ALL-IN-ONE ROBUST ESTIMATOR OF THE GAUSSIAN MEAN

BY ARNAK S. DALALYAN[1] AND ARSHAK MINASYAN[2]

[1]*ENSAE-CREST, arnak.dalalyan@ensae.fr*

[2]*Yerevan State University, YerevaNN, minasyan@yerevann.com*

The goal of this paper is to show that a single robust estimator of the mean of a multivariate Gaussian distribution can enjoy five desirable properties. First, it is computationally tractable in the sense that it can be computed in a time which is at most polynomial in dimension, sample size and the logarithm of the inverse of the contamination rate. Second, it is equivariant by translations, uniform scaling and orthogonal transformations. Third, it has a high breakdown point equal to 0.5, and a nearly-minimax-rate-breakdown point approximately equal to 0.28. Fourth, it is minimax rate optimal, up to a logarithmic factor, when data consists of independent observations corrupted by adversarially chosen outliers. Fifth, it is asymptotically efficient when the rate of contamination tends to zero. The estimator is obtained by an iterative reweighting approach. Each sample point is assigned a weight that is iteratively updated by solving a convex optimization problem. We also establish a dimension-free non-asymptotic risk bound for the expected error of the proposed estimator. It is the first result of this kind in the literature and involves only the effective rank of the covariance matrix. Finally, we show that the obtained results can be extended to sub-Gaussian distributions, as well as to the cases of unknown rate of contamination or unknown covariance matrix.

1. Introduction. Robust estimation is one of the most fundamental problems in statistics. Its goal is to design efficient methods capable of processing data sets contaminated by outliers, so that these outliers have little influence on the final result. The notion of an outlier is hard to define for a single data point. It is also hard, inefficient and often impossible to clean data by removing the outliers. Instead, one can build methods that take as input the contaminated data set and provide as output an estimate which is not very sensitive to the contamination. Recent advances in data acquisition and computational power provoked a revival of interest in robust estimation and learning, with a focus on finite sample results and computationally tractable procedures. This was in contrast to the more traditional studies analyzing asymptotic properties of such statistical methods.

This paper builds on recent advances made in robust estimation and suggests a method that has attractive properties both from asymptotic and finite-sample points of view. Furthermore, it is computationally tractable and its statistical complexity depends optimally on the dimension. As a matter of fact, we even show that what really matters is the intrinsic dimension, defined in the Gaussian model as the effective rank of the covariance matrix.

Note that in the framework of robust estimation, the high-dimensional setting is qualitatively different from the one dimensional setting. This qualitative difference can be shown at two levels. First, from a computational point of view, the running time of several robust methods scales poorly with dimension. Second, from a statistical point of view, while a simple "remove then average" strategy might be successful in low-dimensional settings, it can easily be seen to fail in the high dimensional case. Indeed, assume that for some $\varepsilon \in (0, 1/2)$, $p \in \mathbb{N}$, and $n \in \mathbb{N}$, the data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ consist of $n(1-\varepsilon)$ points (inliers) drawn from a

AMS 2000 subject classifications: Primary 62H12, ; secondary 62F35.

Keywords and phrases: Gaussian mean, robust estimation, breakdown point, minimax rate, computational tractability.

p-dimensional Gaussian distribution $\mathcal{N}_p(0, \mathbf{I}_p)$ (where $\mathbf{I}_p$ is the $p \times p$ identity matrix) and εn points (outliers) equal to a given vector $\boldsymbol{u}$. Consider an idealized setting in which, for a given threshold $r > 0$, an oracle tells the user whether or not $\boldsymbol{X}_i$ is within a distance r of the true mean 0. A simple strategy for robust mean estimation consists of removing all the points of Euclidean norm larger than $2\sqrt{p}$ and averaging all the remaining points. If the norm of $\boldsymbol{u}$ is equal to $\sqrt{p}$, one can check that the distance between this estimator and the true mean $\boldsymbol{\mu} = 0$ is of order $\sqrt{p/n} + \varepsilon\|\boldsymbol{u}\|_2 = \sqrt{p/n} + \varepsilon\sqrt{p}$. This error rate is provably optimal in the small dimensional setting $p = O(1)$, but suboptimal as compared to the optimal rate $\sqrt{p/n} + \varepsilon$ when the dimension p is not constant. The reason of this suboptimality is that the individually harmless outliers, lying close to the bulk of the point cloud, have a strong joint impact on the quality of estimation.

We postpone a review of the relevant prior work to Section 4 in order to ease comparison with our results, and proceed here with a summary of our contributions. In the context of a data set subject to a fully adversarial corruption, we introduce a new estimator of the Gaussian mean that enjoys the following properties (the precise meaning of these properties is given in Section 2):

- it is computable in polynomial time,
- it is equivariant with respect to similarity transformations (translations, uniform scaling and orthogonal transformations),
- it has a high (minimax) breakdown point: $\varepsilon^* = (5 - \sqrt{5})/10 \approx 0.28$,
- it is minimax-rate-optimal, up to a logarithmic factor,
- it is asymptotically efficient when the rate of contamination tends to zero,
- for inhomogeneous covariance matrices, it achieves a better sample complexity than all the other previously studied methods.

In order to keep the presentation simple, all the aforementioned results are established in the case where the inliers are drawn from the Gaussian distribution. We then show that the extension to a sub-Gaussian distribution can be carried out along the same lines. Furthermore, we prove that using Lepski's method, one can get rid of the knowledge of the contamination rate. More precisely, we establish that the rate $\sqrt{p/n} + \varepsilon\sqrt{\log(1/\varepsilon)}$ can be achieved without any information on ε other than $\varepsilon < (5 - \sqrt{5})/10 \approx 0.28$. Finally, we prove that the same order of magnitude of the estimation error is achieved when the covariance matrix $\boldsymbol{\Sigma}$ is unknown but isotropic (*i.e.*, proportional to the identity matrix). When the covariance matrix is an arbitrary unknown matrix with bounded operator norm, our estimator has an error of order $\sqrt{p/n} + \sqrt{\varepsilon}$, which is the best known rate of estimation by a computationally tractable procedure in the case of unknown covariance matrices.

The rest of this paper is organized as follows. We complete this introduction by presenting the notation used throughout the paper. Section 2 describes the problem setting and provides the definitions of the properties of robust estimators such as rate optimality or breakdown point. The iteratively reweighted mean estimator is introduced in Section 3. This section also contains the main facts characterizing the iteratively reweighted mean estimator along with their high-level proofs. A detailed discussion of relation to prior work is included in Section 4. Section 5 is devoted to a formal statement of the main building blocks of the proofs. Extensions to the cases of sub-Gaussian distributions, unknown ε and $\boldsymbol{\Sigma}$ are examined in Section 6. Some empirical results illustrating our theoretical claims are reported in Section 7. Postponed proofs are gathered in Section 8 and in the appendix.

For any vector $\boldsymbol{v}$, we use the norm notation $\|\boldsymbol{v}\|_2$ for the standard Euclidean norm, $\|\boldsymbol{v}\|_1$ for the sum of absolute values of entries and $\|\boldsymbol{v}\|_\infty$ for the largest in absolute value entry of $\boldsymbol{v}$. The tensor product of $\boldsymbol{v}$ by itself is denoted by $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$. We denote by $\boldsymbol{\Delta}^{n-1}$ and by $\mathbb{S}^{n-1}$, respectively, the probability simplex and the unit sphere in $\mathbb{R}^n$. For any symmetric

matrix $\mathbf{M}$, $\lambda_{\max}(\mathbf{M})$ is the largest eigenvalue of $\mathbf{M}$, while $\lambda_{\max,+}(\mathbf{M})$ is its positive part. The operator norm of $\mathbf{M}$ is denoted by $\|\mathbf{M}\|_{\text{op}}$. We will often use the effective rank $\mathbf{r}_{\mathbf{M}}$ defined as $\text{Tr}(\mathbf{M})/\|\mathbf{M}\|_{\text{op}}$, where $\text{Tr}(\mathbf{M})$ is the trace of matrix $\mathbf{M}$. For symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size we write $\mathbf{A} \succeq \mathbf{B}$, if the matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite. For a rectangular $p \times n$ matrix $\mathbf{A}$, we let $s_{\min}(\mathbf{A})$ and $s_{\max}(\mathbf{A})$ be the smallest and the largest singular values of $\mathbf{A}$ defined respectively as $s_{\min}(\mathbf{A}) = \inf_{\boldsymbol{v} \in \mathbb{S}^{n-1}} \|\mathbf{A}\boldsymbol{v}\|_2$ and $s_{\max}(\mathbf{A}) = \sup_{\boldsymbol{v} \in \mathbb{S}^{n-1}} \|\mathbf{A}\boldsymbol{v}\|_2$. The set of all $p \times p$ positive semidefinite matrices is denoted by $\mathcal{S}_+^p$.

