Median of means principle as a divide-and-conquer procedure for robustness, sub-sampling and hyper-parameters tuning

Joon Kwon, Guillaume Lecué and Matthieu Lerasle

December 6, 2018

Abstract

Many learning methods have poor risk estimates with large probability under moment assumptions on data, are sensitive to outliers and require hyper-parameters tuning. The purpose here is to introduce an algorithm whose task is, when fed with such learning methods and possibly corrupted data satisfying at best moment assumptions to: return a robust estimator with good excess risk bounds holding with exponentially large probability estimate, identify large non-corrupted subsamples and automatically tune hyper-parameters. The procedure is tested on the LASSO [48] which is known to be highly sensitive to outliers. The basic tool is the median-of-means principle [1, 24, 40] which can be recast as a divide-and-conquer methodology [25], making this procedure easily scalable.

1 Introduction

Robustness has become an important subject of interest in the machine learning community over the last few years because large datasets are highly sensitive to corruption. These may appear due to hardware, storage or transmission issues when datasets are distributed on several servers. As can be seen, for instance, in Figure 1 in [29] and Figure 1 and 5 in [30], many learning algorithms based on emprical risk minimization including the LASSO, may be completely mislead by a single corrupted data.

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, 47, 51]. Various robust descent algorithms have also been considered recently [43, 41, 42, 23, 22]. Despite these important progresses, the ultimate step of a data-scientist routine, which is the estimator selection, is yet to receive a proper treatment. Actually, practitioners usually have at disposal several procedures, each of these requiring one or several tuning parameters to be properly calibrated. The goal of estimator selection [8, 11, 7] is to select among these candidates a final estimator. An alternative to estimator selection is aggregation [16, 50, 39] where the goal is to select a linear or convex combination of the candidates. Aggregation is also referred to as Ensemble methods in the Machine Learning community. Classical examples of ensemble methods include binning, boosting, bagging or stacking.

The most classical way to select/aggregate candidates is the validation principle. The dataset is partitioned into a "train set" used to build candidates and a "test set" used to measure their ability to generalize by estimating their risks. The final estimator used in practice 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 base estimators are robust, outliers in the test set may break performance of the selection/aggregation step, resulting in a poor final estimator. This raises the question of a robust selection/aggregation step addressed in this work.

There exist many data-driven methods to tune hyperparameters or to select an estimator in a collection of candidates. Among these, one can mention SURE method [46], model selection [8, 11, 12, 35, 9, 38] 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 [50, 44, 52], cross-validation [2, 3] or Lepski [33] and Goldenschluger-Lepski [21] methods to name a few. To the best of our knowledge, all these techniques either use the 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 motivation of this paper is to propose a general method of robust estimator selection, with provable theoretical guarantees. Roughly stated, the method uses median-of-means [1, 24, 40] to build robust pairwise comparisons between candidates, and the final estimator is selected by a minmax procedure in the spirit of [4, 27] or the Goldenshluger-Lepski method [21], see Section 2.4 for details. The method is easy to implement in prediction problems. We focus here on least-squares regression and refer to [34] for other examples including density estimation and classification.

The second motivation is to use estimator selection procedures to build subsampling strategies. Subsampling is used in machine learning for computational reasons: some algorithms require to break large datasets in smaller pieces to run on smaller problems [25]. This has been used 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 chooses randomly 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]. Subsampling is also relevant to work with corrupted datasets since base estimators should be trained on clean parts of the dataset. Searching for large non-corrupted subsamples is another aim of this work.

The idea underlying our approach is that the choice of a subsample is as an hyper-parameter selection problem. Therefore, estimator selection procedures can be used to select a subsample. Overall, the procedure is computationally attractive on huge datasets, robust to the presence of a few corrupted data and selects an efficient subsample for the learning task.

The paper is divided as follows. Section 2 presents the general setting of prediction where our main estimator selection procedure is defined. Theoretical guarantees of the selection procedure are provided in Section 3. Section 4 applies the general strategy more specifically to the problem of subsample selection, and numerical experiments are presented in Section 6. An application to the problem of robustness of linear aggregators is provided in Section 5. The proofs of technical lemmas are outsourced in the appendix in Sections A and B.

2 Setting

2.1 Notation

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. $]\!]k, l]\!] = \{k+1, k+2, \dots, l\}$. Call partition of a set E any family of disjoint subsets of E with union equal to E.

2.2 Preliminaries

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 $Q\left[g\right]:=\mathbb{E}_{Z\sim Q}\left[g(Z)\right]$.

Let F be a linear subspace of $L^2(P_X)$. For any $f\in F$, 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* $R(f):=P\left[\gamma(f)\right]$ and let f^* denote the *oracle*

$$f^*:=\operatorname*{arg\,min}_{f\in F}R(f).$$

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. Consider the following assumption.

Assumption 1. *For every $f\in F$,*

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

2.3 Data

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 $(X_i,Y_i)_{i\in\mathcal{I}}$ are 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 not known by the statistician beforehand.

