*Submitted to the Annals of Statistics*
arXiv: arXiv:1906.03058

# ROBUST SUBGAUSSIAN ESTIMATION OF A MEAN VECTOR IN NEARLY LINEAR TIME

BY JULES DEPERSIN [1] AND GUILLAUME LECUÉ [1]

[1] *CREST, ENSAE, IPParis, jules.depersin@ensae.fr; guillaume.lecue@ensae.fr*

We construct an algorithm for estimating the mean of a heavy tailed random variable when given an adversarial corrupted sample of $N$ independent observations. The only assumption we make on the distribution of the non-corrupted (or *informative*) data is the existence of a covariance matrix $\Sigma$, unknown to the statistician. Our algorithm outputs $\hat{\mu}$ which is robust to the presence of $|\mathcal{O}|$ adversarial outliers and satisfies

$$\|\hat{\mu}-\mu\|_2 \lesssim \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} K}{N}} \tag{1}$$

with probability at least $1-\exp(-c_0 K)-\exp(-c_1 u)$, and runtime $\tilde{\mathcal{O}}(Nd+uKd)$ where $K \in \{600|\mathcal{O}|,\dots,N\}$ and $u \in \mathbb{N}^*$ are two parameters of the algorithm. The algorithm is fully data-dependent and does not use (1) in its construction which combines recently developed tools for median-of-means estimators and covering semidefinite Programming. We also show that this algorithm can automatically adapt to the number of outliers (adaptive choice of $K$) and that it satisfies the same bound in expectation.

**1. Introduction on the robust mean vector estimation problem.** Estimating the mean of a random variable in a $d$-dimensional space when given some of its realizations is arguably the oldest and most fundamental problem of statistics. In the past few years, it has received important attention from two communities: the statistics [7, 48, 10, 9, 47, 49, 46, 32, 13, 42, 14] and computer science [20, 19, 23, 21, 25, 26, 12, 24, 33] communities. Both communities consider the problem of *robust mean estimation*, focusing mainly on different definitions of robustness.

The first work to raise the question of robust mean estimation are Huber's [34, 35], Tukey's [56, 57] or Hampel's [31, 30]. Their concerns was more about robustness to model misspecification and on the breakdown point property ("smallest amount of contamination necessary to upset an estimator entirely" taken from [27]). The computational problem connected to this issue was not of primary interest even though it was already raised, for instance, in Section 5.3 from [27] for the construction of Tukey contours (a $d$-dimensional definition of quantiles).

In recent years, many efforts have been made by the statistics community on the construction of estimators performing in a *subgaussian way* for heavy-tailed data. Such estimators achieve the same statistical properties as the empirical mean $\bar{X}_N$ of $(X_1,\cdots,X_N)$, a $N$-sample of i.i.d. Gaussian variables $\mathcal{N}(\mu,\Sigma)$ where $\mu \in \mathbb{R}^d$ and $\Sigma \succeq 0$ is the covariance matrix. In that case, for a given confidence $1-\delta$, the subgaussian rate as defined in [47] is (up to an absolute multiplicative constant)

$$r_\delta = \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op}\log(1/\delta)}{N}} \tag{2}$$

*MSC 2010 subject classifications:* Primary 62F35, 62G08; secondary 62C20, 62G05, 62G20
*Keywords and phrases:* Empirical processes, robust statistics, Algorithms, heavy-tailed data

where $\mathrm{Tr}(\Sigma)$ is the trace of $\Sigma$ and $\|\Sigma\|_{op}$ is the operator norm of $\Sigma$. Indeed, it follows from Borell-TIS's inequality (see Theorem 7.1 in [40] or pages 56-57 in [41]) that with probability at least $1-\delta$,

$$\left\|\bar{X}_N - \mu\right\|_2 = \sup_{\|v\|_2 \leqslant 1} \left\langle \bar{X}_N - \mu, v \right\rangle \leqslant \mathbb{E} \sup_{\|v\|_2 \leqslant 1} \left\langle \bar{X}_N - \mu, v \right\rangle + \sigma\sqrt{2\log(1/\delta)}$$

where $\sigma = \sup_{\|v\|_2 \leqslant 1} \sqrt{\mathbb{E}\left\langle \bar{X}_N - \mu, v \right\rangle^2}$ is the weak variance of the Gaussian process. It is straightforward to check that $\mathbb{E}\sup_{\|v\|_2 \leqslant 1}\left\langle \bar{X}_N - \mu, v \right\rangle \leqslant \sqrt{\mathrm{Tr}(\Sigma)/N}$ and $\sigma = \sqrt{\|\Sigma\|_{op}/N}$, which leads to the rate in (2) (up to the constant $\sqrt{2}$ on the second term in (2)). In most of the recent works, the effort has been made to achieve the rate $r_\delta$ for i.i.d. heavy-tailed data even under the minimal requirement that the data only have a second moment. Under this second-moment assumption only, the empirical mean cannot[1] achieve the rate (2) and one needs to consider other procedures. Over the years, some procedures have been proposed to achieve such a goal: it started with [8] and [45], then, a Le Cam test estimator, called a tournament estimator in [47], a minmax median-of-means estimator in [46] and a PAC-Bayesian estimator in [9] were constructed. The constructions in [45, 47, 46] are based on the median-of-means principle, a technique that we will also use.

On the other side, the computer science (CS) community mostly considers a different definition of robustness and targets a different goal. In many recent CS papers, tractable algorithms (and not only theoretical estimators) have been constructed and proved to be robust with respect to *adversarial contamination* of the dataset that is when some of the data are replaced by other data which may have nothing to do with the original batch and which can even be adversarial. This covers the Huber $\epsilon$-contamination model [35] and also the $\mathcal{O} \cup \mathcal{I}$ framework from [38, 39, 46]. We recall now this adversarial contamination model together with the heavy-tailed setup which will serve as our unique assumption in this work.

ASSUMPTION 1. There exists $N$ random vectors $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}^d$ which are independent with mean $\mu$ and covariance matrix $\mathbb{E}(\tilde{X}_i - \mu)(\tilde{X}_i - \mu)^\top \preceq \Sigma$ where $\Sigma$ is an unknown covariance matrix. The $N$ random vectors $(\tilde{X}_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives the modified dataset $(X_i)_{i=1}^N$ to the statistician. Hence, the statistician receives an "adversarially" contaminated dataset of $N$ vectors in $\mathbb{R}^d$ which can be partitioned into two groups: the modified data $(X_i)_{i\in\mathcal{O}}$, which can be seen as outliers and the "good data" or inliers $(X_i)_{i\in\mathcal{I}}$ such that $\forall i \in \mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \dots, N\}$ is unknown to the statistician.

In the adversarial contamination model from Assumption 1, the set $\mathcal{O}$ can depend arbitrarily on the initial data $(\tilde{X}_i)_{i=1}^N$; the corrupted data $(X_i)_{i\in\mathcal{O}}$ can have any arbitrary dependance structure; and the informative data $(X_i)_{i\in\mathcal{I}}$ may also be correlated (for instance, it is the case, in general, when the $|\mathcal{O}|$ data $\tilde{X}_i$ with largest $\ell_2^d$-norm are modified by the adversary). The computer science community looks at the problem of robust mean estimation from algorithmic perspectives such as the running time in this contamination model. A typical result in this line of research is Theorem 1.3 from [12] that we recall now.

[1] Under only a second-moment assumption, the empirical mean achieves the rate $\sqrt{\mathrm{Tr}(\Sigma)/(\delta N)}$ which can not be improved in general, see [8].

THEOREM 1.1 (Theorem 1.3, [12]). *Let $X_1, \dots, X_N$ be a data points in $\mathbb{R}^d$ following Assumption 1. We assume that the covariance matrix $\Sigma$ of the inliers satisfies $\Sigma \preceq \sigma^2 I_d$. We assume that $\epsilon = |\mathcal{O}|/N$ is such that $0 < \epsilon < 1/3$ and $N \gtrsim d\log(d)/\epsilon$. There exists an algorithm running in $\tilde{\mathcal{O}}(Nd)/\mathrm{poly}(\epsilon)$ which outputs $\hat{\mu}_\epsilon$ such that with probability at least $9/10$, $\|\hat{\mu}_\epsilon - \mu\|_2 \lesssim \sigma\sqrt{\epsilon}$.*

The notation $\tilde{\mathcal{O}}(Nd)$ stands for the computational running time of an algorithm up to $\log(Nd)$ factors. The first result proving the existence of a polynomial time algorithm robust to adversarial contamination may be found in [20] and the first achieving such a result under only a second moment assumption may be found in [22]. Theorem 1.1 improves upon many existing results since it achieves the optimal information theoretic-lower bound with a (nearly) linear-time algorithm.

Finally, there are two recent papers for which both algorithmic and statistical considerations are important. In [32, 13], algorithms achieving the subgaussian rate in (2) have been constructed. They both run in polynomial time: $\mathcal{O}(N^{24} + Nd)$ for [32] and $\mathcal{O}(N^4 + N^2 d)$ for [13] (see [13] for more details on these running times). They do not consider a contamination of the dataset even though their results easily extend to this setup. Some other estimators which have been proposed in the statistics literature are very fast to compute but they do not achieve the optimal subgaussian rate from (2). A typical example is Minsker's geometric median estimator [48] which achieves the rate $\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)/N}$ in linear time $\tilde{\mathcal{O}}(Nd)$. All the later three papers use the median-of-means principle. We will also use this principle. What we mainly borrow from the literature on MOM estimators is the advantage to work with local block means instead of the data themselves. We will identify two such advantages by doing so: a stochastic one and a computational one (see Remark 4 below for more details).

The aim of this work is to show that a single algorithm can answer the three problems: robustness to heavy-tailed data, to adversarial contamination and computational cost. Assumption 1 covers the two concepts of robustness considered in the statistics and computer science communities since the *informative data* (data indexed by $\mathcal{I}$) are only assumed to have a second moment and there are $|\mathcal{O}|$ adversarial outliers in the dataset. Our aim is to show that the rate of convergence (2) which is the rate achieved by the empirical mean in the ideal i.i.d. Gaussian case can be achieved in the corrupted and heavy-tailed setup from Assumption 1 with a fast algorithm: we construct an algorithm running in time $\tilde{\mathcal{O}}(Nd + u\log(1/\delta)d)$ which outputs an estimator of the true mean achieving the subgaussian rate (2) with confidence $1 - \delta - (1/10)^u$ (for $\exp(-c_0 N) \leqslant \delta \leqslant \exp(-c_1|\mathcal{O}|)$) on a corrupted database and under a second moment assumption only. It is therefore robust to heavy-tailed data and to contamination. Our approach takes ideas from both communities: the median-of-means principle which has been recently used in the statistics community and a SDP relaxation from [12] which can be theoretically computed fast. The baseline idea is to construct $K$ equal size groups of data from the $N$ given ones and to compute their empirical means $\bar{X}_k, k = 1, \dots, K$. These $K$ empirical means are used successively to find a robust descent direction thanks to a SDP relaxation from [12]. We prove the robust subgaussian statistical property of the resulting descent algorithm under only the Assumption 1.

The paper is organized as follows. In the next section, we give a high-level description of the algorithm and summarize its statistical and computation performance in our main result Theorem 2.1. We also clearly identify how it improves upon existing results on the same subject. In Section 3, we prove its statistical properties and give a precise definition of the algorithm. In Section 4, we study the statistical performance of the SDP relaxation at the heart of the descent direction. In Section 5, we fully characterize its computational cost. In Section 6, we construct a procedure achieving the same statistical properties and can

automatically adapt to the number of outliers. This latter adaptive procedure is also proved to satisfy estimation results in expectation.

We will use the following notation $[n] = \{1, \dots, n\}$ for any $n \in \mathbb{N}$ and $\ell_2^d$ stands for the Euclidean space $\mathbb{R}^d$ endowed with its canonical Euclidean norm $\|\cdot\|_2 : x = (x_j)_{j=1}^d \in \mathbb{R}^d \to \left(\sum_j x_j^2\right)^{1/2}$. A $\ell_2^d$-ball centered in $x \in \mathbb{R}^d$ with radius $r > 0$ is denoted by $B_2^d(x, r)$, the $\ell_2^d$ unit ball is denoted by $B_2^d$ and the $\ell_2^d$ unit sphere is denoted by $\mathcal{S}_2^{d-1}$.

**2. Construction of the algorithms and main result.** The construction of our robust subgaussian descent procedure is using two ideas. The first one comes from the median-of-means (MOM) approach which has recently received a lot of attention in the statistical and machine learning communities [6, 45, 18, 50, 48]. The MOM approach [51, 3, 36, 4] often yields robust estimation strategies (but usually at a high computational cost). Let us give the general idea behind that approach: we first randomly split the data into $K$ equal-size blocks $B_1, \dots, B_K$ (if $K$ does not divide $N$, we just remove some data). We then compute the empirical mean within each block: for $k = 1, \dots, K$,

$$\bar{X}_k = \frac{1}{|B_k|} \sum_{i \in B_k} X_i$$

where we set $|B_k| = \mathrm{Card}(B_k) = N/K$. In the one-dimensional case, we then take the median of the latter $K$ empirical means to construct a robust *and subgaussian* estimator of the mean [18]. It is more complicated in the multi-dimensional case, where there is no *definitive* equivalent of the one dimensional median but instead there are several candidates: coordinate-wise median, the geometric median (also known as Fermat point), the Tukey Median, among many others (see [55]). The strength of this approach is the robustness of the median operator, which leads to good statistical properties even on corrupted databases. For the construction of our algorithm, we use the idea of grouping the data and compute iteratively some median of the bucketed means $\bar{X}_k, k = 1, \dots, K$.