2. Desirable properties of a robust estimator. We consider the setting in which the sample points are corrupted versions of independent and identically distributed random vectors drawn from a p-variate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}$. In what follows, we will assume that the rate of contamination and the covariance matrix are known and, therefore, can be used for constructing an estimator of $\boldsymbol{\mu}^*$. We present in Section 6 some additional results which are valid under relaxations of this assumption.

DEFINITION 1. We say that the distribution $\mathbf{P}_n$ of data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is Gaussian with adversarial contamination, denoted by $\mathbf{P}_n \in \text{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ with $\varepsilon \in (0, 1/2)$ and $\boldsymbol{\Sigma} \succeq 0$, if there is a set of n independent and identically distributed random vectors $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ drawn from $\mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ satisfying

$$\big|\{i : \boldsymbol{X}_i \neq \boldsymbol{Y}_i\}\big| \leq \varepsilon n.$$

In what follows, the sample points $\boldsymbol{X}_i$ with indices in the set $\mathcal{O} = \{i : \boldsymbol{X}_i \neq \boldsymbol{Y}_i\}$ are called outliers, while all the other sample points are called inliers. We define $\mathcal{I} = \{1, \dots, n\} \setminus \mathcal{O}$, the set of inliers. Assumption GAC allows both the set of outliers $\mathcal{O}$ and the outliers themselves to be random and to depend arbitrarily on the values of $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$. In particular, we can consider a game in which an adversary has access to the clean observations $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ and is allowed to modify an ε fraction of them before unveiling to the Statistician. The Statistician aims at estimating $\boldsymbol{\mu}^*$ as accurately as possible, the accuracy being measured by the expected estimation error:

$$R_{\mathbf{P}_n}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) = \|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*\|_{\mathbb{L}_2(\mathbf{P}_n)} = \left(\sum_{j=1}^p \mathbf{E}_{\mathbf{P}_n}[(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)_j^2]\right)^{1/2}.$$

Thus, the goal of the adversary is to apply a contamination that makes the task of estimation the hardest possible. The goal of the Statistician is to find an estimator $\widehat{\boldsymbol{\mu}}_n$ that minimizes the worst-case risk

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) = \sup_{\boldsymbol{\mu}^* \in \mathbb{R}^p} \sup_{\mathbf{P}_n \in \text{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)} R_{\mathbf{P}_n}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*).$$

Let $\mathbf{r}_{\boldsymbol{\Sigma}} = \text{Tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_{\text{op}}$ be the effective rank of $\boldsymbol{\Sigma}$. The theory developed by [Chen et al., 2016, 2018], in conjunction with [Bateni and Dalalyan, 2020, Prop. 1], implies that

$$\inf_{\widehat{\boldsymbol{\mu}}_n} R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) \geq c\|\boldsymbol{\Sigma}\|_{\text{op}}^{1/2}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\right) \tag{1}$$

for some constant $c > 0$, where the infimum is over all measurable functions of $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$. A detailed proof of this claim is presented in Appendix E of the supplementary material ([Dalalyan and Minasyan, 2021]). This lower bound naturally leads to the following definition.

DEFINITION 2. We say that the estimator $\widehat{\boldsymbol{\mu}}_n$ is minimax rate optimal (in expectation), if there are universal constants c_1, c_2 and C such that

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) \le C\|\boldsymbol{\Sigma}\|_{\rm op}^{1/2}\left(\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + \varepsilon\right)$$

for every $(n, \boldsymbol{\Sigma}, \varepsilon)$ satisfying $\mathbf{r}_{\boldsymbol{\Sigma}} \le c_1 n$ and $\varepsilon \le c_2$.

The iteratively reweighted mean estimator, introduced in the next section, is not minimax rate optimal but is very close to being so. Indeed, we will prove that it is minimax rate optimal up to a $\sqrt{\log(1/\varepsilon)}$ factor in the second term (ε is replaced by $\varepsilon\sqrt{\log(1/\varepsilon)}$). It should be stressed here that, to the best of our knowledge, none of the results on robust estimation of the Gaussian mean provides rate-optimality in expectation in the high-dimensional setting. Indeed, all those results provide risk bounds that hold with high probability, and either (a) do not say anything about the magnitude of the error on a set of small but strictly positive probability or (b) use the confidence parameter in the construction of the estimator. Both of these shortcomings prevent from extracting bounds for expected loss from high-probability bounds. This being said, it should be noted that most prior work has focused on the Huber contamination, in which case no meaningful (*i.e.*, different from $+\infty$, see Section 2.6 in [Bateni and Dalalyan, 2020]) upper bound on the minimax risk in expectation can be obtained.

DEFINITION 3. We say that $\widehat{\boldsymbol{\mu}}_n$ is an asymptotically efficient estimator of $\boldsymbol{\mu}^*$, if when $\varepsilon = \varepsilon_n$ tends to zero sufficiently fast, as n tends to infinity, we have

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) \le \|\boldsymbol{\Sigma}\|_{\rm op}^{1/2}\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}}\big(1 + o_n(1)\big).$$

If we compare inequalities in Definition 2 and Definition 3, we can see that the constant C present in the former disappeared in the latter. This reflects the fact that asymptotic efficiency implies not only rate-optimality, but also the optimality of the constant factor. We recall here that in the outlier-free situation, the sample mean is asymptotically efficient and its worst-case risk is equal to $\|\boldsymbol{\Sigma}\|_{\rm op}^{1/2}\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}$.
One can infer from (1) that a necessary condition for the existence of asymptotically efficient estimator is $\varepsilon_n^2 = o_n(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$. We show in the next section that this condition is almost sufficient, by proving that the iteratively reweighted mean estimator is asymptotically efficient provided that $\varepsilon_n^2 \log(1/\varepsilon_n) = o_n(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$.
The last notion that we introduce in this section is the breakdown point, the term being coined by Hampel [1968], see also [Donoho and Huber, 1983]. Roughly speaking, the breakdown point of a given estimator is the largest proportion of outliers that the estimator can support without becoming infinitely large. The definition we provide below slightly differs from the original one. This difference is motivated by our goal to focus on studying the expected risk of robust estimators.

DEFINITION 4. We say that $\varepsilon_n^* \in [0, 1/2]$ is the (finite-sample) breakdown point of the estimator $\widehat{\boldsymbol{\mu}}_n$, if

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) < +\infty, \qquad \forall \varepsilon < \varepsilon_n^*$$

and $R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) = +\infty$, for every $\varepsilon > \varepsilon_n^*$.

One can check that the breakdown points of the componentwise median and the geometric median (see the definition of $\widehat{\boldsymbol{\mu}}_n^{\rm GM}$ in (4) below) equal $1/2$. Unfortunately, the minimax rate of

these methods is strongly suboptimal, see [Chen et al., 2018, Prop. 2.1] and [Lai et al., 2016, Prop. 2.1]. Among all rate-optimal (up to a polylogarithmic factor) robust estimators, Tukey's median is[1] the one with highest known breakdown point equal to $1/3$ [Donoho and Gasko, 1992]. It should be noted that this paper deals with the original definition of the breakdown point which, as already mentioned, is slightly different from that of Definition 4.
The notion of breakdown point given in Definition 4, well adapted to estimators that do not rely on the knowledge of ε, becomes less relevant in the context of known ε. Indeed, if a given estimator $\widehat{\boldsymbol{\mu}}_n(\varepsilon)$ is proved to have a breakdown point equal to 0.1, one can consider instead the estimator $\widetilde{\boldsymbol{\mu}}_n(\varepsilon) = \widehat{\boldsymbol{\mu}}_n(\varepsilon)\mathbb{1}(\varepsilon < 0.1) + \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\mathbb{1}(\varepsilon \geq 0.1)$, which will have a breakdown point equal to 0.5. For this reason, it appears more appealing to consider a different notion that we call rate-breakdown point, and which is of the same flavor as the δ-breakdown point defined in [Chen et al., 2016].