For any nonempty subset B of $[N]$, and any measurable function $g:\mathbb{X}\times\mathbb{R}\to\mathbb{R}$, denote

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

2.4 An estimator selection procedure

Let $\mathcal{M}$ be a finite set and $(\hat{f}_m)_{m\in\mathcal{M}}$ be a collection of *estimators*. All along the paper, each index m is a couple $m=(\lambda_m,B_m)$, where λ_m describes the method and the set of parameters and B_m is a nonempty subset of $[N]$ of cardinality less than $N/4$. For each $m\in\mathcal{M}$, estimator $\hat{f}_m$ is built with data $((X_i,Y_i))_{i\in B_m}$ using the algorithm and parameters λ_m, that is there exists a measurable application

$$G_{\lambda_m}:\bigcup_{n\geqslant 1}(\mathbb{X}\times\mathbb{R})^n\longrightarrow F\quad\text{such that}\qquad\hat{f}_m=G_{\lambda_m}((X_i,Y_i)_{i\in B_m}).$$

For computational issues, $(B_m)_{m\in\mathcal{M}}$ is a strict sub-collection of $\mathcal{P}([N])$. Section 4 gives some specific structured sub-samples. The best choice of $m\in\mathcal{M}$ regarding our final objective satisfies

$$m_o:=\operatorname*{arg\,min}_{m\in\mathcal{M}}P\left[\gamma(\hat{f}_m)\right]=\operatorname*{arg\,min}_{m\in\mathcal{M}}\max_{m'\in\mathcal{M}}P\left[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\right]\quad.\tag{1}$$

The real-valued expectations $P\big[\gamma(\hat{f}_m)-\gamma(\hat{f}_{m'})\big],m,m'\in\mathcal{M}$ are unknown and should be estimated.

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 estimators $\mathcal{T}(m, m')$ of $P\big[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\big]$ are defined by:

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

where $\mathrm{med}_{v\in[V]}$ denotes a median of the V empirical means $P_{T_v^{(m,m')}}\big[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\big], v \in [V]$. The selection of the final estimator is obtained by plugging these median-of-means estimators into equation (1): we select $\hat{f}_{\widehat{m}}$, where

$$\hat{m} := \underset{m\in\mathcal{M}}{\arg\min} \max_{m'\in\mathcal{M}} \mathcal{T}(m, m'). \tag{2}$$

Median-of-means have been introduced in [1, 24, 40]. Median-of-means pairwise comparisons have been used to build robust estimators in [36, 28]. Minmax strategies have been used in [4, 27] for least-squares regression and in [5] for density estimation. Finally, the minmax principle has been used for (non-robust) selection of estimators in [21].

2.5 Minmax MOM selection to divide-and-conquer

It is classical to use divide-and-conquer approaches [25] to deal with large databases: the database is divided in small batches, algorithms are run on each batch and the results are "aggregated". Minmax MOM selection (2) can perform this aggregation. Denote by B_m the block of data hosted on server m for $m = 1, \dots, V$. Train estimators $\hat{f}_m$ for all $m \in \mathcal{M} = [V]$ (leaving apart the choice of hyper-parameters for simplicity). For all $m, m' \in \mathcal{M}$ and $v \in [V] \setminus \{m, m'\}$, let $T_v^{(m,m')} = B_v$, compute the V real numbers $P_{B_v}\big[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'})\big]$ and take their median. Then compute the minmax MOM estimator (if there are too many medians, choose at random m and m' in $[V]$ a smaller number of times). Following the map-reduce terminology [18], the mapper is the training of the procedure itself and the V evaluations over the remaining V datasets. The reducer is the computation of the $\binom{V}{2}$ medians of differences of empirical risks and the minmax MOM selection (2).

3 Robust oracle inequality

Theorem 3.1. *Grant Assumption 1 and let $(\hat{f}_m)_{m\in\mathcal{M}}$ be a family of estimators in F. If $V \geqslant 3\,|\mathcal{O}|$, then with probability larger than $1 - |\mathcal{M}|^2\, e^{-V/48}$, the estimator $\hat{f}_{\widehat{m}}$, where $\hat{m}$ is selected by minmax-MOM criterion (2) satisfies, for all $\varepsilon > 0$,*

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

where $f \to \ell(f) = R(f) - R(f^*)$ is the excess loss function,

$$a_{\varepsilon,V} := 8\chi^2 \sqrt{\frac{2V}{N}} + 2\sqrt{2}\varepsilon \quad \text{and} \quad b_{\varepsilon,V} := \frac{64V\sigma^2}{N\varepsilon}.$$

The proof of Theorem 3.1 is postponed to Section B.1. Roughly speaking, Theorem 3.1 states that, with exponentially large probability, the selected estimator (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* oracle inequality 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 $\log|\mathcal{M}|/N$, which is minimax optimal [49]. The oracle inequality 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 [45] without breaking the results.