In [13], the authors propose to use these block means for a gradient descent algorithm: at the current point $x_c$ of the iterative algorithm, a "robust descent direction" well aligned with $x_c - \mu$ is constructed with high probability. Note that $x_c - \mathbb{E}X$ is the best descent direction towards $\mathbb{E}X$ starting from $x_c$; we can also re-write that as a matrix problem: a top eigenvector (i.e. an eigenvector associated with the largest singular value) of $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ is the optimal descent direction $(x_c - \mathbb{E}X)/\|x_c - \mathbb{E}X\|_2$. As a consequence, a top eigenvector of a solution to the optimization problem

$$\underset{M \geq 0, \mathrm{Tr}(M)=1}{\mathrm{argmax}} \left\langle M, (\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top \right\rangle \tag{3}$$

also yields the best descent direction we are looking for (note that $\left\langle A, B \right\rangle = \mathrm{Tr}(A^\top B)$ is the inner product between two matrices $A$ and $B$). Optimization problem (3) may be seen as a SDP relaxation for the problem of finding a top eigenvector and it is the reason why we go into SDP optimization techniques. Recently, this SDP relaxation has been bypassed thanks to the power method in [42] whose aims is also to approximate a top eigenvector.

Of course, we don't know $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ in (3) but we are given a database of $N$ data $X_1, \dots, X_N$ (among which $|\mathcal{I}|$ of them have mean $\mu$). We use these data to estimate in a robust way the unknown quantity $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ in (3). Ideally, we would like to identify the *informative data* and their block means $(1/|\mathcal{K}|) \sum_{k \in \mathcal{K}} (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top$, where $\mathcal{K} = \{k : B_k \bigcap \mathcal{O} = \varnothing\}$, to estimate this quantity but this information is not available either.

To address this problem we use a tool introduced in [12, 20] adapted to the block means. The idea is to endow each block mean $\bar{X}_k$ with a weight $\omega_k$ taken in $\Delta_K$ defined as

$$\Delta_K = \left\{ (\omega_k)_{k=1}^K : 0 \leqslant \omega_k \leqslant \frac{1}{9K/10}, \sum_{k=1}^K \omega_k = 1 \right\}.$$

Ideally we would like to put 0 weights to all block means $\bar{X}_k$ corrupted by outliers. But, we cannot do it since $\mathcal{K}$ is unknown. To overcome this issue, we learn the optimal weights and consider the following minmax optimization problem

$$(E_{x_c}) \qquad \max_{M \geq 0, \mathrm{Tr}(M)=1} \min_{w \in \Delta_K} \left\langle M, \sum_{k=1}^K \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \right\rangle.$$

This is the dual problem from [12] adapted to the block means. The key insight from [12] is that an approximate solution $M_c$ of the maximization problem in $(E_{x_c})$ can be obtained in a reasonable amount of time using a covering SDP approach [12, 53] (see Section 4). We expect a solution (in $M$) to $(E_{x_c})$ to be close to a solution of the minimization problem in (3) – which is $M^* = (\mu - x_c)(\mu - x_c)^\top / \|\mu - x_c\|_2^2$ – and the same for their top eigenvectors (up to the sign). We note that in order to find a good descent direction the authors of [13] also use a (different) SDP relaxation. Theirs costs $\mathcal{O}(N^4 + Nd)$ to be computed.

At a high level description, the robust descent algorithm we perform outputs $\hat{\mu}_K$ after at most $\log d$ iterations of the form $x_c - \theta_c v_1$ where $v_1$ is a top eigenvector of an approximate solution $M_c$ to the problem $(E_{x_c})$ and $\theta_c$ is a step size. It starts at the coordinate-wise median of the bucketed means $\bar{X}_1, \ldots, \bar{X}_K$ . In Algorithm 4, we define precisely the step size and the stopping criteria we use to define the algorithm (it requires too much notation to be defined at this stage). This algorithm outputs the vector $\hat{\mu}_K$ whose running time and statistical performance are gathered in the following result.

THEOREM 2.1. *Grant Assumption 1. Let $K \in \{1, \ldots, N\}$ be the number of equal-size blocks and assume that $K \geqslant 300|\mathcal{O}|$. Let $u \in \mathbb{N}^*$ be a parameter of the covering SDP used at each descent step. With probability at least $1 - \exp(-K/180000) - (1/10)^u$, the descent algorithm finishes in time $\tilde{\mathcal{O}}(Nd + Kud)$ and outputs $\hat{\mu}_K$ such that*

$$\|\hat{\mu}_K - \mu\|_2 \leqslant 808 \left( 1200 \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200 \|\Sigma\|_{op} K}{N}} \right).$$

To make the presentation of the proof of Theorem 2.1 as simple as possible we did not optimize the constants (better constants have been obtained in [8, 9]). Theorem 2.1 generalizes and improves Theorem 1.1 in several ways. We first improve the confidence from a constant "$9/10$" to an exponentially large confidence $1 - \exp(-c_0 K)$ (when $u \sim K$), which was a major technical challenge (note however that the confidence $9/10$ in [11] can be increased to any desired confidence at the expense of deteriorating the rate of convergence – see footnote of page 2 in [11]). We obtain the result for any covariance structure $\Sigma$ and $\hat{\mu}_K$ does not require the knowledge of $\Sigma$ for its construction. We obtain a result which holds for any $N$ (even in the case where $N \leqslant d$). The construction of $\hat{\mu}_K$ does not require the knowledge of the exact proportion of outliers $\epsilon$ in the dataset, but it requires an upper bound in the number of outlier, so that we can chose $K \gtrsim |\mathcal{O}|$. Moreover, using a Lepskii adaptation method [44, 43] it is also possible to automatically choose $K$ and therefore to adapt to the proportion of outliers if we have some extra knowledge on $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ (see Section 6 for more details). Moreover, if we only care about constant $9/10$ confidence, our runtime does not depend on $\epsilon$ and is nearly-linear $\tilde{\mathcal{O}}(Nd)$. We also refer the reader to Corollary 2 for more comparison with Theorem 1.1.

REMARK 1 (Nearly-linear time). We identify two important situations where the algorithm from Theorem 2.1 runs in nearly-linear time, that is, in time $\tilde{\mathcal{O}}(Nd)$. First, when the number of outliers is known to be less than $\sqrt{N}$, we can choose $K \leqslant \sqrt{N}$ and $u = K$. In that case, the algorithm runs in time $\tilde{\mathcal{O}}(Nd)$ and the subgaussian rate is achieved with probability at least $1 - 2\exp(-c_0 K)$ for some constant $c_0$ (see also Corollary 3 for an adaptive to $K$ version of this result). Another widely investigated situation is when we only want to have a constant confidence like $9/10$ as it is the case in the CS community such as in Theorem 1.1. In that case, one may choose $u = 1$ and any values of $K \in [N]$ can be chosen (so we can have any number of outliers up to a $N/300$) to achieve the rate in Theorem 2.1 with constant probability and in nearly-linear time $\tilde{\mathcal{O}}(Nd)$ (see also Corollary 2 for an adaptive to $K$ version of this result). Finally, it is possible to get a subgaussian estimator for the all range of $K \in [N]$ which is also robust to adversarial outliers up to a constant fraction of $N$ when we take $u = K$. In that case, the running time is $\tilde{\mathcal{O}}(Nd + K^2 d)$ which is at worst $\tilde{\mathcal{O}}(N^2 d)$. So algorithm outputs $\hat{\mu}_K$ in time between $\tilde{\mathcal{O}}(Nd)$ and $\tilde{\mathcal{O}}(N^2 d)$ depending on the number of outliers and the probability deviation certifying the result we want.

Theorem 2.1 improves the result from [32, 13] since $\hat{\mu}_K$ runs faster than the polynomial times $\mathcal{O}(N^{24} + Nd)$ and $\mathcal{O}(N^4 + Nd)$ in [32] and [13]. The algorithm $\hat{\mu}_K$ also does not require the knowledge of $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$. Finally, Theorem 2.1 provides running time guarantees on the algorithm unlike in [47, 46, 9] and it improves upon the statistical performance from [48]. The main technical novelty lies in Proposition 1, necessary to improve analysis from [12] toward exponentially large confidence $1 - \exp(-c_0 K)$. Proposition 1 may be of independent interest. Theorem 2.1 also improves the running time in [12] $\tilde{\mathcal{O}}(Nd/\epsilon^6)$ and the constant probability deviation (see Theorem 1.1 for more details) – both probability estimates and computational time have been improved by using bucketed means in place of the data themselves (see Remark 4 below for more details). The computational time improvement from Theorem 2.1 upon the one in [13] is due to the use of covering SDP [1, 53, 11] at each iteration of the robust gradient descent algorithm. Very recent works [42, 33, 16] obtain similar results to the one of Theorem 2.1. They were also able to replace SDPs by spectral methods for the computations of a robust descent direction at each step. Even though cover SDPs are from a theoretical point of view computationally efficient [1, 53] they are notoriously difficult to implement in practice whereas the power methods used in [42, 33, 16] open the door to implementable algorithms. For more references on robust mean estimation, we refer the reader to the survey [24].

**3. Proof of the statistical performance in Theorem 2.1.** In this section, we prove the statistical performance of $\hat{\mu}_K$ as stated in Theorem 2.1. We first identify an event $\mathcal{E}$ onto which we will derive the rate of convergence of the order of (2). This event is also used to compute the running time of $\hat{\mu}_K$ in the next section as announced in Theorem 2.1.

PROPOSITION 1. *Denote by $\mathcal{E}$ the event onto which for all symmetric matrices $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$, there are at least $9K/10$ of the blocks for which $\left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leqslant 8r$ where*

$$r = 1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200\,\|\Sigma\|_{op} K}{N}}. \tag{4}$$

*If Assumptions 1 holds and $K \geqslant 300|\mathcal{O}|$ then $\mathbb{P}[\mathcal{E}] \geqslant 1 - \exp(-K/180000)$.*

Proposition 1 contains all the stochastic arguments we will use in this paper (constants have not been optimized). In other words, after identifying the event $\mathcal{E}$, all the remaining

arguments do not involve any other stochastic tools. The proof of Proposition 1 is based on a rounding argument similar to the one used to prove Grothendieck's inequality [29, 54] or in the Goemans and Williamson's analysis of a SDP relaxation of the Max-Cut problem [28] or in Nesterov's theorem [52]. Before proving Proposition 1, let us first state a result that is of particular interest beyond our problem.

COROLLARY 1. *On the event $\mathcal{E}$, for all symmetric matrices $M \in \mathbb{R}^{d\times d}$ such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ there are at least $9K/10$ blocks $k$ for which $\left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leqslant 8r$ and for all such $k$'s and all $x_c \in \mathbb{R}^d$,*

$$\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r \leqslant \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2 \leqslant \left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r. \tag{5}$$

Let us now turn to a proof of Proposition 1. We first remark that if we were to only consider matrices $M$ of rank 1, Proposition 1 would boil down to showing that for all $v \in \mathcal{S}_2^{d-1}$ (the unit sphere in $\ell_2^d$) on more than $9K/10$ blocks $\left|\left\langle v, \bar{X}_k - \mu\right\rangle\right| \leqslant 8r$. This is a "classical" result in the MOM literature which has been proved in [47] and [46]. We recall now this result and the short proof from [46] adapted to the adversarial contamination setup from Assumption 1. We will use it to prove Proposition 1.

LEMMA 1. *Grant Assumption 1 and assume that $K \geqslant 300|\mathcal{O}|$. With probability at least $1 - \exp(-K/180000)$, for all $v \in \mathcal{S}_2^{d-1}$, there are at least $99K/100$ of the blocks $k$ such that $\left|\left\langle v, \bar{X}_k - \mu\right\rangle\right| \leqslant r$.*

**Proof.** We use the notation introduced in Assumption 1 and we considered the following bucketed means $\overline{\tilde{X}}_k = |B_k|^{-1}\sum_{i\in B_k}\tilde{X}_i$ for $k \in [K]$. They are the $K$ means constructed on the $N$ independent vectors $\tilde{X}_i, i \in [N]$ before contamination (whereas $\bar{X}_k$ are the ones constructed after contamination).

In the following, we show that with probability at least $1 - \exp(-K/180000)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\sum_{k\in[K]} I(\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r) \leqslant \frac{2K}{300}. \tag{6}$$

The result from Lemma 1 follows from (6) because the adversary is allowed to change at most $|\mathcal{O}|$ data points among the $\tilde{X}_i$'s. Hence, there are at most $|\mathcal{O}|$ bucketed means $\overline{\tilde{X}}_k$ containing an outliers and so $K - |\mathcal{O}| \geqslant 299K/300$ means $\overline{\tilde{X}}_k$ which are unchanged that is for which $\overline{\tilde{X}}_k = \bar{X}_k$. So, if (6) holds then they are at least $298K/300$ means $\overline{\tilde{X}}_k$ for which $\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| \leqslant r$ and so, at least $297K/300 = 99K/100$ means $\bar{X}_k$ for which $\left|\left\langle \bar{X}_k - \mu, v\right\rangle\right| \leqslant r$.

As in [37], we define $\phi(t) = 0$ if $t \leqslant 1/2$, $\phi(t) = 2(t - 1/2)$ if $1/2 \leqslant t \leqslant 1$ and $\phi(t) = 1$ if $t \geqslant 1$. We have $I(t \geqslant 1) \leqslant \phi(t) \leqslant I(t \geqslant 1/2)$ for all $t \in \mathbb{R}$ and so

$$\begin{aligned}
&\sum_{k\in[K]} I(\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r) \\
&\leqslant \sum_{k\in[K]} I(\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r) - \mathbb{P}[\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r/2] + \mathbb{P}[\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r/2] \\
&\leqslant \sum_{k\in[K]} \phi\left(\frac{\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right|}{r}\right) - \mathbb{E}\phi\left(\frac{\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right|}{r}\right) + \mathbb{P}[\left|\left\langle \overline{\tilde{X}}_k - \mu, v\right\rangle\right| > r/2]
\end{aligned}$$

$$\leqslant \sup_{v\in\mathcal{S}_2^{d-1}}\left(\sum_{k\in[K]}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)-\mathbb{E}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)\right)+\sum_{k\in[K]}\mathbb{P}[|\langle\overline{\tilde{X}}_k-\mu,v\rangle|>r/2].$$

For all $k\in[K]$, we have

$$\mathbb{P}[|\langle\overline{\tilde{X}}_k-\mu,v\rangle|>r/2]\leqslant\frac{\mathbb{E}\langle\overline{\tilde{X}}_k-\mu,v\rangle^2}{(r/2)^2}\leqslant\frac{4Kv^\top\Sigma v}{Nr^2}$$
$$\leqslant\frac{4K\sup_{v\in\mathcal{S}_2^{d-1}}v^\top\Sigma v}{Nr^2}=\frac{4K\|\Sigma\|_{op}}{Nr^2}\leqslant\frac{1}{300}$$

because $r^2\geqslant 1200K\|\Sigma\|_{op}/N$.