DEFINITION 5. We say that $\varepsilon_r^* \in [0, 1/2]$ is the $r(n, \boldsymbol{\Sigma}, \varepsilon)$-breakdown point of the estimator $\widehat{\boldsymbol{\mu}}_n$ for a given function $r : \mathbb{N} \times \mathcal{S}_+^p \times [0, 1/2)$, if for every $\varepsilon < \varepsilon_r^*$,

$$\sup_{n,p} \frac{R_{\max}(\widehat{\boldsymbol{\mu}}_n(\varepsilon), \boldsymbol{\Sigma}, \varepsilon)}{r(n, \boldsymbol{\Sigma}, \varepsilon)} < +\infty,$$

and ε_r^* is the largest value satisfying this property.

In the context of Gaussian mean estimation, if the previous definition is applied with $r(n, \boldsymbol{\Sigma}, \varepsilon) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\big)$, we call the corresponding value the minimax-rate-breakdown point. Similarly, if $r(n, \boldsymbol{\Sigma}, \varepsilon) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\sqrt{\log(1/\varepsilon)}\big)$, we call the corresponding value the nearly-minimax-rate-breakdown point. It should be mentioned here that the extra $\sqrt{\log(1/\varepsilon)}$ factor cannot be avoided by any Statistical Query (SQ) polynomial-time algorithm, as shown by the SQ lower bound established in [Diakonikolas et al., 2017].

3. Iterative reweighting approach. In this section, we define the iterative reweighting estimator that will be later proved to enjoy all the desirable properties. To this end, we set

$$\bar{\boldsymbol{X}}_{\boldsymbol{w}} = \sum_{i=1}^n w_i \boldsymbol{X}_i, \qquad G(\boldsymbol{w}, \boldsymbol{\mu}) = \lambda_{\max,+}\bigg(\sum_{i=1}^n w_i(\boldsymbol{X}_i - \boldsymbol{\mu})^{\otimes 2} - \boldsymbol{\Sigma}\bigg) \tag{2}$$

for any pair of vectors $\boldsymbol{w} \in [0,1]^n$ and $\boldsymbol{\mu} \in \mathbb{R}^p$. The main idea of the proposed methods is to find a weight vector $\widehat{\boldsymbol{w}}_n$ belonging to the probability simplex

$$\boldsymbol{\Delta}^{n-1} = \Big\{\boldsymbol{w} \in [0,1]^n : w_1 + \ldots + w_n = 1\Big\}$$

that mimics the ideal weight vector $\boldsymbol{w}^*$ defined by $w_j^* = \mathbb{1}(j \in \mathcal{I})/|\mathcal{I}|$, so that the weighted average $\bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}_n}$ is nearly as close to $\boldsymbol{\mu}^*$ as the average of the inliers. Note that, for any weight vector $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ and any vector $\boldsymbol{\mu} \in \mathbb{R}^p$, we have

$$\sum_{i=1}^n w_i(\boldsymbol{X}_i - \boldsymbol{\mu})^{\otimes 2} = \sum_{i=1}^n w_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2} + (\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu})^{\otimes 2} \succeq \sum_{i=1}^n w_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2}.$$

This readily yields $G(\boldsymbol{w}, \boldsymbol{\mu}) \geq G(\boldsymbol{w}, \bar{\boldsymbol{X}}_{\boldsymbol{w}})$, and, therefore

$$G(\boldsymbol{w}, \bar{\boldsymbol{X}}_{\boldsymbol{w}}) = \min_{\boldsymbol{\mu} \in \mathbb{R}^n} G(\boldsymbol{w}, \boldsymbol{\mu}), \qquad \forall \boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}. \tag{3}$$

The precise definition of the proposed estimator is as follows. We start from an initial esti-

[1] Recent results in [Zhu et al., 2020] suggest that an estimator based on the TV-projection may achieve the optimal breakdown point of $1/2$.

Algorithm 1: Iteratively reweighted mean estimator (known ε and $\boldsymbol{\Sigma}$)

Input: data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{R}^p$, contamination rate ε and $\boldsymbol{\Sigma}$
Output: parameter estimate $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$
Initialize: compute $\widehat{\boldsymbol{\mu}}^0$ as a minimizer of $\sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2$

Set $K = 0 \vee \left\lceil \frac{\log(4\mathbf{r}_{\boldsymbol{\Sigma}}) - 2\log(\varepsilon(1-2\varepsilon))}{2\log(1-2\varepsilon) - \log\varepsilon - \log(1-\varepsilon)} \right\rceil$.

For $k = 1 : K$
Compute current weights:
$$\boldsymbol{w} \in \operatorname*{arg\,min}_{(n-n\varepsilon)\|\boldsymbol{w}\|_\infty \le 1} \lambda_{\max}\Big(\sum_{i=1}^n w_i (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})^{\otimes 2} - \boldsymbol{\Sigma}\Big) \vee 0.$$
Update the estimator: $\widehat{\boldsymbol{\mu}}^k = \sum_{i=1}^n w_i \boldsymbol{X}_i$.
EndFor
Return $\widehat{\boldsymbol{\mu}}^K$.

mator $\widehat{\boldsymbol{\mu}}^0$ of $\boldsymbol{\mu}^*$. To give a concrete example, and also in order to guarantee equivariance by similarity transformations, we assume that $\widehat{\boldsymbol{\mu}}^0$ is the geometric median:

$$\widehat{\boldsymbol{\mu}}^0 = \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} \in \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2. \tag{4}$$

DEFINITION 6. We call iteratively reweighted mean estimator, denoted by $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$, the K-th element of the sequence $\{\widehat{\boldsymbol{\mu}}^k; k = 0, 1, \dots\}$ starting from $\widehat{\boldsymbol{\mu}}^0$ in (4) and defined by the recursion

$$\widehat{\boldsymbol{w}}^k \in \arg\min_{(n-n\varepsilon)\|\boldsymbol{w}\|_\infty \le 1} G(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}^{k-1}), \qquad \widehat{\boldsymbol{\mu}}^k = \bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}^k}, \tag{5}$$

where the minimum is over all weight vectors $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ satisfying $\max_j w_j \le 1/(n - n\varepsilon)$ and the number of iteration is

$$K = 0 \bigvee \left\lceil \frac{\log(4\mathbf{r}_{\boldsymbol{\Sigma}}) - 2\log(\varepsilon(1-2\varepsilon))}{2\log(1-2\varepsilon) - \log\varepsilon - \log(1-\varepsilon)} \right\rceil. \tag{6}$$

The idea of computing a weighted mean, with weights measuring the outlyingness of the observations goes back at least to [Donoho, 1982, Stahel, 1981]. Perhaps the first idea similar to that of minimizing the largest eigenvalue of the covariance matrix was that of minimizing the determinant of the sample covariance matrix over all subsamples of a given cardinality [Rousseeuw, 1985, 1984]. It was also observed in [Lopuhaä and Rousseeuw, 1991] that one can improve the estimator by iteratively updating the weights. An overview of these results can be found in [Rousseeuw and Hubert, 2013].

Note that the value of K provided above is tailored to the case where the initial estimator is the geometric median. Clearly, K depends only logarithmically on the dimension and $K = K_\varepsilon$ tends to 2 when ε goes to zero, see Figure 1. Note also that the choice of K in (6) is derived from the condition $\big(\sqrt{\varepsilon(1-\varepsilon)}/(1-2\varepsilon)\big)^K \|\widehat{\boldsymbol{\mu}}^0 - \boldsymbol{\mu}^*\|_{\mathbf{L}_2} \le \varepsilon$, see (10) below. We can use any other initial estimator of $\boldsymbol{\mu}^*$ instead of the geometric median, provided that K is large enough to satisfy the last inequality.