4 Robustness, sub-sampling and hyper-parameters tuning

4.1 Theoretical guarantees

Assume $N \geqslant \max(8,\ 6|\mathcal{O}|)$. Let $K_{\max} \in [\![3,\ \lfloor \log_2 N \rfloor]\!]$. Then for each $K \in [\![3,\ K_{\max}]\!]$, consider a partition $(B_k^{(K)})_{k\in[2^K]}$ of $[N]$ such that for all $k \in [2^K]$, $\lfloor N/2^K \rfloor \leqslant \left|B_k^{(K)}\right|$. $(B_k^{(K)})_{k\in[2^K]}$ is called 2^K-partition. Let Λ be a set of hyper-parameters and let

$$\mathcal{B} = \bigcup_{K=3}^{K_{\max}} \bigcup_{k\in[2^K]} \left\{B_k^{(K)}\right\}, \qquad \mathcal{M} = \Lambda \times \mathcal{B}. \tag{3}$$

Note that each subsample in $\mathcal{B}$ has cardinality less than $N/4$.

Remark 4.1. If computational power is an issue, one can restrict $\mathcal{B}$ by bounding K from below by some integer $K_{\text{comp}} > 3$. One can also consider subpartitions $(B_k^{(K)})_{k\in\mathcal{I}_K}$ for $\mathcal{I}_K \subset [2^K]$. For the sake of conciseness, theoretical results are provided only when $K_{\text{comp}} = 3$ and $\mathcal{I}_K = [2^K]$. The extension to more general situations does not involve new ideas.

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 $v \in [V]$.

Corollary 4.2. *Let $\mathcal{M}$ be defined by (3). Grant Assumption 1 and assume that the family $(\hat{f}_{\lambda,B})_{(\lambda,B)\in\mathcal{M}}$ takes values in F. Let $\rho : \Lambda \times \mathbb{N}^* \to \mathbb{R}_+$ 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 $N \geqslant 8\nu_{max}V$ and $V \geqslant 3|\mathcal{O}|$. Assume that, for all $\lambda \in \Lambda$ and all $B \subset \mathcal{I}$ such that $|B| \geqslant \nu(\lambda)$, with probability larger than $1 - \exp(-1/48)$, $\ell(\hat{f}_{\lambda,B}) \leqslant \rho(\lambda, |B|)$. Then, the estimator $\hat{f}_{\widehat{m}}$ defined in (2) satisfies*

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

with probability larger than

$$1 - (|\Lambda|^2 N^2 + 1)e^{-V/48}. \tag{5}$$

Corollary 4.2 is proved in Section B.2. Let us stress some relevant points.

- 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, and when B is large enough and contains only informative data. For example, this condition is met by ERM when informative data satisfy moment assumptions, see Proposition 5.1 below. With these arguably weak requirement, the minmax MOM procedure (2) achieve the best bound $\rho(\lambda, |B|)$ with exponentially large probability (5).
- 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 $\lambda \to \nu(\lambda)$ is introduced to handle situations where the risk bound holds only when the sample size is larger than $\nu(\lambda)$.

4.2 An efficient partition scheme of the dataset

This section presents partitions $(B_k^{(K)})_{k\in[2^K]}$ ($K \in [\![3,\ K_{\max}]\!]$) and $(T_v^{(m,m')})_{v\in[V]}$ (for $(m,m') \in \mathcal{M}^2$) with computational advantage, used in the numerical experiments of Section 6.

The selection procedure presented in Section 2.4 requires, for all $(m,m') \in \mathcal{M}^2$ and $v \in [V]$, the computation of $P_{T_v^{(m,m')}}\big[\gamma(\hat{f}_m)\big]$ and $P_{T_v^{(m,m')}}\big[\gamma(\hat{f}_{m'})\big]$. Since the partition $(T_v^{(m,m')})_{v\in[V]}$ of $[N] \setminus (B_m \cup B_{m'})$ 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 requires the computation of only $8V|\mathcal{M}|/3$ empirical risks.

For each $K \in [\![3,\ \lfloor \log_2 N \rfloor]\!]$ and each $k \in [2^K]$, let

$$B_k^{(K)} := \left[\!\!\left[\left\lfloor \frac{(k-1)N}{2^K} \right\rfloor,\ \left\lfloor \frac{kN}{2^K} \right\rfloor \right]\!\!\right].$$

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 $k \in [2^K]$, $\lfloor N/2^K \rfloor \leqslant \big|B_k^{(K)}\big| \leqslant N/4$, as required. Moreover, the following key property holds.

Lemma 4.3. *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

$$K_0 := \lceil \log_2(V/3) \rceil + 2. \tag{6}$$

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

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

Lemma 4.4. *For all $3 \leqslant K_1, K_2 \leqslant \lfloor \log_2 N \rfloor$ and $k_1 \in \big[2^{K_1}\big]$, $k_2 \in \big[2^{K_2}\big]$, 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 4.4, 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 2.4).