Next, we use several tools from empirical process theory and in particular, for a symmetrization argument, we consider a family of $N$ independent Rademacher variables $(\epsilon_i)_{i=1}^N$ independent of the $(\tilde{X}_i)_{i=1}^N$. In *(bdi)* below, we use the bounded difference inequality (Theorem 6.2 in [5]). In *(sa-cp)*, we use the symmetrization argument and the contraction principle (Chapter 4 in [41]) – we refer to the supplementary material of [46] for more details. We have, with probability at least $1-\exp(-K/180000)$,

$$\sup_{v\in\mathcal{S}_2^{d-1}}\left(\sum_{k\in[K]}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)-\mathbb{E}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)\right)$$
$$\overset{(bdi)}{\leqslant}\mathbb{E}\sup_{v\in\mathcal{S}_2^{d-1}}\left(\sum_{k\in[K]}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)-\mathbb{E}\phi\left(\frac{|\langle\overline{\tilde{X}}_k-\mu,v\rangle|}{r}\right)\right)+\sqrt{\frac{K^2}{360000}}$$
$$\overset{(sa-cp)}{\leqslant}\frac{4K}{Nr}\mathbb{E}\sup_{v\in\mathcal{S}_2^{d-1}}\langle v,\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu)\rangle+\frac{K}{600}$$
$$=\frac{4K}{\sqrt{N}r}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu)\right\|_2+\frac{K}{600}\leqslant\frac{K}{300}$$

because $r\geqslant 1200\mathbb{E}\left\|\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\right\|_2/\sqrt{N}$ since

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu)\right\|_2\leqslant\sqrt{\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu)\right\|_2^2}\leqslant\sqrt{\mathrm{Tr}(\Sigma)}.$$

As a consequence, when $K\geqslant 300|\mathcal{O}|$, with probability at least $1-\exp(-K/180000)$, for all $v\in\mathcal{S}_2^{d-1}$,

$$\sum_{k\in[K]}I(|\langle\overline{\tilde{X}}_k-\mu,v\rangle|>r)\leqslant\frac{|\mathcal{K}|}{300}+\frac{K}{300}\leqslant\frac{2K}{300},$$

which is (6). ∎

**Proof of Proposition 1:** Let $M\in\mathbb{R}^{d\times d}$ be such that $M\succeq 0$ and $\mathrm{Tr}(M)=1$. Denote by $\mathcal{A}_M=\{k\in[K]:\|M^{1/2}(\bar{X}_k-\mu)\|_2\geqslant 8r\}$ and assume that $|\mathcal{A}_M|\geqslant 0.1K$. Let $G$ be a Gaussian vector in $\mathbb{R}^d$ with mean 0 and covariance matrix $M$ (and independent from $X_1,\ldots,X_N$). We consider the random variable $Z=\sum_{k\in[K]}I\left(|\langle\bar{X}_k-\mu,G\rangle|>5r\right)$. We work conditionally to $X_1,\ldots,X_N$ in this paragraph.

For all $k \in [K]$, $\left\langle \bar{X}_k - \mu, G \right\rangle$ is a centered Gaussian variable with variance $\sigma_k^2 := \left\| M^{1/2}(\bar{X}_k - \mu) \right\|_2^2$. In particular, for all $k \in \mathcal{A}_M$, if we denote by $g$ a standard real-valued Gaussian variable, we have $\mathbb{P}_G\left[ |\left\langle \bar{X}_k - \mu, G \right\rangle| > 5r \right] \geqslant \mathbb{P}_G\left[ |\left\langle \bar{X}_k - \mu, G \right\rangle| > 5\sigma_k/8 \right] = 2\mathbb{P}[g > 5/8] \geqslant 0.528$ (where $\mathbb{P}_G$ (resp. $\mathbb{E}_G$) denotes the probability (resp. expectation) w.r.t. $G$ conditionally on $X_1, \ldots, X_N$). Hence, $\mathbb{E}_G Z \geqslant 0.528 |\mathcal{A}_M| \geqslant 0.0528K$. Since $|Z| \leqslant K$ a.s., it follows from Paley-Zygmund inequality (see Proposition 3.3.1 in [15]) that

$$\mathbb{P}_G[Z > 0.01K] \geqslant \frac{(\mathbb{E}_G Z - 0.01K)^2}{\mathbb{E}_G Z^2} \geqslant (0.0428)^2 = 0.0018.$$

Moreover, it follows from the Borell-TIS inequality (see Theorem 7.1 in [40] or pages 56-57 in [41]) that with probability at least $1 - \exp(-8)$, $\|G\|_2 \leqslant \mathbb{E}\|G\|_2 + 4\sqrt{\|M\|_{op}}$. Moreover, $\mathbb{E}\|G\|_2 \leqslant \sqrt{\mathrm{Tr}(M)} \leqslant 1$ and $\|M\|_{op} \leqslant \mathrm{Tr}(M) \leqslant 1$, so $\|G\|_2 \leqslant 5$ with probability at least $1 - \exp(-8) \geqslant 0.9996$. Since $0.9996 + 0.0018 > 1$ there exists a vector $G_M \in \mathbb{R}^d$ such that $\|G_M\|_2 \leqslant 5$ and $\sum_{k \in [K]} I\left( |\left\langle \bar{X}_k - \mu, G_M \right\rangle| > 5r \right) > 0.01K$. We recall that this latter result holds when we assume that $|\mathcal{A}_M| \geqslant 0.1K$.

Next, we denote by $\Omega_0$ the event onto which for all $v \in \mathcal{S}_2^{d-1}$, there are at least $99K/100$ blocks such that $|\left\langle \bar{X}_k - \mu, v \right\rangle| \leqslant r$. We know from Lemma 1 that $\mathbb{P}[\Omega_0] \geqslant 1 - \exp(-K/180000)$. Let us place ourselves on the event $\Omega_0$ up to the end of the proof. Let $M \in \mathbb{R}^{d \times d}$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ and assume that $|\mathcal{A}_M| \geqslant 0.1K$. It follows from the first paragraph of the proof that there exists $G_M \in \mathbb{R}^d$ such that $\|G_M\|_2 \leqslant 5$ and $\sum_{k \in [K]} I\left( |\left\langle \bar{X}_k - \mu, G_M \right\rangle| > 5r \right) > 0.01K$. Given that we work on the event $\Omega_0$, we have for $v_M = G_M / \|G_M\|_2$, that for more than $99K/100$ blocks $|\left\langle \bar{X}_k - \mu, v_M \right\rangle| \leqslant r$ and so $|\left\langle \bar{X}_k - \mu, G_M \right\rangle| \leqslant \|G_M\|_2 r \leqslant 5r$ which contradicts the fact that $\sum_{k \in [K]} I\left( |\left\langle \bar{X}_k - \mu, G_M \right\rangle| > 5r \right) > 0.01K$. Therefore, we necessarily have $|\mathcal{A}_M| \leqslant 0.1K$, which concludes the proof. ■

**Proof of Corollary 1:** Let us assume that the event $\mathcal{E}$ holds up to the end of the proof. Let $M \in \mathbb{R}^{d \times d}$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Let $\mathcal{K}_M = \{k \in [K] : \left\| M^{1/2}(\bar{X}_k - \mu) \right\|_2 \leqslant 8r\}$. On the event $\mathcal{E}$, we have $|\mathcal{K}_M| \geqslant 9K/10$. Let $x_c \in \mathbb{R}^d$. For all $k \in \mathcal{K}_M$, we have $\left\| M^{1/2}(\mu - \bar{X}_k) \right\|_2 \leqslant 8r$ and so

$$\begin{aligned} \left\| M^{1/2}(\bar{X}_k - x_c) \right\|_2 &\in \left\| M^{1/2}(\mu - x_c) \right\|_2 + \left[ -\left\| M^{1/2}(\mu - \bar{X}_k) \right\|_2, \left\| M^{1/2}(\mu - \bar{X}_k) \right\|_2 \right] \\ &\subset \left\| M^{1/2}(x_c - \mu) \right\|_2 + [-8r, 8r]. \end{aligned}$$

■

Let us now turn to the study of the optimization problem $(E_{x_c})$ on the event $\mathcal{E}$. Like in [12], we denote by $OPT_{x_c}$ the optimal value of $(E_{x_c})$ and by

$$h_{x_c} : M \to \min_{w \in \Delta_K} \left\langle M, \sum_{k \in [K]} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \right\rangle$$

its objective function to be minimized over $\{M \in \mathbb{R}^{d \times d} : M \succeq 0, \mathrm{Tr}(M) = 1\}$.

REMARK 2. For a given $M$, the optimal choice of $w \in \Delta_K$ in the definition of $h_{x_c}(M)$ is straightforward: one just have to put the maximum possible weight on the $9K/10$ smallest $\left\langle M, (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \right\rangle, k \in [K]$. Formally, we set $\mathcal{S}_M = \sigma(\{1, 2, \cdots, 9K/10\})$, where

$\sigma$ is a permutation on $[K]$ that arranges the $(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c), k \in [K]$ in ascending order:

$$\left\|M^{1/2}(\bar{X}_{\sigma(1)} - x_c)\right\|_2 \leqslant \left\|M^{1/2}(\bar{X}_{\sigma(2)} - x_c)\right\|_2 \leqslant \cdots \leqslant \left\|M^{1/2}(\bar{X}_{\sigma(K)} - x_c)\right\|_2.$$

Then we get $h_{x_c}(M) = (1/|\mathcal{S}_M|)\sum_{k\in\mathcal{S}_M}(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)$.

The first lemma deals with the optimal value of ($E_{x_c}$) when the current point $x_c$ is far from the mean $\mu$.

LEMMA 2. *On the event $\mathcal{E}$, for all $x_c \in \mathbb{R}^d$, if $\|x_c - \mu\|_2 > 16r$ then*

$$(8/9)(\|x_c - \mu\|_2 - 8r)^2 \leqslant OPT_{x_c} \leqslant (\|x_c - \mu\|_2 + 8r)^2.$$

**Proof.** Let $M$ be a matrix such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Set $\mathcal{K}_M = \{k \in [K] : \|M^{1/2}(\bar{X}_k - \mu)\|_2 \leqslant 8r\}$. On the event $\mathcal{E}$, we have $|\mathcal{K}_M| \geqslant 9K/10$ and it follows from Corollary 1 that for all $k \in \mathcal{K}_M$ and all $x_c \in \mathbb{R}^d$,

$$\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r \leqslant \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2 \leqslant \left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r. \tag{7}$$

Then we define a weight vector $\tilde{\omega} \in \Delta_K$ by setting for all $k \in [K]$

$$\tilde{\omega}_k = \begin{cases} 1/|\mathcal{K}_M| \text{ if } k \in \mathcal{K}_M \\ 0 \qquad\quad \text{else.} \end{cases}$$

It follows from the definition of $h_{x_c}$ and (7) that

$$\begin{aligned} h_{x_c}(M) &\leqslant \sum_{k\in[K]} \tilde{\omega}_k(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) \\ &= \frac{1}{|\mathcal{K}_M|}\sum_{k\in\mathcal{K}_M}\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \leqslant \left(\left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r\right)^2. \end{aligned} \tag{8}$$

Taking the maximum over all $M \in \mathbb{R}^d$ such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ on both side of the latter inequality yields the right-hand side inequality of Lemma 2.

For the left-hand side inequality of Lemma 2, we let $x_c \in \mathbb{R}^d$ be such that $\|x_c - \mu\|_2 > 16r$ and let $M$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. We use the notation and observation from Remark 2: we note that $|\mathcal{K}_M \bigcap \mathcal{S}_M| \geqslant 8K/10$ so that it follows from Corollary 1 that

$$\begin{aligned} h_{x_c}(M) &= \frac{1}{9K/10}\sum_{k\in\mathcal{S}_M}\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \geqslant \frac{1}{9K/10}\sum_{k\in\mathcal{A}_M\bigcap\mathcal{S}_M}\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \\ &\geqslant \frac{8K/10}{9K/10}\left(\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r\right)^2. \end{aligned}$$

Then, taking the maximum over all $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$ on both sides, finishes the proof. ∎

The next lemma shows that the top eigenvector of an approximate solution to ($E_{x_c}$) is aligned with the best possible descent direction $(\mu - x_c)/\|\mu - x_c\|_2$. It is taken from the proof of Lemma 3.3 in [12]. We reproduce here a short proof for completeness.

PROPOSITION 2. *On the event $\mathcal{E}$, if $M$ is a matrix such that $M \succeq 0$, $\mathrm{Tr}(M) = 1$ and $h_{x_c}(M) \geqslant (\beta\|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leqslant \beta \leqslant 1$, then any top eigenvector $v_1$ of $M$ satisfies*

$$\left|\left\langle v_1, \frac{x_c - \mu}{\|x_c - \mu\|_2}\right\rangle\right| > \sqrt{2\beta^2 - 1}.$$

**Proof.** Let $M$ be a matrix such that $M \succeq 0$ , $\mathrm{Tr}(M) = 1$ and $h_{x_c}(M) \geqslant (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leqslant \beta \leqslant 1$. We use the same argument as in the proof of Lemma 2: on the event $\mathcal{E}$, $|\mathcal{K}_M| \geqslant 9K/10$ where $\mathcal{K}_M = \{k \in [K] : \|M^{1/2}(\bar{X}_k - \mu)\|_2 \leqslant 8r\}$ and so $\tilde{\omega} \in \Delta_K$ where for all $k \in [K], \tilde{\omega}_k = 1/|\mathcal{K}_M|$ if $k \in \mathcal{K}_M$ and $\tilde{\omega}_k = 0$ if $k \notin \mathcal{K}_M$. It follows from the definition of $h_{x_c}$ that

$$h_{x_c}(M) \leqslant \sum_{k\in[K]} \tilde{\omega}_k (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c) = \frac{1}{|\mathcal{K}_M|} \sum_{k\in\mathcal{K}_M} \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2$$

and so from Corollary 1, $h_{x_c}(M) \leqslant \left(\left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r\right)^2$. Since, we assumed that $h_{x_c}(M) \geqslant (\beta \|x_c - \mu\|_2 + 8r)^2$, it follows that $\left\|M^{1/2}(\mu - x_c)\right\|_2^2 \geqslant \beta^2 \|\mu - x_c\|_2^2$.

Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d \geqslant 0$ denote the eigenvalues of $M$ and let $v_1, \ldots, v_d$ denote corresponding eigenvectors. The conditions on $M$ imply that $\sum_j \lambda_j = 1$ and $\mathcal{B}_M = (v_1, \ldots, v_d)$ is an orthonormal basis of $\mathbb{R}^d$. We denote $v = (\mu - x_c)/\|\mu - x_c\|_2$. We decompose $v$ in $\mathcal{B}_M$ as $v = \sum_j \alpha_j v_j$ with $\sum_j \alpha_j^2 = 1$. Using this decomposition, we have $v^\top M v = \sum_j \lambda_j \alpha_j^2$. We have $\lambda_1 = \lambda_1 \sum_j \alpha_j^2 \geqslant \sum_j \lambda_j \alpha_j^2 \geqslant \beta^2$, so $\lambda_1 \geqslant \beta^2$. Moreover, since $\sum_j \lambda_j = 1$, we have $\beta^2 \sum_j \alpha_j^2 \leqslant \sum_j \lambda_j \alpha_j^2 \leqslant \lambda_1 \alpha_1^2 + (1 - \lambda_1)(1 - \alpha_1^2) \leqslant \alpha_1^2 + (1 - \beta^2) \sum_j \alpha_j^2$, so we have $\alpha_1^2 \geqslant (2\beta^2 - 1)$. As we know that $\alpha_1 = \langle v_1, v \rangle$, we get the result. ∎

Proposition 2 is the first tool we need to construct a descent algorithm since it provides a descent/ascent direction (depending on the sign of the top eigenvector of an approximate solution to ($E_{x_c}$)). It remains to specify three other quantities to fully characterize our algorithm: a starting point, a step size and a stopping criteria. We start with the starting point. Here we simply use the coordinate-wise median-of-means. The following statistical guarantee on the coordinate-wise median-of-means is known or folklore but we want to put forward that in our case it holds on the event $\mathcal{E}$. This again shows that $\mathcal{E}$ is the only event we need to fully analyze all the building blocks of the algorithm. We recall that the coordinate-wise median-of-means is the estimator $\hat{\mu}^{(0)} \in \mathbb{R}^d$ whose coordinates are for all $j \in [d], \hat{\mu}_j^{(0)} = \mathrm{med}(\bar{X}_{k,j} : k \in [K])$ where $\bar{X}_{k,j}$ is the $j$-th coordinate of the block mean $\bar{X}_k$ for all $k \in [K]$.

PROPOSITION 3. *On the event $\mathcal{E}$, we have $\|\hat{\mu}^{(0)} - \mu\|_2 \leqslant 8\sqrt{d}r$.*

**Proof.** Let us place ourselves on the event $\mathcal{E}$ during all the proof. For all directions, $v \in \mathcal{S}_2^{d-1}$, there are at least $9K/10$ blocks $k$ such that $|\langle \bar{X}_k - \mu, v \rangle| \leqslant 8r$. In particular, for all $j \in [d], |\langle \bar{X}_k - \mu, e_j \rangle| \leqslant 8r$ where $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{R}^d$. That is for at least $9K/10$ blocks $|\bar{X}_{k,j} - \mu_j| \leqslant 8r$. In particular, the latter result is true for the median of $\{\bar{X}_{k,j} : k \in [K]\}$, that is, for $\hat{\mu}_j^{(0)}$. We therefore have $\|\hat{\mu}^{(0)} - \mu\|_\infty \leqslant 8r$ and so $\|\hat{\mu}^{(0)} - \mu\|_2 \leqslant 8r\sqrt{d}$. ∎

Proposition 3 guarantees that starting from the coordinate-wise median-of-means we are off by a $\sqrt{d}$ proportional factor from the optimal rate $r$. This will play a key role to analyze the number of steps we need to reach $\mu$ within the optimal rate $r$. Indeed, if we prove a geometric decay of the distance to $\mu$ along the descent algorithm then only $\log d$ steps (up to a multiplicative constants) would be enough to reach $\mu$ by a distance at most of the order of $r$.

Let us now specify the step size we use at each iteration. At the current point $x_c$ we compute a top eigenvector $v_1$ of an approximate solution $M$ to ($E_{x_c}$) (i.e. $M$ such that $h_{x_c}(M) \geqslant$

$(\beta \|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leqslant \beta \leqslant 1$). The next iteration is $x_{c+1} = x_c - \theta_c v_1$ where the step size is

$$\theta_c = -\mathrm{Med}\left(\left\langle \bar{X}_k - x_c, v_1 \right\rangle : k \in [K]\right). \tag{9}$$

In particular, since $\theta_c v_1$ does not depend on the sign of $v_1$ (the product $\theta_c v_1$ is the same if we replace $v_1$ by $-v_1$), we do not care which top eigenvector of $M$ we choose.

Let us now prove a geometric decay of the algorithm while $x_c$ is far from $\mu$. Again, this result is proved on the event $\mathcal{E}$.

PROPOSITION 4. *On the event $\mathcal{E}$, the following holds. Let $x_c \in \mathbb{R}^d$ (be the current point of the algorithm). Assume that $M$ is an approximate solution of ($E_{x_c}$): $M$ is such that $h_{x_c}(M) \geqslant (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $0.78 \leqslant \beta \leqslant 1$ and let $v_1$ be one of its top eigenvectors. Then, we have*

$$\|x_{c+1} - \mu\|_2^2 \leqslant 0.8 \|x_c - \mu\|_2^2 + 64r^2$$

*when $x_{c+1} = x_c - \theta_c v_1$ for $\theta_c$ defined in (9).*

**Proof.** Let us assume that the event $\mathcal{E}$ holds up to the end of the proof. Let $M$ be an approximate solution to ($E_{x_c}$) such that $h_{x_c}(M) \geqslant (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $0.78 \leqslant \beta \leqslant 1$ and let $v_1$ be a top eigenvector of $M$.

In direction $v_1$, there are at least $9K/10$ blocks such that $|\langle \bar{X}_k - \mu, v_1 \rangle| \leqslant 8r$ (see Lemma 1). Hence, on these blocks, we also have

$$\begin{aligned} |\theta_c - \langle x_c - \mu, v_1 \rangle| &= |\mathrm{Med}\left(\left\langle \mu - \bar{X}_k, v_1 \right\rangle : k \in [K]\right)| \\ &\leqslant \mathrm{Med}\left(|\langle \mu - \bar{X}_k, v_1 \rangle| : k \in [K]\right) \leqslant 8r. \end{aligned} \tag{10}$$

Let $v = (\mu - x_c)/\|\mu - x_c\|_2$ denote the optimal normalized descent direction. We write $v = \lambda_1 v_1 + \lambda_2 v_1^\perp$ where $v_1^\perp$ is a normalized orthogonal vector to $v_1$. We have $\lambda_1^2 + \lambda_2^2 = 1$ and it follows from Proposition 2 that $|\lambda_1| = |\langle v_1, v \rangle| > \sqrt{2\beta^2 - 1}$. We conclude that

$$\begin{aligned} \|x_{c+1} - \mu\|_2^2 = \|x_c - \mu - \theta_c v_1\|_2^2 &= \left\|(\langle x_c - \mu, v_1 \rangle - \theta_c) v_1 + \langle x_c - \mu, v_1^\perp \rangle v_1^\perp\right\|_2^2 \\ &= (\langle x_c - \mu, v_1 \rangle - \theta_c)^2 + \langle x_c - \mu, v_1^\perp \rangle^2 \leqslant (8r)^2 + \lambda_2^2 \|x_c - \mu\|_2^2 \end{aligned}$$

As $\lambda_2^2 = 1 - \lambda_1^2 < 2 - 2\beta^2 < 0.8$ we get the result. ∎

We now have almost all the building blocks to fully characterize the algorithm. The last and final step is to find a stopping rule. The idea we use to design such a rule is based on Proposition 4: we know that when the current point $x_c$ is not in a $\ell_2^d$-neighborhood of $\mu$ with a radius $80r$ then the $\ell_2^d$-distance between the next iteration $x_{c+1}$ and $\mu$ should be less than $\sqrt{0.81}$ times the $\ell_2^d$-distance between $x_c$ and $\mu$ – that is a geometric decay of the distance to $\mu$. Moreover, if the current iteration $x_c$ is in a $\ell_2^d$-ball centered in $\mu$ with the radius $80r$ then, it follows from Proposition 4 that the next iteration $x_{c+1}$ will also be in a $\ell_2^d$-ball centered in $\mu$ with radius at most $80r$. So once the algorithm enters the ball $B_2^d(\mu, 80r)$ it never leaves it. We therefore have a geometric decay of the distance to $\mu$ along the iterations until we reach the ball $B_2^d(\mu, 80r)$. Starting from the coordinate-wise median(-of-means) which is in a $8\sqrt{d}r$ neighborhood of $\mu$ (see Proposition 3), we only have to do $\log(8\sqrt{d})/\log(1/\sqrt{0.81})$ iterations to output a current point which at most $80r$-close to $\mu$ w.r.t. the $\ell_2^d$-norm.

We are now in a position to write an "almost final" pseudo-code of our algorithm. In the next section, we will dive a bit deeper in this pseudo-code (and in particular on the covering SDP algorithm used to construct an approximate solution to ($E_{x_c}$)) in order to provide a final pseudo-code together with its total running time.

**input :** $X_1, \dots, X_N$ and a number $K$ of blocks
**output:** A robust subgaussian estimator of $\mu$

1 Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$
2 Construct the $K$ empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
3 Compute $\hat{\mu}^{(0)}$ the coordinate-wise median-of-means and put $x_c \leftarrow \hat{\mu}^{(0)}$
4 **for** $T = 1, 2, \cdots, \log(8\sqrt{d})/\log(1/\sqrt{0.81})$ **do**
5 Compute $M_c$ an approximate solution to $(E_{x_c})$ such that

$$h_{x_c}(M_c) \geqslant \left(0.78 \left\|x_c - \mu\right\|_2 + 8r\right)^2$$

6 Compute $v_1$ a top eigenvector of $M_c$
7 Compute a step size $\theta_c = -\operatorname{Med}\left(\left\langle \bar{X}_k - x_c, v_1 \right\rangle : k \in [K]\right)$
8 Update $x_c \leftarrow x_c - \theta_c v_1$
9 **end**
10 **Return** $x_c$

**Algorithm 1:** "Almost final" pseudo-code of the robust sub-gaussian estimator of $\mu$

Algorithm 1 is "almost" our final algorithm. There is one last step we need to check carefully: given a current point $x_c$ we need to find a way to construct $M_c$ satisfying "$h_{x_c}(M_c) \geqslant \left(0.78 \left\|x_c - \mu\right\|_2 + 8r\right)^2$" without knowing $r$ or $\mu$. This is the last issue we need to address in order to explain how step **5** from Algorithm 1 can be realized in a fully data-dependent way in a good time. This issue is answered in the next section together with the computation of its running time.

**4. Approximately solving the SDP ($E_{x_c}$).** The aim of this section is to show that, on the event $\mathcal{E}$, it is possible to construct in a reasonable amount of time a matrix $M_c$ such that "$h_{x_c}(M_c) \geqslant \left(0.78 \left\|x_c - \mu\right\|_2 + 8r\right)^2$" without any extra information than the data. To that end we construct in an efficient way an approximate solution to the optimization problem $(E_{x_c})$ using covering SDP as in [12]. The main result of this section is the following.

THEOREM 4.1. *Let $u \in \mathbb{N}^*$. On the event $\mathcal{E}$, for every $x_c \in \mathbb{R}^d$ such that $\left\|x_c - \mu\right\|_2 \geqslant 800r$, given input $x_c$, we can either compute, in time $\tilde{\mathcal{O}}(Kud)$, with probability $> 1 - (1/10)^{u+5}/\sqrt{d}$ :*

- *A matrix $M_c$ such that*

$$h_{x_c}(M_c) \geqslant \left(0.78 \left\|x_c - \mu\right\|_2 + 8r\right)^2$$

- *Or directly a subgaussian estimate of $\mu$, using only the block means $\bar{X}_1, \dots, \bar{X}_K$ as inputs.*

Theorem 4.1 answers the last issue raised at the end of Section 3 and provides the running time for step **5** of Algorithm 1. It therefore concludes the statement that there exists a fully data-driven robust subgaussian algorithm for the estimation of a mean vector under the only Assumption 1 (the total running time of Algorithm 1 is studied in Section 5).

REMARK 3. Theorem 4.1 states that we either find an approximate solution $M_c$ to $(E_{x_c})$ or a good estimate of $\mu$ (at the current point $x_c$). As we will see in this section, this second case is degenerate as it is not the typical situation.

Before turning to the proof of Theorem 4.1, we recall the definition of the following quantities to ease the reading of the proof:

$$OPT_{x_c} = \min_{M \succeq 0 : \mathrm{Tr}(M)=1} h_{x_c}(M) \text{ where } h_{x_c} : M \to \min_{w \in \Delta_K} \langle M, \sum_{k \in [K]} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle$$

and $(E_{x_c})$ refers to the optimization problem $\min_M \left( h_{x_c}(M) : M \succeq 0, \mathrm{Tr}(M) = 1 \right)$.