We have to emphasize that $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ relies on the knowledge of both ε and $\boldsymbol{\Sigma}$ (the dependence on $\boldsymbol{\Sigma}$ is through the effective rank, which coincides with the dimension for non degenerate

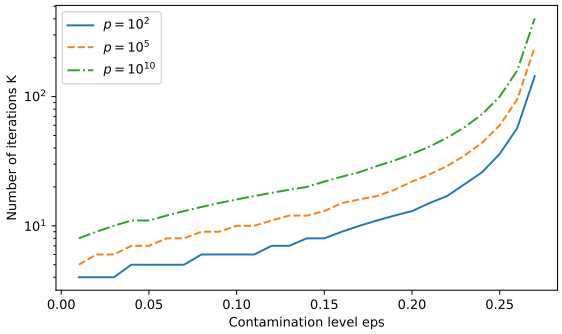

FIG 1. *The behavior of the number of iterations $K = K_\varepsilon$, given by (6), as a function of the contamination rate ε for different values of the dimension $p = \mathbf{r}_{\boldsymbol{\Sigma}}$.*

covariance matrices). Indeed, the number of iterations K depends on both ε and $\boldsymbol{\Sigma}$. Additionally, $\boldsymbol{\Sigma}$ is used in the cost function $G(\boldsymbol{w}, \boldsymbol{\mu})$ and ε is used for specifying the set of feasible weights in optimization problem (5). We present some extensions to the case of unknown ε and $\boldsymbol{\Sigma}$ in Section 6.

The rest of this section is devoted to showing that the iteratively reweighted estimator enjoys all the desirable properties announced in the introduction. An estimator is called computationally tractable, if its computational complexity is at most polynomial in n, p and $1/\varepsilon$.

Fact 1

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is computationally tractable.

In order to check computational tractability, it suffices to prove that each iteration of the algorithm can be performed in polynomial time. Since the number of iterations depends logarithmically on $\mathbf{r}_{\boldsymbol{\Sigma}} \leq p$, this will suffice. Note now that the optimization problem in (5) is convex and can be cast into a semi-definite program. Indeed, it is equivalent to minimizing a real value t over all the pairs $(t, \boldsymbol{w})$ satisfying the constraints

$$t \geq 0, \quad \boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}, \quad \|\boldsymbol{w}\|_\infty \leq \frac{1}{n(1-\varepsilon)}, \quad \sum_{i=1}^n w_i (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})^{\otimes 2} \preceq \boldsymbol{\Sigma} + t\mathbf{I}_p.$$

The first two constraints can be rewritten as a set of linear inequalities, while the third constraint is a linear matrix inequality. Given the special form of the cost function and the constraints, it is possible to design specific optimization routines which will find an approximate solution to the problem in a faster way than the out-of-shelf SDP-solvers. However, we will not pursue this line of research in this work.

Fact 2

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is translation, uniform scaling and orthogonal transformation equivariant.

The equivariance mentioned in this statement should be understood as follows. If we denote by $\widehat{\boldsymbol{\mu}}_{n,X}^{\mathrm{IR}}$ the estimator computed for data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and by $\widehat{\boldsymbol{\mu}}_{n,X'}^{\mathrm{IR}}$ the one computed for data $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$, with $\boldsymbol{X}'_i = \boldsymbol{a} + \lambda \mathbf{U} \boldsymbol{X}_i$, where $\boldsymbol{a} \in \mathbb{R}^p$, $\lambda > 0$ and $\mathbf{U}$ is a $p \times p$ orthogonal

matrix, then $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_{n,X'} = \boldsymbol{a} + \lambda \mathbf{U} \widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_{n,X}$. To prove this property, we first note that

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}'_i - \boldsymbol{\mu}\|_2 = \lambda \min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}_i - \lambda^{-1}\mathbf{U}^\top(\boldsymbol{\mu} - \boldsymbol{a})\|_2.$$

This implies that $\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X} = \lambda^{-1}\mathbf{U}^\top(\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X'} - \boldsymbol{a})$, which is equivalent to $\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X'} = \boldsymbol{a} + \lambda\mathbf{U}\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X}$. Therefore, the initial value of the recursion is equivariant. If we add to this the fact that[2] $G_X(\boldsymbol{w}, \boldsymbol{\mu}) = \lambda^2 G_{X'}(\boldsymbol{w}, \boldsymbol{a} + \lambda\mathbf{U}\boldsymbol{\mu})$ for every $(\boldsymbol{w}, \boldsymbol{\mu})$, we get the equivariance of $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n$.

Fact 3

The breakdown point ε^*_n and the nearly-minimax-rate-breakdown point ε^*_r of $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n$ satisfy, respectively, $\varepsilon^*_n = 0.5$ and $\varepsilon^*_r \geq (5 - \sqrt{5})/10 \approx 0.28$.

We prove later in this paper (see (14)) that if $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ satisfy GAC$(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$, there is a random variable Ξ depending only on $\boldsymbol{\zeta}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}^*$, $i = 1, \dots, n$, such that

$$\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*\|_2 \leq \frac{\sqrt{\varepsilon(1-\varepsilon)}}{1-2\varepsilon} G(\boldsymbol{w}, \boldsymbol{\mu})^{1/2} + \Xi, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^p, \tag{7}$$

for every $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ such that $n(1-\varepsilon)\|\boldsymbol{w}\|_\infty \leq 1$. Inequality (7) is one of the main building blocks of the proof of Facts 3 to 5. This inequality, as well as inequalities (11) and (12) below will be formally stated and proved in subsequent sections. To check Fact 3, we set $\alpha_\varepsilon = \sqrt{\varepsilon(1-\varepsilon)}/(1-2\varepsilon)$ and note that[3]

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}^k - \boldsymbol{\mu}^*\|_2 = \|\bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}^k} - \boldsymbol{\mu}^*\|_2 &\leq \alpha_\varepsilon\, G(\widehat{\boldsymbol{w}}^k, \widehat{\boldsymbol{\mu}}^{k-1})^{1/2} + \Xi \\
&\leq \alpha_\varepsilon\, G(\boldsymbol{w}^*, \widehat{\boldsymbol{\mu}}^{k-1})^{1/2} + \Xi \\
&\leq \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \bar{\boldsymbol{X}}_{\boldsymbol{w}^*}) + \|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\|_2^2\right)^{1/2} + \Xi \\
&\leq \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \boldsymbol{\mu}^*) + \|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\|_2^2\right)^{1/2} + \Xi \\
&\leq \alpha_\varepsilon\, \|\widehat{\boldsymbol{\mu}}^{k-1} - \boldsymbol{\mu}^*\|_2 + \widetilde{\Xi},
\end{aligned} \tag{8}$$

where $\widetilde{\Xi} = \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \boldsymbol{\mu}^*)^{1/2} + \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}^*}\|_2\right) + \Xi$. Unfolding this recursion, we get [4]

$$\|\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n - \boldsymbol{\mu}^*\|_2 = \|\widehat{\boldsymbol{\mu}}^K - \boldsymbol{\mu}^*\|_2 \leq \alpha_\varepsilon^K\, \|\widehat{\boldsymbol{\mu}}^0 - \boldsymbol{\mu}^*\|_2 + \frac{\widetilde{\Xi}}{1-\alpha_\varepsilon}. \tag{9}$$

The geometric median $\widehat{\boldsymbol{\mu}}^0 = \widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_n$ having a breakdown point equal to $1/2$, we infer from the last display that the error of the iteratively reweighted estimator remains bounded after altering ε-fraction of data points provided that $\alpha_\varepsilon < 1$. This implies that the breakdown point is at least equal to the solution of the equation $\sqrt{\varepsilon(1-\varepsilon)} = 1 - 2\varepsilon$, which yields $\varepsilon^* \geq (5 - \sqrt{5})/10$. Moreover, if $\varepsilon \in [(5 - \sqrt{5})/10, 1/2]$, then the number of iterations K equals zero and the iteratively reweighted mean coincides with the geometric median. Therefore, its breakdown point is $1/2$.

[2] We use here the notation $G_X(\boldsymbol{w}, \boldsymbol{\mu})$ to make clear the dependence of G in (2) on $\boldsymbol{X}_i$s. We also stress that when the estimator is computed for the transformed data $\boldsymbol{X}'_i$, the matrix $\boldsymbol{\Sigma}$ is naturally replaced by $\lambda^2\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$.

[3] See Section 8.1 for more detailed explanations.

[4] Here and in the sequel α_ε^K stands for K-th power of α_ε.

Fact 4

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is nearly minimax rate optimal, in the sense that its worst-case risk is bounded by $C\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}+\varepsilon\sqrt{\log(1/\varepsilon)}\big)$, where C is a universal constant.