Lemma 4.5. *For all $(m,m') \in \mathcal{M}^2$ and $v \in [V]$, $\big|T_v^{(m,m')}\big| \geqslant N/(4V)$.*

The empirical risk of each estimator $\hat{f}_m$ has to be computed on the 2^{K_0}-partition only, which thanks to (6) means the computation of at most $8V|\mathcal{M}|/3$ empirical risk as advertised.

5 Applications

5.1 ERM and linear aggregation

This section illustrates Corollary 4.2 considering non-robust linear aggregates of preliminary estimators as baseline estimators. 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:

$$f_\lambda^* := \operatorname*{arg\,min}_{f\in F_\lambda} R(f).$$

For any nonempty set $B \subset \mathcal{I}$, denote $\hat{f}_{\lambda,B}$ the empirical risk minimizer (ERM) on F_λ associated with data $(X_i, Y_i)_{i\in B}$:

$$\hat{f}_{\lambda,B} := \operatorname*{arg\,min}_{f\in F_\lambda} \frac{1}{|B|}\sum_{i\in B}(Y_i - f(X_i))^2. \tag{7}$$

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

Proposition 5.1. *[31, Theorem 1.3] Let $\lambda \in \Lambda$. Assume that there exists $\chi_\lambda > 0$ such that for all $f \in F_\lambda$, $(Pf^4)^{1/4} \le \chi_\lambda (Pf^2)^{1/2}$. Denote $\zeta_\lambda := Y - f_\lambda^*(X)$ and assume that $(P\zeta_\lambda^4)^{1/4} \le \sigma_\lambda$. 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 (7) 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 5.1, 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 [31, Proposition 1.5]: ERM are not robust to "stochastic outliers" in general.

Let $\mathcal{M} = \Lambda \times \mathcal{B}$ be as in (3) and N, V and $K_{\max}$ be such that $N \ge \max(8, 6|\mathcal{O}|)$, $3 \le V \le N/8$ and $K_{\max} \ge \log_2 V$. Consider the minmax selection procedure $\hat{f}_{\widehat{m}}$ from (2) on the family $(f_{\lambda,B})_{(\lambda,B)\in\mathcal{M}}$ of linear aggregate obtained by ERM in (7). The following result combines Corollary 4.2 and Proposition 5.1 (see the proof in the Appendix).

Corollary 5.2. *Grant Assumption 1 on F and assume that for all $\lambda \in \Lambda$ and all $f \in F_\lambda$, $(Pf^4)^{1/4} \le \chi_\lambda (Pf^2)^{1/2}$ and $(P\zeta_\lambda^4)^{1/4} \le \sigma_\lambda$ for $\zeta_\lambda := Y - f_\lambda^*(X)$. Assume also that $N \ge \max_{\lambda\in\Lambda}(1600\chi_\lambda^4)^2 d_\lambda$. Then, with probability at least $1 - (|\Lambda|^2 N^2 + 1)\exp(-V/48)$, for all $\varepsilon > 0$,*

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

While Proposition 5.1 shows statistical guarantee with constant probability for the estimators $\hat{f}_{\lambda,B}$ trained on clean data, Corollary 5.2 shows that minmax MOM selection improves the constant probability into an exponential probability, allows $|\mathcal{O}|$ outliers as long as $V \ge 3|\mathcal{O}|$ and selects simultaneously the best method and set of hyperparameters $\lambda \in \Lambda$.

5.2 Application to LASSO

Consider here as baseline estimators the LASSO $\hat{\beta}_{\lambda,B}$ with regularization parameter λ trained on $B \subset [N]$. Assume that $(\lambda, B) \in \Lambda \times \mathcal{B}$ where $\Lambda \subset \mathbb{R}^+$ is any finite grid and $\mathcal{B}$ is defined in (3). Statistical guarantees of these baseline estimators have been obtained in [32, Theorem 1.4] with a regularization parameter $\lambda \asymp \|\zeta\|_{L_q} \sqrt{s \log(ed/s)/N}$ instead of $\|\zeta\|_{L_q} \sqrt{s \log(ed)/N}$. This choice is 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 $p \geq 1$ and $t \in \mathbb{R}^d, (\mathbb{E}|\langle X, t\rangle|^p)^{1/p} \leq L \|t\|_2$. Moreover, there exists $q > 2$ such that, for any $\beta^* \in \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, \beta\rangle)^2$, $\zeta := Y - \langle X, \beta^*\rangle \in L_q$.*

Proposition 5.3. *Grant Assumption 2. Assume that β^* is s-sparse for some $1 \leq s \leq d$. Let $B \subset \mathcal{I}$ be such that $|B| \geq s \log(ed/s)$. Then, there exist absolute constants c_0 and c_1 such that the LASSO with regularization parameter $\lambda = c_0 \|\zeta\|_{L_q} \sqrt{s \log(ed/s)/N}$ satisfies, with probability at least $1 - \exp(-1/48)$,*

$$\left\|\hat{\beta}_{\lambda,B} - \beta^*\right\|_2^2 = R(\hat{\beta}_{\lambda,B}) - R(\beta^*) = \ell(\hat{\beta}_{\lambda,B}) \leq c_1 \|\zeta\|_{L_q}^2 \frac{s \log(ed/s)}{|B|} \quad . \tag{8}$$

Proposition 5.3 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 $\lambda \sim \|\zeta\|_{L_q} \sqrt{s \log(ed/s)/N}$. This regularization parameter requires the knowledge of the sparsity s. Proposition 5.3 shows that the risk bound only holds with constant probability because the noise is only assumed to have finite L_q-moment. Finally, LASSO has to be trained with good data; a single outlier completely breaks down its statistical properties (see Figure 1 in [27]). Let us now apply Corollary 4.2 and Proposition 5.3 to this example.

