac{k'N}{2^{K'}} \right\rfloor.$$

This means $i \in B^{(K')}_{\lfloor (k-1)2^{K'-K} \rfloor + 1}$.

The proof of *(ii)* proceeds as follows. Let $k' \in [2^{K'}]$ and $i \in B_{k'}^{(K')}$, meaning that

$$\left\lfloor \frac{(k'-1)N}{2^{K'}} \right\rfloor < i \leqslant \left\lfloor \frac{k'N}{2^{K'}} \right\rfloor.$$

This can be rewritten as

$$\left\lfloor \frac{2^{K-K'}(k'-1)N}{2^K} \right\rfloor < i \leqslant \left\lfloor \frac{2^{K-K'}k'N}{2^K} \right\rfloor,$$

As $(B_k^{(K)})_{k \in [2^K]}$ is a partition of $[N]$, there exists a unique $k \in [2^K]$ such that $i \in B_k^{(K)}$, i.e.

$$\left\lfloor \frac{(k-1)N}{2^K} \right\rfloor < i \leqslant \left\lfloor \frac{kN}{2^K} \right\rfloor.$$

Combining the two previous displays implies that $(k'-1)2^{K-K'} < k \leqslant k'2^{K-K'}$, meaning that $i \in B_k^{(K)}$. Moreover, applying (i) to k gives that $B_k^{(K)}$ is a subset of $B_{k'}^{(K')}$, hence the result.

B.4. Proof of Lemma 4.3

Using (i) from Lemma 4.2, we have:

$$\begin{aligned} B_{k_1}^{K_1} &\subset B_{k_1'}^{(3)}, \quad \text{with} \quad k_1' = \lfloor (k_1-1)2^{3-K_1} \rfloor + 1, \\ B_{k_2}^{K_2} &\subset B_{k_2'}^{(3)}, \quad \text{with} \quad k_2' = \lfloor (k_2-1)2^{3-K_2} \rfloor + 1. \end{aligned}$$

In addition, K_0 is by definition larger than 3 (see (4.4)) and (ii) from Lemma 4.3 then gives:

$$\begin{aligned} &(B_k^{(K_0)})_{k \in [\![(k_1'-1)2^{K_0-3},\ k_1'2^{K_0-3}]\!]} \quad \text{is a partition of} \quad B_{k_1'}^{(3)}, \\ &(B_k^{(K_0)})_{k \in [\![(k_2'-1)2^{K_0-3},\ k_2'2^{K_0-3}]\!]} \quad \text{is a partition of} \quad B_{k_2'}^{(3)}. \end{aligned}$$

Then, using the latter result and starting from the definition of $\mathcal{K}_0(K_1, k_1, K_2, k_2)$, we can write

$$\begin{aligned} &\mathcal{K}_0(K_1, k_1, K_2, k_2) \\ &= \left\{ k \in \left[2^{K_0}\right] \,\middle|\, B_k^{(K_0)} \cap (B_{k_1}^{(K_1)} \cup B_{k_2}^{(K_2)}) = \varnothing \right\} \\ &\supset \left\{ k \in \left[2^{K_0}\right] \,\middle|\, B_k^{(K_0)} \cap (B_{k_1'}^{(3)} \cup B_{k_2'}^{(3)}) = \varnothing \right\} \\ &= \left[2^{K_0}\right] \setminus \left([\![(k_1'-1)2^{K_0-3},\ k_1'2^{K_0-3}]\!] \cup [\![(k_2'-1)2^{K_0-3},\ k_2'2^{K_0-3}]\!] \right). \end{aligned}$$

The latter result together with the definition of K_0 yield the lower bound on the cardinality of $\mathcal{K}_0(k_1, K_1, K_2, k_2)$:

$$|\mathcal{K}_0(K_1, k_1, K_2, k_2)| \geqslant 2^{K_0} - 2^{K_0-3} - 2^{K_0-3} = \frac{3}{4}2^{K_0} \geqslant \frac{3}{4}2^{\log_2(V/3)+2} = V.$$