Without loss of generality, we assume that $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}=1$ so that $\mathbf{r}_{\boldsymbol{\Sigma}}=\mathrm{Tr}(\boldsymbol{\Sigma})$. We can always reduce the initial problem to this case by considering scaled data points $\boldsymbol{X}_i/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$ instead of $\boldsymbol{X}_i$. Combining (9) and the triangle inequality, we get

$$\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}-\boldsymbol{\mu}^*\|_{\mathbb{L}_2}\le\alpha_\varepsilon^K\,\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}-\boldsymbol{\mu}^*\|_{\mathbb{L}_2}+\frac{\|\widetilde{\Xi}\|_{\mathbb{L}_2}}{1-\alpha_\varepsilon}. \tag{10}$$

It is not hard to check that $\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}-\boldsymbol{\mu}^*\|_{\mathbb{L}_2}\le 2\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}/(1-2\varepsilon)$, see Lemma 2 below. Furthermore, the choice of K in (6) entails $2\alpha_\varepsilon^K\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}\le\varepsilon(1-2\varepsilon)$. This implies that

$$\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}-\boldsymbol{\mu}^*\|_{\mathbb{L}_2}\le\varepsilon+\frac{\|\widetilde{\Xi}\|_{\mathbb{L}_2}}{1-\alpha_\varepsilon}.$$

The last two building blocks of the proof are the following[5] inequalities:

$$\mathbf{E}[G(\boldsymbol{w}^*,\boldsymbol{\mu}^*)]\le C\big(1+\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big)\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}, \tag{11}$$

$$\|\Xi\|_{\mathbb{L}_2}\le\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}(1+C\sqrt{\varepsilon})+C\sqrt{\varepsilon}(\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4}+C\varepsilon\sqrt{\log(1/\varepsilon)}, \tag{12}$$

where $C>0$ is a universal constant. In what follows, the value of C may change from one line to the other. We have

$$\begin{aligned}
\|\widetilde{\Xi}\|_{\mathbb{L}_2}&\le\alpha_\varepsilon\big(\|G(\boldsymbol{w}^*,\boldsymbol{\mu}^*)^{1/2}\|_{\mathbb{L}_2}+\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}^*}\|_{\mathbb{L}_2}\big)+\|\Xi\|_{\mathbb{L}_2}\\
&\le C\sqrt{\varepsilon}\,\big(\mathbf{E}^{1/2}[G(\boldsymbol{w}^*,\boldsymbol{\mu}^*)]+\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big)+\|\Xi\|_{\mathbb{L}_2}^2\\
&\le C\varepsilon\,\big((\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4}+\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big)+\|\Xi\|_{\mathbb{L}_2}\\
&\le\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}(1+C\sqrt{\varepsilon})+C\sqrt{\varepsilon}(\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4}+C\varepsilon\sqrt{\log(1/\varepsilon)}\\
&\le C\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}+C\varepsilon\sqrt{\log(1/\varepsilon)}.
\end{aligned} \tag{13}$$

Returning to (9) and combining it with (13), we get the claim of Fact 4 for every $\varepsilon\le\varepsilon_0$, where ε_0 is any positive number strictly smaller than $(5-\sqrt{5})/10$. This also proves the second claim of Fact 3.

Fact 5

In the setting $\varepsilon=\varepsilon_n\to 0$ so that $\varepsilon^2\log(1/\varepsilon)=o_n(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$ when $n\to\infty$, the estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is asymptotically efficient.

The proof of this fact follows from (9) and (12). Indeed, if $\varepsilon^2\log(1/\varepsilon)=o_n(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$, (12) implies that

$$\|\widetilde{\Xi}\|_{\mathbb{L}_2}^2\le\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}\big(1+o_n(1)\big).$$

Injecting this bound in (9) and using the fact that ε tends to zero, we get the claim of Fact 5.

[5] Inequality (11) is [Koltchinskii and Lounici, 2017, Th. 4], while (12) is the claim of Proposition 2 below.

4. Relation to prior work and discussion. Robust estimation of a mean is a statistical problem studied by many authors since at least sixty years. It is impossible to give an overview of all existing results and we will not try to do it here. The interested reader may refer to the books [Maronna et al., 2006] and [Huber and Ronchetti, 2009]. We will rather focus here on some recent results that are the most closely related to the present work. Let us just recall that Huber and Ronchetti [2009] enumerates three desirable properties of a statistical procedure: efficiency, stability and breakdown. We showed here that iteratively reweighted mean estimator possesses these features and, in addition, is equivariant and computationally tractable.

To the best of our knowledge, the form $\sqrt{p/n} + \varepsilon$ of the minimax risk in the Gaussian mean estimation problem has been first obtained by Chen et al. [2018]. They proved that this rate holds with high probability for the Tukey median, which is known to be computationally intractable in the high-dimensional setting. The first nearly-rate-optimal and computationally tractable estimators have been proposed by Lai et al. [2016] and Diakonikolas et al. [2016][6]. The methods analyzed in these papers are different, but they share the same idea: If for a subsample of points the empirical covariance matrix is sufficiently close to the theoretical one, then the arithmetic mean of this subsample is a good estimator of the theoretical mean. Our method is based on this idea as well, which is mathematically formalized in (7), see also Proposition 1 below.

Further improvements in running times—up to obtaining a linear in np computational complexity in the case of a constant ε—are presented in [Cheng et al., 2019]. Some lower bounds suggesting that the log-factor in the term $\varepsilon\sqrt{\log(1/\varepsilon)}$ cannot be removed from the rate of computationally tractable estimators are established in [Diakonikolas et al., 2017]. In a slightly weaker model of corruption, Diakonikolas et al. [2018] propose an iterative filtering algorithm that achieves the optimal rate ε without the extra factor $\sqrt{\log(1/\varepsilon)}$. On a related note, [Collier and Dalalyan, 2019] shows that in a weaker contamination model termed as parametric contamination, the carefully trimmed sample mean can achieve a better rate than that of the coordinatewise/geometric median.

An overview of the recent advances on robust estimation with a focus on computational aspects can be found in [Diakonikolas and Kane, 2019]. Extensions of these methods to the sparse mean estimation are developed in [Balakrishnan et al., 2017, Diakonikolas et al., 2019b]. All these results are proved to hold on an event with a prescribed probability, see [Bateni and Dalalyan, 2020] for a relation between results in expectation and those with high probability, as well as for the definitions of various types of contamination.

The proposed estimator shares some features with the adaptive weights smoothing [Polzehl and Spokoiny, 2000]. Adaptive weights smoothing (AWS) iteratively updates the weights assigned to observations, similarly to Algorithm 1. The main difference is that the weights in AWS are not measuring the outlyingness but the relevance for interpolating a function at a given point. There are also many other statistical problems in which robust estimation has been recently revisited from the point of view of minimax rates. This includes scale and covariance matrix estimation [Chen et al., 2018, Comminges et al., 2020], matrix completion [Klopp et al., 2017], multivariate regression [Chinot, 2020, Dalalyan and Thompson, 2019, Gao, 2020], classification [Cannings et al., 2020, Li and Bradic, 2018], subspace clustering [Soltanolkotabi and Candès, 2012], community detection [Cai and Li, 2015], etc. Properties of robust M-estimators in high-dimensional settings are studied in [Elsener and van de Geer, 2018, Loh, 2017]. There is also an increasing body of literature on the robustness to heavy tailed distributions [Devroye et al., 2016, Lecué and Lerasle, 2019, Lecué and Lerasle, 2020, Lugosi and Mendelson, 2019, 2020, Minsker, 2018] and the computationally tractable

[6] See [Diakonikolas et al., 2019a] for the extended version

methods in this context [Cherapanamjeri et al., 2019, Depersin and Lecué, 2019, Dong et al., 2019, Hopkins, 2018].
A potentially useful observation, from a computational standpoint, is that it is sufficient to solve the optimization problem in Equation (5) up to an error proportional to $\sqrt{\mathbf{r}_{\mathbf{\Sigma}}/n} + \sqrt{\varepsilon}$. Indeed, one can easily repeat all the steps in (8) to check that this optimization error does not alter the order of magnitude of the statistical error.