Corollary 5.4. *Grant Assumption 2. Denote by s^* the largest integer such that $\lfloor N/(4V) \rfloor \geq s^* \log(ed/s^*)$. Assume that β^* is s-sparse for some $s \leq s^*$ and that there are constants $c < C$ such that $c \leq \|\zeta\|_{L_q} \leq C$. Let $(\hat{\beta}_{\lambda,B})_{(\lambda,B)\in\Lambda\times\mathcal{B}}$ be a family of Lasso estimators where $\mathcal{B}$ is defined in (3) and Λ is a regular $1/N$-grid of $[c\sqrt{\log(ed)/N}, C\sqrt{s^* \log(ed/s^*)/N}]$. Let $(\hat{\lambda}, \hat{B})$ be selected according to the minmax MOM selection method (2). Then, with probability at least $1 - (CN^{5/2}\sqrt{s^* \log(ed/s^*)} + 1)\exp(-V/48)$, for all $\varepsilon > 0$,*

$$(1 - a_{\varepsilon,V}) \left\|\hat{\beta}_{\hat{\lambda},\hat{m}} - \beta^*\right\|_2^2 \leq (1 + 3a_{\varepsilon,V}) c_1 \|\zeta\|_{L_q}^2 \frac{s \log(ed/s)}{N/(4V)} + 2b_{\varepsilon,V}.$$

The improvements provided by minmax MOM selection procedure are similar here as those in the previous section. The selected estimator achieves optimal risk bounds, with exponential probability, even if the dataset has been corrupted. The proof of Corollary 5.4 is identical to the one of Corollary 5.2 and therefore not repeated.

6 Application to the LASSO with numerical experiments

In this section, the minmax MOM selection procedure from Section 4 is implemented for LASSO baseline estimators such as in Section 5.2. Numerical experiments[1] are performed with various numbers and "types" of outliers in the dataset in order to investigate their effects on the selected estimator $\hat{f}_{\hat{m}}$ and the selected parameter $(\lambda_{\hat{m}}, B_{\hat{m}})$.

[1] All codes are available at https://github.com/lecueguillaume/MOMpower/tree/master/MOM_selection

6.1 Dataset

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 $|\mathcal{O}| = 0, 4, 8 \ldots, 151$. 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:

$$X_i = (1, \ldots, 1) \in \mathbb{R}^{2000}, \quad Y_i = 10000, \quad i \in \mathcal{O}_1,$$

and second type, which we call *heavy-tail outliers* are constructed as

$$X_i \sim \mathcal{N}(0, I_{2000}), \quad Y_i = \langle X_i, \beta_0 \rangle + \zeta_i, \quad i \in \mathcal{O}_2,$$

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 are drawn according to

$$X_i \sim \mathcal{N}(0, I_{2000}), \quad Y_i = \langle X_i, \beta_0 \rangle + \zeta_i, \quad i \in \mathcal{I},$$

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 by the distribution of the noise, and should not affect too much the performance of 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 Figures 1c and 1d).

6.2 Median-of-mean subsampling procedure with LASSO

We consider the following grid of values for the regularization parameter of the LASSO:

$$\Lambda = \left\{ e^k \,\middle|\, k \in \frac{1}{2} [\![-2, 4]\!] \right\} \tag{9}$$

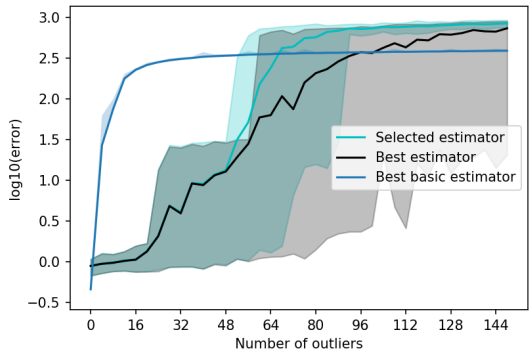
We consider the MOM-based subsampling procedure presented in Section 4 with parameters $V = 40$ and $K_{max} = 4$. The set $\mathcal{B}$ of subsamples is constructed as in Section 4.2 and we set $\mathcal{M} = \Lambda \times \mathcal{B}$. For each $m = (\lambda, B) \in \mathcal{M}$, we compute the LASSO estimator $\hat{\beta}_m$ with hyper-parameter λ based on subsample B:

$$\hat{\beta}_m = \hat{\beta}_{\lambda, B} = \underset{\beta \in \mathbb{R}^{2000}}{\arg\min} \left\{ \frac{1}{2\,|B|} \sum_{i \in B} (Y_i - \langle \beta, X_i \rangle)^2 + \lambda \, \|\beta\|_1 \right\}.$$