We now turn to the proof of Theorem 4.1. It is decomposed into several lemmas adapted from techniques developed by [12] to approximately solve the SDP problem $(E_{x_c})$ in time $\tilde{\mathcal{O}}(Kud)$ as announced in Theorem 1.1. To that end, we first introduce the following covering SDP

$$(C_\rho) \qquad \begin{aligned} \underset{M', y'}{\text{minimize}} \quad & \mathrm{Tr}(M') + \|y'\|_1 \\ \text{subject to} \quad & M' \succeq 0,\ y' \geqslant 0, \\ & \forall k \in [K],\ \rho (\bar{X}_k - x_c)^\top M' (\bar{X}_k - x_c) + 9K/10\ y'_k \geqslant 1 \end{aligned}$$

where $\rho > 0$ is some parameter that we will show how to fine-tune later. Then, we show that, for a good choice of $\rho$, we can turn a good approximate solution for $(C_\rho)$ into a good approximate solution for $(E_{x_c})$.

We denote by $g(\rho)$ the optimal objective value of $(C_\rho)$. We begin with a first lemma that shows how to link the two optimization problems $(E_{x_c})$ and $(C_\rho)$. The proof can be found in Lemma 4.2 from [12]. We adapt it here for our purpose.

LEMMA 3. *Let $\rho > 0$. From a feasible solution $(M', y')$ for $(C_\rho)$ that achieves $\mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$, we can construct a feasible solution $M$ for $(E_{x_c})$ with objective value $h_{x_c}(M) \geqslant 1/\rho$. The reverse is also true. In particular, if $g(\rho)$ (resp. $OPT_{x_c}$) denotes the optimal value achieved by the objective function in $(C_\rho)$ (resp. $(E_{x_c})$), we have $g(\rho) \leqslant 1$ iff $1/\rho \geqslant OPT_{x_c}$.*

**Proof.** We first note that the optimization problem $(E_{x_c})$ is equivalent to the following one:

$$(\tilde{E}_{x_c}) \qquad \begin{aligned} \underset{M, y, z}{\text{maximize}} \quad & z - \frac{\|y\|_1}{9K/10} \\ \text{subject to} \quad & M \succeq 0,\ \mathrm{Tr}(M) = 1,\ y \geqslant 0,\ z \geqslant 0 \\ & \forall k \in [K],\ (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c) + \ y_k \geqslant z \end{aligned}$$

Indeed, for a given $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$, one can notice that the optimal value is achieved in $(\tilde{E}_{x_c})$ for $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c)), k \in [K]$ and $z = \mathcal{Q}_{9/10}\left( (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c) \right)$ the 9/10-th quantile of $\{(\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c) : k \in [K]\}$, so that $z - \|y\|_1 / (9K/10) = h_{x_c}(M)$ which gives the equivalence between $(E_{x_c})$ and $(\tilde{E}_{x_c})$.

Then, let a feasible solution $(M', y')$ for $(C_\rho)$ be such that $\mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$. We define

$$M = \frac{M'}{\mathrm{Tr}(M')}, z = \frac{1}{\rho\, \mathrm{Tr}(M')} \text{ and } y = \frac{(9K/10)}{(\rho\, \mathrm{Tr}(M'))} y'.$$

We can check that $(M, y, z)$ is feasible for $(\tilde{E}_{x_c})$ and $z - \|y\|_1 / (9K/10) \geqslant 1/\rho$. Hence, given the equivalence between $(E_{x_c})$ and $(\tilde{E}_{x_c})$, we obtain that $M$ is feasible for $(E_{x_c})$ and that $h_{x_c}(M) \geqslant 1/\rho$.

Conversely, if $M$ is feasible for $(E_{x_c})$ such that $h_{x_c}(M) \geqslant 1/\rho$ then we define $y$ and $z$ such that for all $k \in [K]$, $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c)), k \in [K]$ and

$z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right)$. We check that $(M, y, z)$ is feasible for $(\tilde{E}_{x_c})$ with objective values equals to $h_{x_c}(M)$ and so it is larger than $1/\rho$. Next, by defining

$$M' = \frac{M}{\rho z} \text{ and } y' = \frac{y}{(9K/10)z},$$

we see that $(M', y')$ is feasible for $(C_\rho)$ and its objective values is less than 1. ∎

From Lemma 3, it is enough to solve $(C_\rho)$ – for a good choice of $\rho$ – to find a good approximate solution for $(E_{x_c})$. It therefore remains to find such a good $\rho$. To do so, we rely on the next two lemmas. The first one is adapted from Lemma 4.3 in [12]; we recall that $g(\rho)$ is the optimal value achieved by the objective function in $(C_\rho)$.

LEMMA 4. *For every $\rho > 0$ and $\alpha \in (0,1)$, $g((1-\alpha)\rho) \geqslant g(\rho) \geqslant (1-\alpha)g((1-\alpha)\rho)$.*

**Proof.** A feasible pair $(M', y')$ for $(C_{(1-\alpha)\rho})$ is also feasible for $(C_\rho)$, which gives the first inequality. If $(M', y')$ is a feasible pair for $(C_\rho)$, then $(M'/(1-\alpha), y'/(1-\alpha))$ is a feasible pair for $(C_{(1-\alpha)\rho})$, which gives the second inequality. ∎

It follows from Lemma 4 that $g$ is continuous, non increasing and $g(1/OPT_{x_c}) = 1$ (this follows from Lemma 3 since we have that $g(\rho) \leqslant 1$ iff $1/\rho \geqslant OPT_{x_c}$ and the continuity of $g$). So in order to find a good solution, we must find a $\rho$ such that $g(\rho)$ is as close to 1 as possible. Unfortunately, we do not know how to solve $(C_\rho)$ exactly for a given $\rho > 0$, but we can compute efficiently a good approximation $(M', y')$ and a top eigenvector of $M'$ thanks to the following result which can be found in [53] or [2] and is detailed in [12] (see Section 4 and Remark 3.4).

LEMMA 5. *[[53], [2]] Let $u \geqslant 1$ be an integer. For every $\rho > 0$ and every fixed $\eta > 0$, we can find with probability $> 1 - (1/10)^{u+10}/d$ a feasible solution to $(C_\rho)$ that is $\eta$-close to the optimal, that is to say a feasible pair $(M', y')$ so that $\mathrm{Tr}(M') + \|y'\|_1 \leqslant (1+\eta)g(\rho)$ in time $\tilde{\mathcal{O}}(uKd)$. Moreover, it is possible to find an approximate top eigenvector of $M'$ in $\tilde{\mathcal{O}}(Kd)$.*

We compute $(u + 3\log(d) + 10)$ times independently the (randomized) algorithm from [53] (or the one from [2]) that has a runtime of $\tilde{\mathcal{O}}(Kd)$ and that outputs an $\eta$-close feasible solution with probability $9/10$. By taking the largest of the output's objective value, we have an $\eta$-close feasible solution with probability $1 - (1/10)^{u+3\log(d)+10}$, in time $\tilde{\mathcal{O}}(uKd)$, proving Lemma 5.

Let us call $\texttt{ALG}_\rho$ the algorithm from Lemma 5, that takes as input $((\bar{X}_k)_{k=1}^K, x_c, \rho, \eta, u)$ and returns a feasible pair $(M', y')$ for $(C_\rho)$ satisfying $\mathrm{Tr}(M') + \|y'\|_1 \leqslant (1+\eta)g(\rho)$ in $\tilde{\mathcal{O}}(uKd)$, with probability $> 1 - (1/10)^{u+10}/d$. Next, in order to find a good $\rho$, we have to get some additional information on the function $g$. We will get it on the event $\mathcal{E}$.

LEMMA 6. *On the event $\mathcal{E}$, for all $x_c \in \mathbb{R}^d$, if $\|x_c - \mu\|_2 > 8r$ then*

$$g(\rho) \leqslant \frac{1}{\rho\, OPT_{x_c}}\left(1 + \rho OPT_{x_c}\left(\frac{9(\|x_c-\mu\|_2 + 8r)^2}{8(\|x_c-\mu\|_2 - 8r)^2} - 1\right)\right).$$

**Proof.** We use the same notation as in the proof of Lemma 3. For any $\nu > 0$, we can choose a triplet $(M, y, z)$ feasible for $(\tilde{E}_{x_c})$ such that $z - \|y\|_1/(9K/10) > OPT_{x_c} - \nu$ and $z$ and $y$ are the optimal solutions of the problem $(\tilde{E}_{x_c})$ given by $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)), k \in [K]$ and $z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right)$ the $9/10$-th quantile of $\{(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) : k \in [K]\}$.

On the event $\mathcal{E}$, Lemma 2 yields $OPT_{x_c} > (8/9)(\|x_c-\mu\|_2 - 8r)^2$ and we have from Corollary 1 that

$$z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right) = \mathcal{Q}_{9/10}\left(\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2\right)$$
$$\leqslant \left(\left\|M^{1/2}(x_c-\mu)\right\|_2 + 8r\right)^2 \leqslant (\|x_c-\mu\|_2 + 8r)^2$$

because $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Let $M' = M/(\rho z), y' = y/[z(9K/10)]$. Since $(M', y')$ is feasible for $(C_\rho)$, we have

$$g(\rho) \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant \frac{1+\rho\|y\|_1/(9K/10)}{\rho z}$$
$$< \frac{1+\rho(z - OPT_{x_c} + \nu)}{\rho z} \leqslant \frac{1+\rho\nu + \rho OPT_{x_c}\left(\frac{9(\|x_c-\mu\|_2+8r)^2}{8(\|x_c-\mu\|_2-8r)^2} - 1\right)}{\rho(OPT_{x_c} - \nu)}.$$

By taking $\nu \to 0$, we get the result. ∎

**Proof of Theorem 4.1.** Let us place ourselves on the event $\mathcal{E}$ so that we can apply Lemma 6. Let $x_c \in \mathbb{R}^d$ and assume that $\|x_c - \mu\|_2 > 800r$. It follows from Lemma 6 that $g(\rho) \leqslant 1/(\rho\, OPT_{x_c}) + 0.171$. Therefore, if we can find a $\rho$ such that $g(\rho) \geqslant 1 - \epsilon + 0.171$ for some $0 < \epsilon < 1$, then necessarily $1/\rho \geqslant OPT_{x_c}(1-\epsilon)$. Let us take $\epsilon = 0.173$, and $\eta = 0.0001$. Then if $\texttt{ALG}_\rho$ returns a feasible pair $(M', y')$ for $(C_\rho)$ so that $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$, then, since $0.9981 > 1.0001 \times 0.998 = (1+\eta)(1-\epsilon+0.171)$ we will know that, with probability $> 1 - (1/10)^{u+10}/d$,

$$(1+\eta)g(\rho) \geqslant \mathrm{Tr}(M') + \|y'\|_1 \geqslant (1+\eta)(1-\epsilon+0.171)$$

hence $1/\rho \geqslant OPT_{x_c}(1-\epsilon)$, and by Lemma 3, we can construct a feasible solution $M_c$ for $(E_{x_c})$ with objective value satisfying $h_{x_c}(M_c) \geqslant OPT_{x_c}(1-\epsilon)$. Next, using Lemma 2, we obtain that when $\|x_c - \mu\|_2 \geqslant 800r$,

$$h_{x_c}(M_c) \geqslant OPT_{x_c}(1-\epsilon) \geqslant (1-\epsilon)(8/9)\left(\|x_c-\mu\|_2 - 8r\right)^2 \geqslant \left(0.78\,\|x_c-\mu\|_2 + 8r\right)^2$$

for $\epsilon = 0.173$, solving step **5** from Algorithm 1.

Therefore, it only remains to show how to find a $\rho$ such that $\texttt{ALG}_\rho$ returns a pair $(M', y')$ (feasible for $(C_\rho)$) satisfying $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$. We do it first by assuming that we have access to an initial $\rho_0$ such that $\texttt{ALG}_{\rho_0}$ returns a feasible pair $(M', y')$ for $(C_\rho)$ (for $\rho = \rho_0$) so that $\mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$ and to a maximal number $T$ of iterations (we will also see later how to choose such $\rho_0$ and $T$). The following algorithm (which is a binary search) taking as input $(\bar{X}_1, \ldots, \bar{X}_K, x_c, \rho_0, u, T)$ returns a feasible pair $(M', y')$ for $(C_\rho)$ so that $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$ (when $T$ is large enough). This is simply due to the fact that $g$ is continuous, non increasing, $g(0) = 10/9 > 1$ and $g(\rho) \leqslant 2/8$ when $\rho \to +\infty$ and $\|x_c - \mu\|_2 > 800r$ (because of Lemma 6). For this to work, we need that for each iteration, $\texttt{ALG}_\rho$ returns a feasible pair $(M', y')$ for $(C_\rho)$ (for $\rho = \rho_0$) so that $\mathrm{Tr}(M') + \|y'\|_1 \leqslant (1 + 0.0001)g(\rho)$. We will suppose that it is the case for the rest of the proof. By union bound, this happens with probability at least $> 1 - T(1/10)^{u+10}/d$

**input :** $\bar{X}_1, \dots, \bar{X}_K, x_c, \rho_0, u, T$
**output:** A feasible pair $(M', y')$ for $(C_\rho)$ satisfying $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$

$\rho_m \leftarrow 0$, $\rho_M \leftarrow \rho_0$, $V \leftarrow \mathtt{ALG}_{\rho_0}(x_c, u, \eta = 0.0001)$ , $i \leftarrow 0$
**while** $V \notin [0.9981, 1]$ *and* $i < T$ **do**
  **if** $V < 0.9981$ **then**
    $\rho_M \leftarrow (\rho_M + \rho_m)/2$
  **end**
  **else**
    $\rho_m \leftarrow (\rho_M + \rho_m)/2$
  **end**
  $V \leftarrow objective(\mathtt{ALG}_{\frac{\rho_m+\rho_M}{2}}(x_c, u, \eta = 0.0001))$ , $i \leftarrow i + 1$
**end**
**Return** $\mathtt{ALG}_{\frac{\rho_m+\rho_M}{2}}(x_c, u, \eta = 0.0001)$

**Algorithm 2:** The `BinarySearch` algorithm to find a $\rho$ so that $\mathtt{ALG}_\rho$ returns a pair $(M', y')$ feasible for $(C_\rho)$ satisfying $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$.