5. Formal statement of main building blocks. The first building block, inequality (7), used in Section 3 to analyze the risk of $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$, upper bounds the error of estimating the mean by the error of estimating the covariance matrix. In order to formally state the result, we need some additional notation.
Let $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ be a vector of weights and let I be a subset of $\{1,\dots,n\}$. We use the notation $\boldsymbol{w}_I$ for the vector obtained from $\boldsymbol{w}$ by zeroing all the entries having indices outside I. Considering $\boldsymbol{w}$ as a probability on $\{1,\dots,n\}$, we define $\boldsymbol{w}_{|I}$ as the corresponding conditional probability on I that is

$$\boldsymbol{w}_{|I} \in \boldsymbol{\Delta}^{n-1}, \qquad (\boldsymbol{w}_{|I})_i = (w_i/\|\boldsymbol{w}_I\|_1)\mathbb{1}(i \in I).$$

We will make repeated use of the notation

$$\bar{\boldsymbol{X}}_{\boldsymbol{w}} = \sum_{i=1}^n w_i \boldsymbol{X}_i, \qquad \bar{\boldsymbol{\xi}}_{\boldsymbol{w}_{|I}} = \sum_{i\in I} (\boldsymbol{w}_{|I})_i \boldsymbol{\xi}_i, \qquad \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}} = \sum_{i\in I} (\boldsymbol{w}_{|I})_i \boldsymbol{\zeta}_i.$$

PROPOSITION 1. *Let $\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_n$ be a set of vectors such that $\boldsymbol{\zeta}_i = \boldsymbol{X}_i - \boldsymbol{\mu}^*$ for every $i \in \mathcal{I}$, where $\mathcal{I}$ is a subset of $\{1,\dots,n\}$. For every weight vector $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ such that $\sum_{i\notin\mathcal{I}} w_i \le \varepsilon_w \le 1/2$ and for every $p \times p$ matrix $\mathbf{\Sigma}$, it holds*

$$\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*\|_2 \le \frac{\sqrt{\varepsilon_w}}{1-\varepsilon_w}\, \lambda_{\max,+}^{1/2}\Big(\sum_{i=1}^n w_i (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2} - \mathbf{\Sigma}\Big) + R(\boldsymbol{\zeta}, \boldsymbol{w}, \mathcal{I}),$$

with the remainder term

$$R(\boldsymbol{\zeta}, \boldsymbol{w}, \mathcal{I}) = 2\sqrt{\|\mathbf{\Sigma}\|_{\mathrm{op}}\varepsilon_w} + \sqrt{2\varepsilon_w}\, \lambda_{\max,+}^{1/2}\Big(\sum_{i\in\mathcal{I}} (\boldsymbol{w}_{|\mathcal{I}})_i (\mathbf{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big) + (1+\sqrt{2\varepsilon_w})\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\|_2.$$

The proof of this result is postponed to the last section. In simple words, the claim of proposition is that the estimation error of the weighted mean $\bar{\boldsymbol{X}}_{\boldsymbol{w}}$ is, up to a remainder term, governed by the quantity $G(\boldsymbol{w}, \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{1/2}$. It turns out that the remainder term is bounded by a small quantity uniformly in $\boldsymbol{w}$ and $\mathcal{I}$, provided that these two satisfy suitable conditions. For $\mathcal{I}$, it is enough to constrain the cardinality of its complement $\mathcal{I}^c = \mathcal{O}$. For $\boldsymbol{w}$, it appears to be sufficient to assume that its sup-norm is small. In that respect, the following lemma plays a key role in the proof.

LEMMA 1. *For any integer $\ell > 0$, let $\mathcal{W}_{n,\ell}$ be the set of all $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ such that $\max_i w_i \le 1/\ell$. The following facts hold:*
i) *For every $J \subset \{1,\dots,n\}$ such that $|J| \ge \ell$, the uniform weight vector $\boldsymbol{u}^J \in \mathcal{W}_{n,\ell}$.*
ii) *The set $\mathcal{W}_{n,\ell}$ is the convex hull of the uniform weight vectors $\{\boldsymbol{u}^J : |J| = \ell\}$.*
iii) *For every convex mapping $G : \boldsymbol{\Delta}^{n-1} \to \mathbb{R}$, we have*

$$\sup_{\boldsymbol{w}\in\mathcal{W}_{n,\ell}} G(\boldsymbol{w}) = \max_{|J|=\ell} G(\boldsymbol{u}^J),$$

where the last maximum is over all subsets J of cardinality ℓ of the set $\{1,\dots,n\}$.

iv) *If $\boldsymbol{w} \in \mathcal{W}_{n,\ell}$ then for any I such that $|I| \geq \ell' > n - \ell$, we have $\boldsymbol{w}_{|I} \in \mathcal{W}_{n,\ell+\ell'-n}$.*

Let us denote by $\mathcal{W}_n(\varepsilon)$ the set $\mathcal{W}_{n,n(1-\varepsilon)}$. This is exactly the feasible set in the optimization problem defining the iterations of Algorithm 1. It is clear that for $\boldsymbol{w} \in \mathcal{W}_n(\varepsilon)$ and for $|I^c| \leq n\varepsilon$, we have $\sum_{i \notin I} w_i \leq \varepsilon/(1-\varepsilon)$. We now infer from Proposition 1 and (3) that

$$\begin{aligned}\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*\|_2 &\leq \alpha_\varepsilon \, \lambda_{\max,+}^{1/2}\Big(\sum_{i=1}^n w_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2} - \boldsymbol{\Sigma}\Big) + \Xi \\ &= \alpha_\varepsilon \inf_{\boldsymbol{\mu} \in \mathbb{R}^p} G(\boldsymbol{w}, \boldsymbol{\mu})^{1/2} + \Xi, \end{aligned} \tag{14}$$

with Ξ being the largest value of $R(\boldsymbol{\zeta}, \boldsymbol{w}, I)$ over all possible weights $\boldsymbol{w} \in \mathcal{W}_n(\varepsilon)$ and subsets $I \subset \{1, \dots, n\}$ satisfying $|I^c| \leq n\varepsilon$. The second building block, formally stated in the next proposition, provides a suitable upper bound for the random variable Ξ.

PROPOSITION 2. *Let $R(\boldsymbol{\zeta}, \boldsymbol{w}, I)$ be defined in Proposition 1 and $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n$ be i.i.d. centered Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}$ satisfying $\lambda_{\max}(\boldsymbol{\Sigma}) = 1$. If $\varepsilon \leq 0.28$, then the random variable*

$$\Xi = \sup_{\boldsymbol{w} \in \mathcal{W}_n(\varepsilon)} \max_{|I| \geq n(1-\varepsilon)} R(\boldsymbol{\zeta}, \boldsymbol{w}, I)$$

satisfies, for a universal constant $C > 0$, the inequalities

$$\|\Xi\|_{\mathbb{L}_2} \leq \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}(1 + C\sqrt{\varepsilon}) + C\sqrt{\varepsilon}(\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4} + C\varepsilon\sqrt{\log(1/\varepsilon)} \tag{15}$$

$$\|\Xi\|_{\mathbb{L}_2} \leq \sqrt{p/n}(1 + 16\sqrt{\varepsilon}) + 5\sqrt{3\varepsilon}\,(p/n)^{1/4} + 32\varepsilon\sqrt{\log(2/\varepsilon)}, \tag{16}$$

where for the second inequality we assumed that $p \geq 2$ and $n \geq p \vee 4$.

The second inequality is weaker than the first one, since obviously $\mathbf{r}_{\boldsymbol{\Sigma}} \leq p$. However, the advantage of the second inequality is that it comes with explicit constants and shows that these constants are not excessively large.

To close this section, we state a theorem that rephrases Fact 4 in a way that might be more convenient for future references. Its proof is omitted, since it follows the lines of the proof of Fact 4 presented above.