B.5. Proof of Lemma 4.4

Let $(m, m') \in \mathcal{M}^2$ and $v \in [V]$. By construction of $T_v^{(m,m')}$, there exists $k \in \left[2^{K_0}\right]$ such that $T_v^{(m,m')} = B_k^{(K_0)}$. Then, it follows from (4.4) that $2^{K_0} \leqslant 8V/3$, so that we can write:

$$\left| T_v^{(m,m')} \right| \geqslant \left\lfloor \frac{N}{2^{K_0}} \right\rfloor \geqslant \left\lfloor \frac{3N}{8V} \right\rfloor = \left\lfloor \frac{N}{4V} + \frac{N}{8V} \right\rfloor \geqslant \left\lfloor \frac{N}{4V} + 1 \right\rfloor \geqslant \frac{N}{4V},$$

where we used the fact that $V \leqslant N/8$ by assumption.

B.6. Proof of Corollary 5.2

Let us apply Corollary 4.1 with well-chosen functions $\nu\colon \lambda \to \mathbb{R}_+^*$ and $\rho\colon \Lambda \times \mathbb{N}^* \to \mathbb{R}_+ \cup \{+\infty\}$. For $s \in [s^*]$, denote,

$$\lambda_s = c_0 \|\zeta\|_{L_q} \sqrt{\log(ed/s)}$$

so that $\Lambda = \{\lambda_s\}_{s\in[s^*]}$. For $s \in [s^*]$, we define

$$\nu(\lambda_s) = s \log(ed/s).$$

Then, by definition of s^*, we get $\nu_{\max} = \lceil \max_{\lambda\in\Lambda} \nu(\lambda) \rceil = \lceil s^* \log(ed/s^*) \rceil$, and the assumption $N/(8V, 2^{K_{\min}+1}) \geqslant s_* \log(ed/s^*)$ implies $N \geqslant \nu_{\max} \max(8V, 2^{K_{\min}+1})$ which is required to apply Corollary 4.1. Besides, we define ρ as follows. For $s \in [s^*]$ and $c \in \mathbb{N}^*$,

$$\rho(\lambda_s, c) = \begin{cases} s\lambda_s^2 c_1 c_0^{-2} c^{-1} & \text{if } s \geqslant \|\beta^*\|_0 \\ +\infty & \text{otherwise.} \end{cases}$$

For all $s \geqslant \|\beta^*\|_0$, β^* is of course s-sparse, and then according to Proposition 5.1, for $B \subset \mathcal{I}$ such that $|B| \geqslant \nu(\lambda_s)$, it holds with probability higher than $1 - e^{-1/48}$ that

$$\ell(\hat{f}_{\lambda_s,B}) = \ell(\hat{\beta}_{\lambda_s |B|^{-1/2},B}) \leqslant \rho(\lambda_s, |B|).$$

The above inequality is also true for $s < \|\beta^*\|_0$ since the right-hand side is then infinite. We can now apply Corollary 4.1 which gives that with probability higher than $1 - ((s^*)^2 N^2 + 1)e^{-V/48}$, it holds that:

$$(1 - a_{\varepsilon,V})\ell(\hat{f}_{\hat{m}}) \leqslant (1 + 3a_{\varepsilon,V}) \min_{\lambda\in\Lambda} \rho\left(\lambda, \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor\right) + 2b_{\varepsilon,V},$$

and the result follows by noting that:

$$\begin{aligned} \min_{\lambda\in\Lambda} \rho\left(\lambda, \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor\right) &= \min_{s\in[s^*]} \rho\left(\lambda_s, \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor\right) \\ &= \min_{s\geqslant\|\beta^*\|_0} s\lambda_s^2 c_1 c_0^{-2} \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor^{-1} \\ &= \min_{s\geqslant\|\beta^*\|_0} \|\zeta\|_{L_q}^2 \, s \log(ed/s) c_1 \left\lfloor \frac{N}{\max(4V, 2^{K_{\min}})} \right\rfloor^{-1} \\ &= c_1 \|\zeta\|_{L_q}^2 \frac{\|\beta^*\|_0 \log(ed \|\beta^*\|_0^{-1})}{\lfloor N/\max(4V, 2^{K_{\min}}) \rfloor}. \end{aligned}$$