If we can find a $\rho_0$ (such that $\mathtt{ALG}_{\rho_0}$ returns a feasible pair $(M', y')$ for $(C_\rho)$ so that $\mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$) and a large enough number of iterations $T$ in `BinarySerach`, Algorithm 2 returns a feasible pair $(M', y')$ for $(C_\rho)$ from which we can construct an approximating solution $M_c$ for $(E_{x_c})$ with objective value $h_{x_c}(M_c)$ larger than $(0.78\|x_c - \mu\|_2 + 8r)^2$ whenever $\|x_c - \mu\|_2 \geqslant 800r$. This is exactly what we expect in step **5** of Algorithm 1. Next, the last and final step that remains to be explained is to show how one can get such a $\rho_0$ and $T$ using only the block means $(\bar{X}_k)_{k=1}^K$ in $\tilde{\mathcal{O}}(Nd + uKd)$.

Let us consider $\hat{\mu}^{(0)}$ the coordinate-wise median(-of-means) and let us define $\delta = \mathrm{Med}(\|\bar{X}_k - \hat{\mu}^{(0)}\|_2 : k \in [K])$ – both quantities can be computed in time $\tilde{\mathcal{O}}(Kd)$. On the event $\mathcal{E}$, it follows from Corollary 1 (for $M = I_d/d$) and Proposition 3 that $\delta \leqslant 16\sqrt{d} \times r$. So if one takes $\rho_0 = d/\delta^2 \geqslant 1/[(16)^2 r^2]$, and if $\|x_c - \mu\|_2 > 800r$, Lemma 2 and Lemma 6 guarantee that $OPT_{x_c} \geqslant (8/9)\left(\|x_c - \mu\|_2 - 8r\right)^2 \geqslant (8/9)(792)^2 r^2$ and so

$$g(\rho_0) \leqslant \frac{1}{\rho\, OPT_{x_c}} + 0.171 \leqslant \frac{16^2}{(8/9)(792)^2} + 0.171 < 0.18$$

so $\mathtt{ALG}_{\rho_0} \leqslant (1+\eta)g(\rho) < 1.0001 \times 0.18 < 1$ (for the same choice of $\eta = 0.0001$).

Now we tackle the question of the number $T$ of iterations, which is crucial for the runtime. We know from Lemma 4 and Lemma 6 that the interval $I$ of all $\rho$'s such that $0.9981 \leqslant objective(\mathtt{ALG}_\rho) \leqslant 1$ is at least of size $0.001/OPT_{x_c}$ when $\|x_c - \mu\|_2 > 800r$. Indeed, since $g(\rho) \leqslant objective(\mathtt{ALG}_\rho) \leqslant (1+\eta)g(\rho)$, if $\rho$ is such that $0.9981 \leqslant g(\rho) \leqslant 1/(1+\eta)$ then $0.9981 \leqslant objective(\mathtt{ALG}_\rho) \leqslant 1$. Now, if we let $\rho_1 > 0$ and $0 < \alpha < 1$ be such that $g(\rho_1) = 0.9981$ and $g((1-\alpha)\rho_1) = 1/(1+\eta)$ the interval $I$ is at least of size $\alpha\rho_1$. Moreover, from Lemma 4 we have $1/(1+\eta) \leqslant g((1-\alpha)\rho_1) \leqslant g(\rho_1)/(1-\alpha)$ and so $0.9981 = g(\rho_1) \geqslant (1-\alpha)/(1+\eta)$, i.e. $\alpha \geqslant 1 - 0.9981(1+\eta) > 0.001$. Finally, since $g(\rho_1) \leqslant 1$, $g(1/OPT_{x_c}) = 1$ and $g$ is non-increasing, we conclude that $\rho_1 \geqslant 1/OPT_{x_c}$ and so the length of $I$ is at least $\alpha\rho_1 \geqslant 0.001/OPT_{x_c}$.

So, in the case where $\|x_c - \mu\|_2 > 800r$, $\log_2(\rho_0 \times OPT_{x_c}/0.001)$ iterations are enough to ensure that `BinarySearch` outputs $(M', y')$ (from $\mathtt{ALG}_\rho$ for a well-chosen $\rho$) feasible for $(C_\rho)$ and such that $0.9981 \leqslant \mathrm{Tr}(M') + \|y'\|_1 \leqslant 1$. Moreover, on the event $\mathcal{E}$ it is possible to show that for all iterations $x_c$ of the algorithm we have $\|x_c - \mu\|_2 < C\sqrt{d}r$ for a constant

$C \leqslant 800$ (we may take that as an induction hypothesis for the firsts iterates $x_c$, and the proof of Theorem 2.1 below in Section 5 shows that it will still holds for $x_{c+1}$). So if $\delta > r/d$ then $\rho_0 < d^3/r^2$, and since $OPT_{x_c} < (C^2 d + 8)r^2$ (this follows from Lemma 2), the binary search ends in time $T = \log_2(\tilde{C}d^4)$ with $\tilde{C} < 10^6$.

Thus, if the binary search has not ended in that time, we have either $\delta < r/d$ (which is a degenerate case) or $\|x_c - \mu\|_2 < 800r$ (or both). If $\|x_c - \mu\|_2 > 800r$ and $\delta < r/d$, then, taking $\rho_1 = 1/(d\delta)^2$, we have, by Lemma 6, $\texttt{ALG}_{\rho_1} < 1/2$. So, if we can not end our binary search in time $\log_2(\tilde{C}d^4)$, we compute $\texttt{ALG}_{1/(d\delta)^2}$: if this gives something smaller than 1, that means that $1/(d\delta)^2 > 1/OPT_{x_c} \Rightarrow \delta < \sqrt{(C^2 d + 8)}r/d < (C+1)r/\sqrt{d}$. We notice that on $\mathcal{E}$, $\left\|\hat{\mu}^{(0)} - \mu\right\|_2 < \delta + 8r$, so if $\texttt{ALG}_{1/(d\delta)^2} < 1$, then $\hat{\mu}^{(0)}$ is a good estimate for $\mu$. If on the contrary we have $\texttt{ALG}_{\rho_1} > 1$, it means that $\|x_c - \mu\|_2 < 800r$, so we stop the algorithm and return $x_c$.

Let us write now in pseudo-code the procedure we just described. This is an algorithm, named `SolveSDP`, running in $\tilde{\mathcal{O}}(Kud)$ which takes as inputs $\bar{X}_1, \ldots, \bar{X}_K$, $x_c$, $u$ and which outputs, on the event $\mathcal{E}$, with probability $> 1 - \log(\tilde{C}d^4)(1/10)^{u+10}/d$, for every $x_c \in \mathbb{R}^d$ such that $\|x_c - \mu\|_2 \geqslant 800r$ either a matrix $M_c$ such that

$$h_{x_c}(M_c) \geqslant (0.78\|x_c - \mu\|_2 + 8r)^2$$

or a subgaussian estimate of $\mu$. It therefore describes step **5** from Algorithm 1.

**input :** $\bar{X}_1, \ldots, \bar{X}_K$, $x_c$ and $u$
**output:** A feasible solution for ($E_{x_c}$)

Compute the coordinate wise MOM $\hat{\mu}^{(0)}$ and $\delta = \mathrm{Med}(\left\|\bar{X}_k - \hat{\mu}^{(0)}\right\|_2 : k \in [K])$
$T \leftarrow \log(\tilde{C}d^4)$, $\rho_0 \leftarrow d/\delta^2$
$(M', y') \leftarrow \mathrm{BinarySearch}(T, \rho_0, u, x_c)$
**if** $\mathrm{Tr}(M') + \|y\|_1 \in [0.9981, 1]$ **then**
  $M \leftarrow M'/\mathrm{Tr}(M')$
  **Return** (True, $M$)
**end**
**else**
  **if** $\texttt{ALG}_{1/(d\delta)^2}(x_c, u, \eta = 0.0001) < 1$ **then**
    **Return** (False, $\hat{\mu}^{(0)}$)
  **end**
  **else**
    **Return** (False, $x_c$)
  **end**
**end**

**Algorithm 3:** SolveSDP

REMARK 4. [Two advantages of block means] During the whole algorithm, we solve the program ($C_\rho$) up to a factor $(1+\eta)$ where $\eta$ is *fixed* (here we take it equal to $0.0001$). This differs crucially from the work of [12] where $\eta$ depends on the fraction of outliers, which decreases the performance of the algorithm in Lemma 5, the true running time being $\tilde{\mathcal{O}}(Kd/\mathrm{Poly}(\eta))$. This is a first advantage of using bucketed means instead of the data themselves: we work with a constant fraction of corrupted blocks (we took it equal to $1/10$). The second advantages is of stochastic nature, it is revealed by Proposition 1 or Lemma 1: most

of the bucketed means have a nice subgaussian behavior in all directions. Working with bucketed means has therefore two advantages: a stochastic one, which is to exhibit a subgaussian behavior for $9K/10$ blocks even under a $L_2$-moment assumption and a computational one, which is to make the proportion of corrupted blocks constant.

**5. The final algorithm and its computational cost: proof of Theorem 2.1.** We are now in a position to fully describe our robust subgaussian descent algorithm running in $\tilde{\mathcal{O}}(Nd + uKd)$. One may check that its construction is fully data-dependent, in particular, we do not need to know the value of $r$ or the proportion of outliers.

**input :** $X_1, \dots, X_N$, $K \in [N]$ and $u \in \mathbb{N}^*$
**output:** A robust subgaussian estimator of $\mu$

Construct an equipartition $B_1 \sqcup \ldots \sqcup B_K = \{1, \dots, N\}$
Construct the $K$ empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
Compute $\hat{\mu}^{(0)}$ the coordinate-wise median
$x_c \leftarrow \hat{\mu}^{(0)}$, Bool $\leftarrow$ True, $T \leftarrow 0$
**while** *Bool and* $T < \log(8\sqrt{d})/\log(1/0.81)$ **do**
    Bool, $A \leftarrow$SolveSDP($\bar{X}_1, \dots, \bar{X}_K, x_c, u$)
    **if** *Bool* **then**
        $M_c \leftarrow A$
        Compute $v_1$ a top eigenvector of $M_c$
        Compute a step size $\theta_c = -\operatorname{Med}\left(\left\langle \bar{X}_k - x_c, v_1 \right\rangle : k \in [K]\right)$
        Update $x_c \leftarrow x_c - \theta_c v_1$
        $T \leftarrow T + 1$
    **end**
    **else**
        $x_c \leftarrow A$
    **end**
**end**
**Return** $x_c$

**Algorithm 4:** Final Algorithm: covSDPofMeans

**Proof of Theorem 2.1.** From Theorem 4.1, we know that on $\mathcal{E}$, when, $\|x_c - \mu\|_2 > 800r$, we get, with probability $> 1 - (1/10)^{u+5}/\sqrt{d}$, an $M_c$ so that $h_{x_c}(M_c) \geqslant (0.8\|x_c - \mu\|_2 + 8r)^2$ (or directly a subgaussian estimate, in which case our work is done). Proposition 4, states that in that case $\|x_{c+1} - \mu\|_2^2 \leqslant 0.8\|x_c - \mu\|_2^2 + 64r^2 \leqslant 0.81\|x_c - \mu\|_2^2$. So we have a geometric decays and Proposition 3 guarantees that our starting point is at most $8\sqrt{d}r$ far away from the mean so that in at most $\log(8\sqrt{d})/\log(1/0.81))$ steps the algorithm outputs its current point which is $r$-close to $\mu$, with probability $> 1 - (1/10)^{u+5}\log(8\sqrt{d})/(\log(1/0.81))\sqrt{d}) > 1 - (1/10)^u$ (by union bound).

The last thing to do is to control what happens when $\|x_c - \mu\|_2 < 800r$. Then, we have no guarantees on $v_1$, but using the similar argument as in the proof of Proposition 4 we know that

(11)
$$|\theta_c - \langle x_c - \mu, v_1\rangle| = |\operatorname{Med}\left(\left\langle \mu - \bar{X}_k, v_1\right\rangle : k \in [K]\right)| \leqslant \operatorname{Med}\left(|\left\langle \mu - \bar{X}_k, v_1\right\rangle| : k \in [K]\right) \leqslant 8r$$

and (for some $v_1^\perp$ a normalized orthogonal vector to $v_1$)

$$\|x_{c+1}-\mu\|_2^2 = \|x_c-\mu-\theta_c v_1\|_2^2 = \left\|(\langle x_c-\mu, v_1\rangle - \theta_c)v_1 + \langle x_c-\mu, v_1^\perp\rangle v_1^\perp\right\|_2^2$$
$$= (\langle x_c-\mu, v_1\rangle - \theta_c)^2 + \langle x_c-\mu, v_1^\perp\rangle^2 \leqslant (8r)^2 + \|x_c-\mu\|_2^2 .$$

Hence, $\|x_{c+1}-\mu\|_2 \leqslant (8r) + \|x_c-\mu\|_2$. Therefore, in the worst case scenario where $\|x_c-\mu\|_2 > 800r$ at the last iteration, the algorithm outputs the next iteration $\hat{\mu}_K = x_{c+1}$ so that $\|\hat{\mu}_K-\mu\|_2 \leqslant 808r$.

We end this proof with the computation of the running time of Algorithm 4. We detail the computation cost for each line of Algorithm 4: line **1** cost $N$, line **2** costs $Nd$, line **3** costs $\mathcal{O}(dK\log(K))$. The while loop in line **5** is running at least $\log d$ times (up to constant) so that the computational cost of all remaining lines of Algorithm 4 are at worst to be multiplied by $\log d$. Line **6** costs $\log(\tilde{C}d^4)$ steps, each of cost $\tilde{\mathcal{O}}(Kud)$ (that comes from Lemma 5). Line **9** can be computed in $\tilde{\mathcal{O}}(Nd)$ thanks to Lemma 5. Finally, line **10** costs $\mathcal{O}(Kd)$. Other lines take time at most $d$. We thus recover the running time announced in Theorem 2.1. ∎

**6. Adaptive choice of $K$ and results in expectation.** Given a number of blocks $K \in \{1,\dots,N\}$, a parameter $u \geqslant 1$ (so that the covering SDPs from [53] (used in Lemma 5) run in $u + 3\log d + 10$ times) and the (adversarially corrupted and heavy-tailed) dataset $\{X_1,\dots,X_N\}$, Algorithm 4 returns a vector $\hat{\mu}_K$ in $\mathbb{R}^d$ and Theorem 2.1 ensures that $\hat{\mu}_K$ estimates the true mean $\mu$ at the subgaussian rate (1) with large probability as long as $K \geqslant 300|\mathcal{O}|$. As a consequence, we have certified statistical guarantees for $\hat{\mu}_K$ only when some a priori knowledge on the number $|\mathcal{O}|$ of outliers is provided (such as "the corruption of this database is less than $5\%$" ) or if we choose $K$ like $N$- but, in this later case the rate (1) may be too pessimistic. The aim of this section is to overcome this issue by constructing a procedure which can automatically adapt to the number of outliers. The resulting procedure (denoted later by $\hat{\mu}^{(\hat{J})}$) satisfies the same statistical bounds as $\hat{\mu}_K$ for all $K \geqslant 300|\mathcal{O}|$ without knowing $|\mathcal{O}|$ (up to constants). We also show that it satisfies results in expectation.