THEOREM 1. *Let $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ be the iteratively reweighted mean defined in Definition 6 and Algorithm 1. There is a universal constant $C > 0$ such that for any $n, p \geq 1$ and for every $\varepsilon < (5 - \sqrt{5})/10$, we have*

$$\sup_{\boldsymbol{\mu}^* \in \mathbb{R}^p} \sup_{\mathbf{P}_n \in \mathrm{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)} \mathbf{E}^{1/2}[\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_2^2] \leq \frac{C\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}}{1 - 2\varepsilon - \sqrt{\varepsilon(1-\varepsilon)}}\left(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\sqrt{\log(1/\varepsilon)}\right),$$

where $\mathbf{P}_n \in \mathrm{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ means that the data points $\boldsymbol{X}_i$ are Gaussian with adversarial contamination, see Definition 1. If, in addition, $p \geq 2$ and $n \geq p \vee 10$, then

$$\sup_{\boldsymbol{\mu}^* \in \mathbb{R}^p} \sup_{\mathbf{P}_n \in \mathrm{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)} \mathbf{E}^{1/2}[\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_2^2] \leq \frac{10\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}}{1 - 2\varepsilon - \sqrt{\varepsilon(1-\varepsilon)}}\left(\sqrt{p/n} + \varepsilon\sqrt{\log(1/\varepsilon)}\right).$$

To the best of our knowledge, this is the first result in the literature that provides an upper bound on the expected error of an outlier-robust estimator, which is of nearly optimal rate.

6. Sub-Gaussian distributions, high-probability bounds and adaptation. Risk bounds stated in Fact 4 and Fact 5 and formalized in Theorem 1 hold for the expected error under the condition that the reference distribution is Gaussian. Furthermore, the proposed procedure relies on the knowledge of both the contamination rate ε and the covariance matrix $\mathbf{\Sigma}$. The goal of this section is to show how some of these restrictions can be alleviated.

6.1. *High-probability risk bound for a sub-Gaussian reference distribution.* As expected, the risk bounds established in previous sections can be extended to the case of sub-Gaussian distributions. Furthermore, risk bounds holding with high-probability can be proved using the same techniques as those employed for proving the in-expectation bounds. In order to be more precise, we state in this subsection the high-probability counterpart of the second claim of Theorem 1. The price to pay for covering the more general sub-Gaussian case is that the constant in the right hand side of the inequality is no longer explicit.
Recall that a zero-mean random vector $\boldsymbol{\xi}$ is called sub-Gaussian with parameter $\tau > 0$ (also known as the variance proxy), if

$$\mathbf{E}[e^{\boldsymbol{v}^\top \boldsymbol{\xi}}] \leq e^{\tau/2\|\boldsymbol{v}\|_2^2}, \qquad \forall \boldsymbol{v} \in \mathbb{R}^p.$$

We write $\boldsymbol{\xi} \sim SG_p(\tau)$. If $\boldsymbol{\xi}$ is standard Gaussian then it is sub-Gaussian with parameter 1. Similarly, if $\boldsymbol{\xi}$ is centered and belongs almost surely to the unit ball, then $\boldsymbol{\xi}$ is sub-Gaussian with parameter 1. Let us describe now the set of data-generating distributions that we consider in this section.

DEFINITION 7. We say that the joint distribution $\mathbf{P}_n$ of the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ belongs to the sub-Gaussian model with adversarial contamination with mean $\boldsymbol{\mu}^*$, covariance matrix $\mathbf{\Sigma}$ and contamination rate ε, if there are independent random vectors $\boldsymbol{\xi}_i \sim SG_p(\tau)$ such that

$$\left| \left\{ i := 1, \ldots, n : \boldsymbol{X}_i \neq \boldsymbol{\mu}^* + \mathbf{\Sigma}^{1/2} \boldsymbol{\xi}_i \right\} \right| \leq \varepsilon n.$$

We then write $\mathbf{P}_n \in \mathrm{SGAC}(\boldsymbol{\mu}^*, \mathbf{\Sigma}, \varepsilon)$.

It is clear that a Gaussian model with adversarial contamination defined in Definition 1 is a particular case of the sub-Gaussian model with adversarial contamination. In other terms, the set $\mathrm{SGAC}(\boldsymbol{\mu}^*, \mathbf{\Sigma}, \varepsilon)$ is strictly larger than the set $\mathrm{GAC}(\boldsymbol{\mu}^*, \mathbf{\Sigma}, \varepsilon)$. Nevertheless, as shows the result below, the risk bounds established for the iteratively reweighted mean algorithms remain valid uniformly over this extended class $\mathrm{SGAC}(\boldsymbol{\mu}^*, \mathbf{\Sigma}, \varepsilon)$.

THEOREM 2. *Let $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ be the iteratively reweighted mean defined in Definition 6 and in Algorithm 1. Let $\delta \in (4e^{-n}, 1)$ be a tolerance level. There exists a constant A_5 depending only on the variance proxy τ such that if $n \geq p \geq 2$ and $\varepsilon < (5 - \sqrt{5})/10$, then for every $\boldsymbol{\mu}^* \in \mathbb{R}^p$ and every $\mathbf{P}_n \in \mathrm{SGAC}(\boldsymbol{\mu}^*, \mathbf{\Sigma}, \varepsilon)$, we have*

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_2 \leq \frac{A_5 \|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}}{1 - 2\varepsilon - \sqrt{\varepsilon(1-\varepsilon)}} \left(\sqrt{\frac{p + \log(4/\delta)}{n}} + \varepsilon \sqrt{\log(1/\varepsilon)} \right) \right) \geq 1 - 4\delta.$$

The proof of this theorem is postponed to the supplementary material. Let us just mention that [Cheng et al., 2019, Section 1.2] claim that the rate $\sqrt{p/n} + \varepsilon\sqrt{\log(1/\varepsilon)}$ is optimal for sub-Gaussian distributions, meaning that, unlike the Gaussian case, the $\sqrt{\log(1/\varepsilon)}$ factor cannot be removed. A formal proof of this fact can be found in the last remark of Section 2 in [Lugosi and Mendelson, 2021].

6.2. *Adaptation to unknown contamination rate ε.* An appealing feature of the risk bounds that hold with high probability is that they allow us to apply Lepski's method [Lepski and Spokoiny, 1997, Lepskii, 1992] for obtaining an adaptive estimator with respect to ε. The obtained adaptive estimator enjoys all the five properties enumerated in Section 3 except the asymptotic efficiency, since the adaptation results in an inflation of the risk bound by a factor 3. The precise description of the algorithm, already used in the framework of robust estimation by Collier and Dalalyan [2019], is presented below. We will denote by $\mathbb{B}(\boldsymbol{\mu}, r)$ the ball with center $\boldsymbol{\mu}$ and radius r in the Euclidean space $\mathbb{R}^p$.

DEFINITION 8. We choose a geometric grid $\varepsilon_\ell = a^\ell \varepsilon_0$, $\ell = 1, 2, \dots, \ell_{\max}$, of possible values of the contamination rate. Here, $a \in (0,1)$ is a real number, $\varepsilon_0 = (5-\sqrt{5})/10$ and $\ell_{\max} = [0.5 \log_a(p/n)]$. For each $\ell = 1, \dots, \ell_{\max}$, we denote by $\widehat{\boldsymbol{\mu}}_n^{\text{IR}}(\varepsilon_\ell)$ the iteratively reweighted mean computed for $\varepsilon = \varepsilon_\ell$, see Algorithm 1, and we set

$$R_\delta(z) = \frac{A_5 \|\boldsymbol{\Sigma}\|_{\text{op}}^{1/2}}{1 - 2z - \sqrt{z(1-z)}} \left(\sqrt{\frac{p + \log(4\ell_{\max}/\delta)}{n}} + z\sqrt{\log(1/z)} \right), \qquad z \in [0, \varepsilon_0),$$

where $\delta \in (0,1)$ is a tolerance level and A_5 is the constant from Theorem 2. The adaptively chosen iteratively reweighted mean estimator $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$ is defined by $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}} = \widehat{\boldsymbol{\mu}}_n^{\text{IR}}(\varepsilon_{\widehat{\ell}})$ where

$$\widehat{\ell} = \max \Big\{ \ell \le \ell_{\max} : \bigcap_{j=1}^{\ell} \mathbb{B}\big(\widehat{\boldsymbol{\mu}}_n^{\text{IR}}(\varepsilon_j); R_\delta(\varepsilon_j)\big) \neq \varnothing \Big\}.$$

The estimator $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$ can be computed without the knowledge of the true contamination rate ε. Furthermore, its computational complexity nearly of the same order as the complexity of computing a single instance of the iteratively reweighted mean as defined by Algorithm 1. Indeed, to compute $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$, one needs to apply Algorithm 1 at most $\ell_{\max} = [0.5 \log_a(p/n)]$ times, and to solve a second-order cone program for checking whether the intersection of a small number of balls is empty. The next theorem, proved in the supplementary material, shows that the estimation error of this estimator $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$ is of the optimal rate, up to a logarithmic factor. To the best of our knowledge, this is the first result of this kind in the literature.