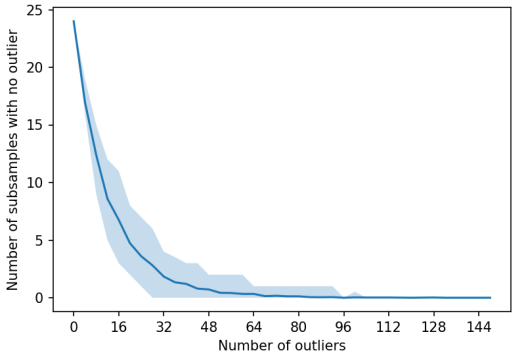
We compute $\hat{\beta}_{\widehat{m}}$ the estimator selected by the MOM-based sub-sampling procedure, which uses partitions $(T_v^{(m,m')})_{v \in [V]}$ (for $m, m' \in \mathcal{M}$) constructed as in Section 4.2 Let us denote $\hat{\beta}_{\widetilde{m}}$ the best oracle estimator among $(\hat{\beta}_m)_{m \in \mathcal{M}}$, 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 based on the whole dataset, which we will call *basic estimators*:

$$\hat{\beta}_{\lambda,[N]} = \underset{\beta \in \mathbb{R}^{2000}}{\arg\min} \left\{ \frac{1}{2N} \sum_{i \in [N]} (Y_i - \langle \beta, X_i \rangle)^2 + \lambda \, \|\beta\|_1 \right\}, \quad \lambda \in \Lambda,$$

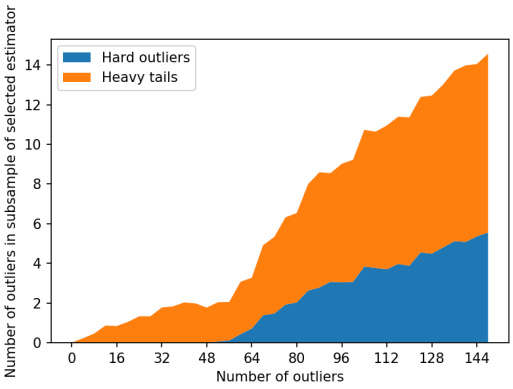
and let $\hat{\beta}_{\widetilde{\lambda},[N]}$ be the best among the latter, so that $\tilde{\lambda} := \arg\min_{\lambda \in \Lambda} R(\hat{\beta}_{\lambda,[N]})$.



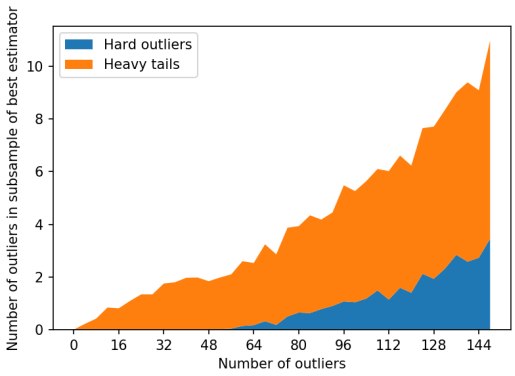
(a) Error of estimators $\hat{\beta}_{\widehat{m}}$, $\hat{\beta}_{\widetilde{m}}$ and $\hat{\beta}_{\widetilde{\lambda},[N]}$



(b) Number of subsamples with no outlier



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



(d) Number of outliers in $B_{\widetilde{m}}$

Figure 1: MOM-subsampling procedure run with $N = 1000$, $V = 40$, $K_{\max} = 4$, and averaged over 200 experiments.

6.3 On the choices of V and $K_{\max}$

The choices of V and $K_{\max}$ have an impact on both the performance the selected estimator and the computation time.

The higher is V, the higher is the number of outliers that the MOM-selection can handle, and as a matter of fact, Theorem 3.1 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 3.1). Here we choose $V = 40$, so we can expect the 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 computation of estimators on 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 us to hope for good selection performance up to $|\mathcal{O}| = 13$ outliers, we choose $K_{\max} = 4$ which indeed satisfies $2^{K_{\max}} > |\mathcal{O}|$.

6.4 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 estimator $\hat{\beta}_{\widehat{m}}$ selected by the MOM-subsampling procedure, 2) of $\hat{\beta}_{\widetilde{m}}$, the best estimator among $(\hat{\beta}_m)_{m\in\mathcal{M}}$, and 3) of the best *basic* estimator $\hat{\beta}_{\widetilde{\lambda},[N]}$. As soon as the dataset contains outliers, basic estimators $(\hat{\beta}_{\lambda,[N]})_{\lambda\in\Lambda}$ have larger errors than $\hat{\beta}_{\widetilde{m}}$ (the best estimator computed on a subsample). For $|\mathcal{O}| \leqslant 48$ the MOM-based subsampling procedure has the same error as the best estimator among $(\hat{\beta}_m)_{m\in\mathcal{M}}$.