The adaptation method we use is based on the Lepski method [43, 44] which is another tool used by the "statistical community" working on robustness issues since [47, 8]. The price we pay for this adaptation is the a priori knowledge of the rate (1) for all $K$ which means that we know in advance $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ – this is for instance the case when it is known that $\Sigma$ is the identity matrix $I_d$. Of course, one can design robust estimators for $\mathrm{Tr}(\Sigma)$ (see [17]) and $\|\Sigma\|_{op}$ but this requires stronger assumptions (more than four moments) that we want to avoid at this stage.

Lepski's method proceeds as follows. We set for all $K \in \{1,\dots,N\}$ and all $j \in \{0,1,\dots,\log_2 N\}$

$$r_K^* = 808\left(1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200\,\|\Sigma\|_{op}K}{N}}\right) \text{ and } r^{(j)} = r^*_{\lceil N/2^j\rceil}$$

the rate of convergence from Theorem 2.1. For a given parameter $u_j \in \mathbb{N}^*$, we construct from Algorithm 4

$$\hat{\mu}^{(j)} \leftarrow covSDPofMeans(X_1,\dots,X_N, K=\lceil N/2^j\rceil, u=u_j). \tag{12}$$

Classical Lepski's method considers the largest $J$ such that $\bigcap_{j=0}^J B_2^d(\hat{\mu}^{(j)}, r^{(j)})$ is none-empty and then take any point $\hat{\mu}$ in this none-empty intersection. Standard analysis of Lepski's method shows that $\hat{\mu}$ estimates $\mu$ at the rate $r_K^*$ (up to an absolute constant) simultaneously for all $K \in \{300|\mathcal{O}|,\dots,N\}$ without knowing $|\mathcal{O}|$. Given that checking that the

intersection of several $\ell_2^d$-balls may not be straightforward, we use a slightly modified version of Lepski's method as described in the following algorithm.

**input :** $X_1, \dots, X_N$ and $\{u_j : j = 0, 1, 2, \dots, \log_2 N\} \subset \mathbb{N}^*$
**output:** A robust subgaussian estimator of $\mu$ with adaptive choice of $K$
**init :** $J = 0$ and $\hat{\mu}^{(0)} = covSDPofMeans(X_1, \dots, X_N, K = N, u = u_0)$
**while** $\left\|\hat{\mu}^{(J)} - \hat{\mu}^{(j)}\right\|_2 \leqslant r^{(J)} + r^{(j)}, j = J-1, J-2, \dots, 0$ **do**
  $J \leftarrow J + 1$
  $\hat{\mu}^{(J)} \leftarrow covSDPofMeans(X_1, \dots, X_N, K = \lceil N/2^J \rceil, u = u_J)$
**end**
**Return** $\hat{\mu}^{(J)}$

**Algorithm 5:** Adaptive choice of $K$ in covSDPofMeans

Unlike for the traditional Lepski's method we check that $\hat{\mu}^{(J)}$ is in $\bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)})$ instead of checking that $\bigcap_{j=0}^{J} B_2^d(\hat{\mu}^{(j)}, r^{(j)})$ is none-empty – this simplifies the adaptation step. It is also possible to speed up the whole procedure by constructing iteratively the bucketed means. Indeed, given that we consider a dyadic grid for $K$, i.e. $K \in \{N, \lceil N/2 \rceil, \lceil N/4 \rceil, \dots\}$, for all $j \in \mathbb{N}$, we can construct the block means $\{\bar{X}_k^{(j+1)}, k = 1, \dots, \lceil N/2^{j+1} \rceil\}$ at step $K = \lceil N/2^{j+1} \rceil$ using the block means from the previous step $K = \lceil N/2^j \rceil$ by simply averaging two successive block means: $\bar{X}_k^{(j+1)} \leftarrow (\bar{X}_{2k}^{(j)} + \bar{X}_{2k+1}^{(j)})/2$.

Let us now turn to the statistical analysis of the output $\hat{\mu}^{(\hat{J})}$ from Algorithm 5 where

$$\hat{J} = \max\left( J \in \{0, 1, \dots, \log_2 N\} : \hat{\mu}^{(J)} \in \bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)}) \right).$$

THEOREM 6.1. *Let $\{u_j : j = 0, 1, 2, \dots, \log_2 N\} \subset \mathbb{N}^*$ be the family of parameters used to construct the family of estimators $\{\hat{\mu}^{(j)}, j = 0, 1, \dots\}$ in Algorithm 5 (see also (12)). For all $K \in \{600|\mathcal{O}|, \dots, N\}$, with probability at least*

$$1 - 2\exp(-K/360000) - \sum_{j=0}^{\log_2(N/(K-1))} (1/10)^{u_j} \tag{13}$$

*the output $\hat{\mu}^{(\hat{J})}$ of Algorithm 5 is such that $\left\|\hat{\mu}^{(\hat{J})} - \mu\right\|_2 \leqslant 3r_K^*$.*

**Proof.** For all $j \in \{0, 1, \dots, \log_2 N\}$ denote by $\mathcal{E}_j$ the event onto which Theorem 2.1 is valid for $K = \lceil N/2^j \rceil$ and for $u = u_j$: that is on $\mathcal{E}_j$, if $\lceil N/2^j \rceil \geqslant 300|\mathcal{O}|$, $\left\|\hat{\mu}^{(j)} - \mu\right\|_2 \leqslant r^{(j)}$ and $\mathbb{P}[\mathcal{E}_j] \geqslant 1 - \exp(-\lceil N/2^j \rceil/180000) - (1/10)^{u_j}$. Let $K \in \{600|\mathcal{O}|, \dots, N\}$ and $J \in \{0, 1, \dots, \log_2 N\}$ be such that $\lceil N/2^J \rceil \leqslant K < \lceil N/2^{J-1} \rceil$. On the event $\bigcap_{j=0}^{J} \mathcal{E}_j$, we have $\left\|\hat{\mu}^{(j)} - \mu\right\|_2 \leqslant r^{(j)}$ for all $j = 0, 1, \dots, J$, in particular, for all $j = 0, 1, \dots, J-1$, $\left\|\hat{\mu}^{(J)} - \hat{\mu}^{(j)}\right\|_2 \leqslant r^{(J)} + r^{(j)}$ and so $\hat{\mu}^{(J)} \in \bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)})$. As a consequence $\hat{J} \geqslant J$ therefore $\left\|\hat{\mu}^{(\hat{J})} - \hat{\mu}^{(J)}\right\|_2 \leqslant r^{(\hat{J})} + r^{(J)} \leqslant 2r^{(J)} \leqslant 2r_K^*$. Finally, we have

$$\mathbb{P}\Big[\bigcap_{j=0}^{J} \mathcal{E}_j\Big] \geqslant 1 - \sum_{j=0}^{J} \exp(-\lceil N/2^j \rceil/180000) - (1/10)^{u_j}$$

$$\geqslant 1 - 2\exp(-K/360000) - \sum_{j=0}^{\log_2(N/(K-1))} (1/10)^{u_j}.$$

■

We can see in Algorithm 5 that $\hat{\mu}^{(\hat{J})}$ does not use any information on the number of outliers $|\mathcal{O}|$ for its construction but it can still estimate $\mu$ at the optimal rate $r_K^*$ for all deviation parameters $K$ in $\{600|\mathcal{O}|, \dots, N\}$. The maximum total running time of Algorithm 5 is achieved when $\hat{J} = \log_2 N$; in that case, it is at most $\tilde{\mathcal{O}}(Nd + \sum_{j=0}^{\log_2 N} \lceil N/2^j \rceil u_j d)$. In particular, if one chooses $u_j = 2^j$ for all $j = 0, 1, \dots, \log_2 N$ then the total running time for the construction of $\hat{\mu}^{(\hat{J})}$ is nearly-linear $\tilde{\mathcal{O}}(Nd)$. For this choice of $u_j$, the probability deviation in (13) is constant and so one should choose the smallest possible $K$ allowed in Theorem 6.1, that is $K = 600|\mathcal{O}|$. Let us write formally this result.

COROLLARY 2. *If one takes $u_j = 2^j$ for all $j = 0, 1, \dots, \log_2 N$ in Algorithm 5 then, in nearly-linear time $\tilde{\mathcal{O}}(Nd)$, with probability at least $1 - 2\exp(-600|\mathcal{O}|/360000) - 1/11$, the output $\hat{\mu}^{(\hat{J})}$ from Algorithm 5 satisfies*

$$\left\| \hat{\mu}^{(\hat{J})} - \mu \right\|_2 \leqslant 2r^*_{600|\mathcal{O}|} = 1616 \left( 1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + 850\sqrt{\frac{\|\Sigma\|_{op}\,|\mathcal{O}|}{N}} \right).$$

In particular, considering the setup from Theorem 1.1, if $|\mathcal{O}| = \epsilon N$ for some $\epsilon \leqslant 1/600$ then the rate achieved by $\hat{\mu}^{(\hat{J})}$ in Corollary 2 is of the order of

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\|\Sigma\|_{op}\,\epsilon} \tag{14}$$

which is like $\sqrt{\|\Sigma\|_{op}\,\epsilon}$ when $N \geqslant (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$. As a consequence, the result from Corollary 2 improves the one from Theorem 1.1 by removing an extra $\log d$ factor in the sample complexity in the case considered in Theorem 1.1 that is when $\Sigma \preceq \sigma^2 I_d$. Moreover, Corollary 2 also shows that the sample complexity depends on the *effective rank* $\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op}$ of $\Sigma$. This ratio can be much smaller than $d$ if the spectrum of $\Sigma$ decays sufficiently fast. Finally, Corollary 2 also covers the case where the sample size $N$ is less than the sample complexity – that is when $N \leqslant (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$. In that case, the estimation rate is given by $\sqrt{\mathrm{Tr}(\Sigma)/N}$ which is the complexity coming from the estimation of $\mu$ in the none corrupted case. As a consequence, Corollary 2 exhibits a phase transition happening at $N \sim (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$ above which corruption is the main source of estimation mistakes and below which corruption does not play any role.

Corollary 2 covers the case where $\hat{\mu}^{(\hat{J})}$ is computed in nearly-linear time and with statistical guarantees happening with constant probability. In the following final result, we show that $\hat{\mu}^{(\hat{J})}$ can estimate $\mu$ at the optimal rate $r_K^*$ for all $K \geqslant 600|\mathcal{O}|$ with a subgaussian deviation $1 - 2\exp(-K/360000)$ if we perform more iterations $u_j$ of the covering SDP from Lemma 5. The price we pay for this subgaussian behavior of $\hat{\mu}^{(\hat{J})}$ is on the total running time which goes from nearly-linear time $\tilde{\mathcal{O}}(Nd)$ to $\tilde{\mathcal{O}}(N^2 d)$ by taking $u_j = \lceil N/2^j \rceil$ for $j = 0, 1, \dots, \log_2 N$ ($u_j = N$ would do as well). We write formally this statement in the next corollary which follows directly from Theorem 6.1.

COROLLARY 3. *If one takes* $u_j = \lceil N/2^j \rceil$ *for all* $j = 0, 1, \dots, \log_2 N$ *in Algorithm 5 then, in time* $\tilde{\mathcal{O}}(N^2 d)$*, for all* $K \geqslant 600|\mathcal{O}|$*, with probability at least* $1 - 4\exp(-K/360000)$*, the output* $\hat{\mu}^{(\hat{J})}$ *from Algorithm 5 satisfies*

$$\left\| \hat{\mu}^{(\hat{J})} - \mu \right\|_2 \leqslant 2r_K^* = 1616 \left( 1200 \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200 \left\| \Sigma \right\|_{op} K}{N}} \right).$$

As a consequence $\hat{\mu}^{(\hat{J})}$ is a subgaussian estimator of $\mu$ for all range of $K$ from $600|\mathcal{O}|$ to $N$ which can handle up to $|\mathcal{O}|$ outliers in the database (even when $|\mathcal{O}| \sim N$) and that can be constructed in time $\tilde{\mathcal{O}}(N^2 d)$. It does not require any knowledge on $|\mathcal{O}|$ for its construction.

Let us now show that the algorithm $\hat{\mu}^{(\hat{J})}$ constructed in Corollary 3 also satisfies estimation results in expectation. So far all the statistical properties have been given with large probability; for $\hat{\mu}^{(\hat{J})}$ it is also possible to obtain a result in expectation.

The benchmark result we use here is the rate achieved by the empirical mean in a non-corrupted setup but unlike the result in deviation we don't need i.i.d. Gaussian variables since $\mathbb{E} \left\| \overline{\tilde{X}}_n - \mu \right\|_2 \leqslant \sqrt{\mathrm{Tr}(\Sigma)/N}$ where $\overline{\tilde{X}}_n = n^{-1} \sum_i \tilde{X}_i$ and $\tilde{X}_1, \dots, \tilde{X}_N$ are the non-corrupted data points from Assumption 1. Hence, $\sqrt{\mathrm{Tr}(\Sigma)/N}$ is the rate we aim to achieve but we also may expect a price to pay for the adversarial corruption, in particular, when $\epsilon = |\mathcal{O}|/N$ is above the phase transition exhibited in (14), that is for $\epsilon \geqslant (\mathrm{Tr}(\Sigma)/\left\| \Sigma \right\|_{op})/N$.