For a given value of parameter V, the MOM-based selection procedure 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 3.1 holds for $|\mathcal{O}| \leqslant V/3$, that is $|\mathcal{O}| \leqslant 13$ here. It seems here that the MOM-based selection procedure performs satisfactorily for $|\mathcal{O}| \leqslant 48$ and even selects a reasonably good estimator when $|\mathcal{O}| \leqslant 56$.

Figures 1c and 1d show the number of each type of outliers in the selected subsample $B_{\widehat{m}}$ and in the best subsample $B_{\widetilde{m}}$. The selection procedure manages to rule out hard outliers when $|\mathcal{O}| \leqslant 48$, and the selected 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 $|\mathcal{O}| \geq 48$. Besides, both selected $B_{\widehat{m}}$ and oracle $B_{\widetilde{m}}$ subsamples contain 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 selection procedure use this extra information automatically in an optimal way. In particular, the minmax MOM selection rule distinguishes between non informative hard outliers and possibly informative heavy-tailed outliers.

Overall, the MOM-based selection procedure shows very strong robustness to the presence of outliers and outputs an estimator with the best possible performance among the given class of estimators.

7 Conclusion

We propose a systematic selection procedure based on the median-of-means principle that automatically improves robustness of any estimator with respect to adversarial and stochastic outliers, selects a large uncorrupted sub-sample and automatically tune hyper-parameters. This selection procedure is natively a divide-and-conquer method and is therefore highly scalable.

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).

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

[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.

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

[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.

[6] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011.

[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.

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

[9] Pierre C. Bellec. Optimal bounds for aggregation of affine estimators. *Ann. Statist.*, 46(1):30–59, 2018.

[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.

[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.

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

[13] Jelena Bradic. Randomized maximum-contrast selection: subagging for large-scale regression. *Electronic Journal of Statistics*, 10(1):121–170, 2016.

[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.

[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.

[17] Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.

[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.

[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.

[21] Goldenshluger, A. and Lepski, O. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 11 2008.

[22] Matthew J Holland. Classification using margin pursuit. *arXiv preprint arXiv:1810.04863*, 2018.

[23] Matthew J Holland. Robust descent using smoothed multiplicative noise. *arXiv preprint arXiv: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.

[25] Michael I. Jordan. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013.

[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. arXiv:1112.5016.

[27] G. Lecué and M. Lerasle. Robust machine learning by median-of-means : theory and practice. Technical report, CNRS, ENSAE, Paris-sud, 2017.

[28] G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud, 2017. arXiv:1701.01961. To appear in Stochastic Processes and their applications.

[29] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

[30] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

[31] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[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.

[33] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.

[34] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv: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.

[36] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS*.

[37] Lester Mackey, Ameet Talwalkar, and Michael I. Jordan. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 16:913–960, 2015.

[38] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[39] 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.

[40] 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.

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

[42] 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 arXiv:1711.04734*, 2017.

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

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

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

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

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

[48] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[49] 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.

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

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

[52] 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.

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

$$\left| \mathcal{V}^{(m,m')} \right| \geqslant V - |\mathcal{O}| \,. \tag{10}$$

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 $\mathrm{Var}_{m,m'}(\,\cdot\,) := \mathrm{Var}\left(\,\cdot\,\middle|\,(X_i,Y_i)_{i\in B_m\cup B_{m'}}\right)$, we have

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

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

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

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:

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

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 $1 - 1/8$,

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

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

$$\ell(\hat{f}_m) - \ell(\hat{f}_{m'}) - \sqrt{8\, C_{m,m'}} \leqslant P_{T_v^{(m,m')}} \left[\gamma(\hat{f}_m) - \gamma(\hat{f}_{m'}) \right] \leqslant \ell(\hat{f}_m) - \ell(\hat{f}_{m'}) + \sqrt{8\, C_{m,m'}}. \quad (11)$$

Denote by $\Omega_v^{(m,m')}$ the event defined by (11) 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 conditionnally 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 (10),

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

Then, on $\Omega^{(m,m')}$, using (10) and the assumption $V \geq 3|\mathcal{O}|$,

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

In other words, inequalities (11) 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 $v \in [V]$:

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

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 $2xy \leqslant x^2/\varepsilon' + \varepsilon' y^2$,

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

■

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

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

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 $b_{\varepsilon,V}$. ■

B Proofs of the main results

B.1 Proof of Theorem 3.1

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

$$(1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) = (1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) - (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}) + (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}). \tag{12}$$

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

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

It follows from Lemma A.4 and (12) that, on Ω,

$$(1-a_{\varepsilon,V})\ell(\hat{f}_{\widehat{m}}) \leqslant \max_{m\in\mathcal{M}} \mathcal{T}(\hat{m},m) + b_{\varepsilon,V} + (1+a_{\varepsilon,V})\ell(\hat{f}_{m_o}). \tag{13}$$

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 (13) yields the result.