THEOREM 6.2. *Under Assumption 1, and if* $N \geqslant 600|\mathcal{O}|$*, the following holds. If one takes* $u_j = \lceil N/2^j \rceil$ *for all* $j = 0, 1, \dots, \log_2 N$ *in Algorithm 5 then, in time* $\tilde{\mathcal{O}}(N^2 d)$*, Algorithm 5 outputs* $\hat{\mu}^{(\hat{J})}$ *satisfying*

$$\mathbb{E} \left\| \hat{\mu}^{(\hat{J})} - \mu \right\|_2 \leqslant (3 + 16c_0^2) r^*_{600|\mathcal{O}|} \leqslant (3 + 16c_0^2) 808 \times 1200 \left( \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\left\| \Sigma \right\|_{op} |\mathcal{O}|}{2N}} \right)$$

*as long as and* $N \geqslant 4c_0 \log(c_0 d + c_0)$ *where* $c_0 = 360000$.

**Proof.** We denote $\tilde{\mu} = \hat{\mu}^{(\hat{J})}$ and $c_0 = 360000$. We know from Corollary 3 that for all $600|\mathcal{O}| \leqslant K \leqslant N$, with probability at least $1 - 4\exp(-K/c_0)$, $\|\tilde{\mu} - \mu\|_2 \leqslant 2r_K^*$. So we know how to control the estimation property of $\tilde{\mu}$ up to an event of probability measure at most $4\exp(-N/c_0)$. On that event, we only need a crude upper bound on $\|\tilde{\mu} - \mu\|_2$ to get the result. This is what we do now.

We know that by construction that $\tilde{\mu} \in B_2^d(\hat{\mu}^{(N)}, 2r_N^*)$. Moreover, $\hat{\mu}^{(N)}$ starts from $\hat{\mu}_0^{(N)}$, the coordinate wise median of the data $X_i$ (because $K = N$ blocks here) and makes at most $T = \log(8\sqrt{d})/\log(1/0.81)$ descent iterations like $x_{c+1} = x_c - \theta_c v_1$ where $v_1 \in \mathcal{S}_2^{d-1}$ and $\theta_c = -\mathrm{Med}\left( \langle X_i - x_c, v_1 \rangle : i \in [N] \right)$. In particular, one has at every iteration

$$\|x_{c+1} - \mu\|_2 \leqslant 2\|x_c - \mu\|_2 + \mathrm{Med}(\|X_i - \mu\|_2 : k \in [K]).$$

Hence, $\hat{\mu}^{(N)}$ satisfies

$$\begin{aligned} \left\| \hat{\mu}^{(N)} - \mu \right\|_2 &\leqslant 2^{T+1} \left( \left\| \hat{\mu}_0^{(j)} - \mu \right\|_2 + \mathrm{Med}(\|X_i - \mu\|_2 : i \in [N]) \right) \\ &\leqslant 16d \left( \left\| \hat{\mu}_0^{(N)} - \mu \right\|_\infty + \mathrm{Med}(\|X_i - \mu\|_\infty : i \in [N]) \right). \end{aligned} \tag{15}$$

In the adversarial contamination model from Assumption 1, as we assumed that $N \geqslant 600|\mathcal{O}|$, there are at least $N - |\mathcal{O}| \geqslant (599/600)N$ indices $i$ such that $X_i = \tilde{X}_i$, hence for at least $(599/600)N$ $i$'s we have, for all $p \in [d]$,

$$|X_{i,p} - \mu_p| \leqslant \max_{i\in[N]} |\tilde{X}_{i,p} - \mu_p| \text{ and } \|X_i - \mu\|_\infty \leqslant \max_{i\in[N]} \left\|\tilde{X}_i - \mu\right\|_\infty$$

where $X_{i,p}$ (resp. $\mu_p$) denotes the $p$-th coordinate of $X_i$ (resp. $\mu$). Hence, in (15), we get

$$\left\|\hat{\mu}^{(N)} - \mu\right\|_2 \leqslant 32d \max_{i\in[N]} \max_{p\in[d]} |X_{i,p} - \mu_p|.$$

Let us now turn to the stochastic argument to upper bound the right-hand side in the last inequality.

$$\mathbb{E}(\max_{i\in[N]} \max_{p\in[d]} |X_{i,p} - \mu_p|^2) \leqslant \mathbb{E}(\max_{i\in[N]} ||X_i - \mu||_2^2) \leqslant N \operatorname{Tr}(\Sigma).$$

Hence,

$$\mathbb{E}(\|\tilde{\mu} - \mu\|_2^2) \leqslant 2048 d^2 N \operatorname{Tr}(\Sigma) + 8(r_N^*)^2. \tag{16}$$

We are now in a position to obtain an estimation result in expectation for $\tilde{\mu}$. We denote $K_{\mathcal{O}} = 600|\mathcal{O}|$:

$$\begin{aligned}
\mathbb{E}\,\|\tilde{\mu} - \mu\|_2 &= \sum_{k=K_{\mathcal{O}}}^{N-1} \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(2r_k^* \leqslant \|\tilde{\mu} - \mu\|_2 \leqslant 2r_{k+1}^*)\right] \\
&\quad + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \leqslant 2r_{K_{\mathcal{O}}}^*)\right] + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geqslant 2r_N^*)\right] \\
&\leqslant 2r_{K_{\mathcal{O}}}^* + \sum_{k=K_{\mathcal{O}}}^{N-1} 2r_{k+1}^* \times 4\exp(-k/c_0) + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geqslant 2r_N^*)\right] \\
&\leqslant 2r_{K_{\mathcal{O}}}^* + 16c_0^2 r_{K_{\mathcal{O}}}^* \exp(-K_{\mathcal{O}}/c_0) + 25c_0 d\sqrt{N \operatorname{Tr}(\Sigma)} \exp(-N/(2c_0))
\end{aligned}$$

where, in the last inequality, we used that

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geqslant 2r_N^*)\right] \leqslant \left(\mathbb{E}\left[\|\tilde{\mu} - \mu\|_2^2\right]\right)^{1/2} \left(\mathbb{P}\left[\|\tilde{\mu} - \mu\|_2 \geqslant 2r_N^*\right]\right)^{1/2} \\
&\leqslant (64d\sqrt{N \operatorname{Tr}(\Sigma)} + 3r_N^*) \times 2\exp(-N/(2c_0)) \leqslant 25c_0 d\sqrt{N \operatorname{Tr}(\Sigma)} \exp(-N/(2c_0))
\end{aligned}$$

from (16). When $N \geqslant 4c_0 \log(c_0 d + c_0)$, then $N \geqslant 2c_0 \log[c_0 dN]$, so $\mathbb{E}\,\|\tilde{\mu} - \mu\|_2 \leqslant (3 + 16c_0^2) r_{K_{\mathcal{O}}}^*$. ∎

We therefore recover the same rate of convergence in expectation in Theorem 6.2 as the one in deviation in Corollary 3 for the adaptive estimator $\hat{\mu}^{(\hat{J})}$, it is also the rate achieved by the non adaptive estimator $\hat{\mu}_K$ for the minimal value of $K = 600|\mathcal{O}|$. In particular, the same phase transition phenomena occurs in expectation as in the discussion following Equation (14).

*Acknowledgements.* We would like to thank Yeshwanth Cherapanamjeri, Ilias Diakonikolas, Yihe Dong, Nicolas Flammarion, Sam Hopkins and Jerry Li for helpful comments on our work.

Guillaume Lecué is supported by a grant overseen by the French National Research Agency (ANR) as part of the "Investments d'Avenir" Program (LabEx ECODEC; ANR-11-LABX-0047), by the Médiamétrie chair on "Statistical models and analysis of high-dimensional data" and by the French ANR PRC grant ADDS (ANR-19-CE48-0005).

## REFERENCES


[1] ALLEN-ZHU, Z., GELASHVILI, R. and RAZENSHTEYN, I. (2014). The restricted isometry property for the general p-norms. *arXiv:1407.2178*.

[2] ALLEN-ZHU, Z., LEE, Y. T. and ORECCHIA, L. (2015). Using Optimization to Obtain a Width-Independent, Parallel, Simpler, and Faster Positive SDP Solver.

[3] ALON, N., MATIAS, Y. and SZEGEDY, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** 137–147. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996). MR1688610

[4] BIRGÉ, L. (1984). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.* **20** 201–223. MR762855

[5] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR3185193

[6] BUBECK, S., CESA-BIANCHI, N. and LUGOSI, G. (2013). Bandits with heavy tail. *IEEE Trans. Inform. Theory* **59** 7711–7717. MR3124669

[7] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'I.H.P. Probabilités et statistiques* **48** 1148-1185. MR3052407

[8] CATONI, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

[9] CATONI, O. and GIULINI, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression Technical Report, CNRS and LSPM.

[10] CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006

[11] CHENG, Y., DIAKONIKOLAS, I. and GE, R. (2018). High-Dimensional Robust Mean Estimation in Nearly-Linear Time. *arXiv preprint arXiv:1811.09380*.

[12] CHENG, Y., DIAKONIKOLAS, I. and GE, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2755–2771. SIAM, Philadelphia, PA. MR3909640

[13] CHERAPANAMJERI, Y., FLAMMARION, N. and BARTLETT, P. L. (2019). Fast Mean Estimation with Sub-Gaussian Rates.

[14] DALALYAN, A. S. and MINASYAN, A. (2020). All-In-One Robust Estimator of the Gaussian Mean. *arXiv preprint arXiv:2002.01432*.

[15] DE LA PEÑA, V. H. and GINÉ, E. (1999). *Decoupling*. *Probability and its Applications (New York)*. Springer-Verlag, New York From dependence to independence, Randomly stopped processes. $U$-statistics and processes. Martingales and beyond. MR1666908

[16] DEPERSIN, J. (2020). A spectral algorithm for robust regression with subgaussian rates Technical Report, CREST - ENSAE.

[17] DEPERSIN, J. and LECUÉ, G. (2019). Fast algorithms for robust estimation of a mean vector.

[18] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558

[19] DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., MOITRA, A. and STEWART, A. (2019). Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM J. Comput.* **48** 742–864. MR3945261

[20] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016* 655–664. IEEE Computer Soc., Los Alamitos, CA. MR3631028

[21] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on* 655–664. IEEE.

[22] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2017). Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*.

[23] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2018). Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 2683–2702. Society for Industrial and Applied Mathematics.

[24] DIAKONIKOLAS, I. and KANE, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

[25] DIAKONIKOLAS, I., KANE, D. M. and STEWART, A. (2018). List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* 1047–1060. ACM, New York. MR3826316

[26] DIAKONIKOLAS, I., KONG, W. and STEWART, A. (2019). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2745–2754. SIAM, Philadelphia, PA. MR3909639

[27] DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313

[28] GOEMANS, M. X. and WILLIAMSON, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)* **42** 1115–1145.

[29] GROTHENDIECK, A. (1953). Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo* **8** 1–79. MR94682

[30] HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896. MR0301858

[31] HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **27** 87–104. MR0359096

[32] HOPKINS, S. B. (2018). Sub-Gaussian Mean Estimation in Polynomial Time. *arXiv preprint arXiv:1809.07425*.

[33] HOPKINS, S. B., LI, J. and ZHANG, F. (2020). Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization. *arXiv preprint arXiv:2007.15839*.

[34] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[35] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust statistics*, second ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ. MR2488795

[36] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188. MR855970

[37] KOLTCHINSKII, V., PANCHENKO, D., LOZANO, F. et al. (2003). Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. *The Annals of Applied Probability* **13** 213–252.

[38] LECUÉ, G. and LERASLE, M. (2019). Learning from MOM's principles: Le Cam's approach. *Stochastic Processes and their applications* **129** 4385–4410.

[39] LECUÉ, G. and LERASLE, M. (2020). Robust machine learning by median-of-means: theory and practice. *Annals of Statistics* **48** 906–931.

[40] LEDOUX, M. (2001). *The concentration of measure phenomenon*. *Mathematical Surveys and Monographs* **89**. American Mathematical Society, Providence, RI. MR1849347 (2003k:28019)

[41] LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach spaces*. *Classics in Mathematics*. Springer-Verlag, Berlin Isoperimetry and processes, Reprint of the 1991 edition. MR2814399

[42] LEI, Z., LUH, K., VENKAT, P. and ZHANG, F. (2020). A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory* 2598–2612.

[43] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.* **35** 459–470. MR1091202

[44] LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* **36** 645–659. MR1147167

[45] LERASLE, M. and OLIVEIRA, R. (2011). Robust empirical mean Estimators Technical Report, IMPA and CNRS.

[46] LERASLE, M., SZABO, Z., MATHIEU, T. and LECUÉ, G. (2017). MONK – Outliers-Robust Mean Embedding Estimation by Median-of-Means Technical Report, CNRS, University of Paris 11, Ecole Polytechnique and CREST.

[47] LUGOSI, G., MENDELSON, S. et al. (2019). Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics* **47** 783–794.

[48] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335.

[49] MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758

[50] MINSKER, S. and STRAWN, N. (2017). Distributed statistical estimation and rates of convergence in normal approximation. Technical Report, arXiv: 1704.02658.

[51] NEMIROVSKY, A. S. and YUDIN, D. B. A. (1983). *Problem complexity and method efficiency in optimization*. *A Wiley-Interscience Publication*. John Wiley & Sons, Inc., New York Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. MR702836

[52] NESTEROV, Y. (1997). Semidefinite relaxation and non-convex quadratic optimization. *Optimization Methods and Software* **12** 1–20.
[53] PENG, R., TANGWONGSAN, K. and ZHANG, P. (2012). Faster and Simpler Width-Independent Parallel Algorithms for Positive Semidefinite Programming.
[54] PISIER, G. (2012). Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society* **49** 237–323.
[55] SMALL, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique* 263–277.
[56] TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to probability and statistics* 448–485. Stanford Univ. Press, Stanford, Calif. MR0120720
[57] TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67. MR0133937