B.2 Proof of Corollary 4.2

First, it follows from $N \geqslant 8\nu_{\max}V$ that there exists $K \geqslant 3$ such that

$$2V < 2^K \leqslant \frac{N}{\nu_{\max}}. \tag{14}$$

Let $K_0 \geqslant 3$ be the smallest integer satisfying (14). In particular, $2^{K_0} \leqslant (4V)$. Let

$$\lambda_0 := \arg\min_{\lambda\in\Lambda} \rho(\lambda, \lfloor N/2^{K_0} \rfloor) \qquad \text{and} \qquad \rho_0 := \rho(\lambda_0, \lfloor N/2^{K_0} \rfloor).$$

$$\mathcal{K}_0 := \left\{ k \in [2^{K_0}] \,\middle|\, B_k^{(K_0)} \subset \mathcal{I} \right\},$$

which is nonempty because $|\mathcal{K}_0| \geqslant 2^{K_0} - V \geqslant V$ thanks to the first inequality from (14). Let

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

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

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

Here the second inequality comes from the definition of Ω_2 and the last inequality from $2^{K_0} \leq (4V)$ combined with ρ nonincreasing in its second variable. Combining the above with the definition of Ω_1 yields the desired inequality:

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

To conclude the proof, let us bound from below the probability of $\Omega_1 \cap \Omega_2$. By Theorem 3.1,

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

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

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

The third inequality follows from the excess risk bound on $\hat{f}_{\lambda_0,B_k^{(K_0)}}$, which holds as $\left|B_k^{(K_0)}\right| \geqslant \lfloor N/2^{K_0}\rfloor \geqslant \nu_{\max} \geqslant \nu(\lambda_0)$ thanks to (14). Therefore, $\mathbb{P}\left[\Omega_1 \cap \Omega_2\right] \geqslant 1 - (|\Lambda|^2 N^2 + 1)\exp(-V/48)$.

B.3 Proof of Lemma 4.3

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 4.4

Using (i) from Lemma 4.3, we have:

$$\begin{aligned} B^{K_1}_{k_1} &\subset B^{(3)}_{k'_1}, \quad \text{with} \quad k'_1 = \lfloor (k_1-1)2^{3-K_1} \rfloor + 1, \\ B^{K_2}_{k_2} &\subset B^{(3)}_{k'_2}, \quad \text{with} \quad k'_2 = \lfloor (k_2-1)2^{3-K_2} \rfloor + 1. \end{aligned}$$

In addition, K_0 is by definition larger than 3 (see (6)) and (ii) from Lemma 4.4 then gives:

$$\begin{aligned} (B^{(K_0)}_k)_{k\in\rrbracket (k'_1-1)2^{K_0-3},\, k'_1 2^{K_0-3}\rrbracket} &\quad \text{is a partition of} \quad B^{(3)}_{k'_1}, \\ (B^{(K_0)}_k)_{k\in\rrbracket (k'_2-1)2^{K_0-3},\, k'_2 2^{K_0-3}\rrbracket} &\quad \text{is a partition of} \quad B^{(3)}_{k'_2}. \end{aligned}$$

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

$$\begin{aligned} \mathcal{K}_0(K_1,k_1,K_2,k_2) &= \left\{ k \in \left[2^{K_0}\right] \,\middle|\, B^{(K_0)}_k \cap (B^{(K_1)}_{k_1} \cup B^{(K_2)}_{k_2}) = \varnothing \right\} \\ &\supset \left\{ k \in \left[2^{K_0}\right] \,\middle|\, B^{(K_0)}_k \cap (B^{(3)}_{k'_1} \cup B^{(3)}_{k'_2}) = \varnothing \right\} \\ &= \left[2^{K_0}\right] \setminus \left(\rrbracket (k'_1-1)2^{K_0-3},\, k'_1 2^{K_0-3}\rrbracket \cup \rrbracket (k'_2-1)2^{K_0-3},\, k'_2 2^{K_0-3}\rrbracket \right). \end{aligned}$$

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 4.5

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 (6) 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

The proof follows from Proposition 5.1 and Corollary 4.2. Let us check the assumption and the features of both results. For $x = 2\exp(1/48)$ and when $|B| \geq (1600\chi_\lambda^4)^2 d_\lambda$ we have $1 - \exp(-|B|/(64\chi_\lambda^8)) - 1/x \geq 1 - \exp(-1/48)$ therefore, $\hat{f}_{\lambda,B}$ satisfies an (exact) oracle inequality with probability larger than $1 - \exp(-1/48)$ when $|B| \geq \nu(\lambda) := (1600\chi_\lambda^4)^2 d_\lambda$ 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 4.2 are satisfied and the result follows from a direct application of the latter result.