THEOREM 3. *Let $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$ be the estimator defined in Definition 8. Let $\delta \in (4e^{-n}, 1)$ be a tolerance level. Let $n \ge p \ge 2$ and $\varepsilon \le (5-\sqrt{5})a/10$, where $a \in (0,1)$ is the parameter used in Definition 8. Then, for every $\boldsymbol{\mu}^* \in \mathbb{R}^p$ and every $\mathbf{P}_n \in \text{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$, we have*

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\mu}}_n^{\text{AIR}} - \boldsymbol{\mu}^*\|_2 \le \frac{3A_5 \|\boldsymbol{\Sigma}\|_{\text{op}}^{1/2}}{a - 2\varepsilon - \sqrt{\varepsilon(a-\varepsilon)}} \left(\sqrt{\frac{p + \log(4\ell_{\max}/\delta)}{n}} + \varepsilon\sqrt{\log(1/\varepsilon)} \right) \right) \le 1 - 4\delta.$$

The breakdown point of the adaptive estimator $\widehat{\boldsymbol{\mu}}_n^{\text{AIR}}$ inferred from the last theorem is slightly smaller than the one of $\widehat{\boldsymbol{\mu}}_n^{\text{IR}}$. Indeed, there is a factor $a < 1$ between these two quantities. Note that one can choose a to be very close to one. The only drawback of choosing a too close to one is the higher computational complexity of the resulting estimator.

6.3. *Extension to unknown covariance $\boldsymbol{\Sigma}$.* The iteratively reweighted mean estimator defined in Algorithm 1 requires the knowledge of the covariance matrix $\boldsymbol{\Sigma}$. Let us discuss what happens when this matrix is unknown, by considering two qualitatively different situations. The first situation is when the covariance matrix is isotropic, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$, with unknown $\sigma > 0$. One can easily check that all the claims of Sections 3 and 5 hold true for a slight modification of $\widehat{\mu}_n^{\text{IR}}$ obtained by minimizing $\lambda_{\max}(\sum_i w_i (\boldsymbol{X}_i - \widehat{\mu}^{k-1})^{\otimes 2} - \boldsymbol{\Sigma})$ instead of $G(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}^{k-1})$ in

Algorithm 2: Iteratively reweighted mean estimator (known ε, unknown Σ)

Input: data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{R}^p$, contamination rate ε
Output: parameter estimate $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$
Initialize: compute $\widehat{\boldsymbol{\mu}}^0$ as a minimizer of $\sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2$

Set $K = 0 \vee \left\lceil \frac{\log(4p) - 2\log(\varepsilon(1-2\varepsilon))}{2\log(1-2\varepsilon) - \log\varepsilon - \log(1-\varepsilon)} \right\rceil$.

For $k = 1 : K$
Compute current weights:

$$\boldsymbol{w} \in \operatorname*{arg\,min}_{(n-n\varepsilon)\|\boldsymbol{w}\|_\infty \le 1} \lambda_{\max}\left(\sum_{i=1}^n w_i (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})^{\otimes 2} \right).$$

Update the estimator: $\widehat{\boldsymbol{\mu}}^k = \sum_{i=1}^n w_i \boldsymbol{X}_i$.
EndFor
Return $\widehat{\boldsymbol{\mu}}^K$.

(5). The advantage of considering $G(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}^{k-1})$ was computational. Indeed, if for the weights $\boldsymbol{w}$ at some iteration k in Algorithm 1 the aforementioned largest eigenvalue $\lambda_{\max}$ is nonpositive, then one can stop the iterations and output the corresponding weighted mean.
On the other hand, replacing $\lambda_{\max,+}$ by $\lambda_{\max}$, we obtain that for isotropic matrices $\boldsymbol{\Sigma}$,

$$\operatorname*{arg\,min}_{\boldsymbol{w}} \lambda_{\max}\left(\sum_i w_i (\boldsymbol{X}_i - \widehat{\mu}^{k-1})^{\otimes 2} - \sigma^2 \mathbf{I}_p \right) = \operatorname*{arg\,min}_{\boldsymbol{w}} \lambda_{\max}\left(\sum_i w_i (\boldsymbol{X}_i - \widehat{\mu}^{k-1})^{\otimes 2} \right),$$

This means that the computation of the weight $\boldsymbol{w}$ does not require the knowledge of $\boldsymbol{\Sigma}$ and the modified estimator defined in Algorithm 2 satisfies inequalities of Theorems 1 and 2.
The second situation is when $\boldsymbol{\Sigma}$ is unknown and arbitrary. In this case, to the best of our knowledge, there is no known computationally tractable estimator of $\boldsymbol{\mu}^*$ achieving a rate faster than $\sqrt{p/n} + \sqrt{\varepsilon}$. It turns out that a slightly modified version of the iteratively reweighted mean estimator defined in Algorithm 2 achieves this rate as well. The formal statement of the result entailing this claim is presented below.

THEOREM 4. *Let $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ be the iteratively reweighted mean defined in Algorithm 2. Let $\delta \in (4e^{-n}, 1)$ be a tolerance level. There exists a constant A_5 depending only on the variance proxy τ such that if $n \ge p \ge 2$ and $\varepsilon < (5 - \sqrt{5})/10$, then for every $\boldsymbol{\mu}^* \in \mathbb{R}^p$ and every $\mathbf{P}_n \in \mathrm{SGAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$, we have*

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_2 \le \frac{A_5 \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}}{1 - 2\varepsilon - \sqrt{\varepsilon(1-\varepsilon)}} \left(\sqrt{\frac{p + \log(4/\delta)}{n}} + 3\sqrt{\varepsilon} \right) \right) \le 1 - 4\delta.$$

Robust estimators of the unknown mean in the case of unknown covariance matrix have been proposed in several papers, see, for instance, [Cheng et al., 2019, Diakonikolas et al., 2016, Lai et al., 2016]. The dependence of the risk bound on ε, obtained in these papers, is of order $\sqrt{\varepsilon}$, up to possible logarithmic factors.
The proof of Theorem 4 is deferred to the supplementary material. Akin the case of known $\boldsymbol{\Sigma}$, the proof hinges on (7), which shows that the error of estimating the mean is dominated by the error of estimating $\boldsymbol{\Sigma}$. However, since we merely need an upper bound of order $\sqrt{\varepsilon}$, it suffices to show that the term $G(\boldsymbol{w}, \boldsymbol{\mu})$ remains bounded. Showing this boundedness turns out to be easier than proving that the same term is small in the case of known $\boldsymbol{\Sigma}$.

Note that Theorem 4 can be used to construct an adaptive (with respect to ε) estimator of $\boldsymbol{\mu}^*$ in the case of unknown $\boldsymbol{\Sigma}$ using Lepski's method detailed in Section 6.2. For this construction, it suffices to know an upper bound $\sigma_{\max}$ on the operator norm $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$. The resulting estimator has an error of the order $\sigma_{\max}(\sqrt{p/n}+\sqrt{\varepsilon})$.
One can also consider intermediate cases, in which the covariance matrix is not arbitrary but has a more general form than the simple isotropic one. In such a situation, it might be of interest to extend the method proposed in Algorithm 1 by using an initial estimator of $\boldsymbol{\Sigma}$ and by updating its value at each step. Indeed, when a weight vector is computed, it can be used for updating not only the mean but also the covariance matrix. The study of this estimator is left for future research.

7. Empirical results. This section showcases empirical results obtained by applying the iteratively reweighted mean estimator described in Algorithm 1 to synthetic data sets. We stress right away that there are multiple ways of solving the optimization problem involved in Algorithm 1, and the implementation we used in our experiments is not the most efficient one. As already mentioned, the aforementioned optimization problem can be seen as a semi-definite program and out-of-shelf algorithm can be applied to solve it. We implemented this approach in R using the MOSEK solver. All the results reported in this section are obtained using this implementation. We are currently working on an improved implementation using the dual sub-gradient algorithm of [Cox et al., 2014].
We applied Algorithm 1 to $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ from $\mathbf{P}_n\in \mathrm{GAC}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma},\varepsilon)$ with various types of contamination schemes. In the numerical experiments below, we illustrate (a) the evolution of the estimation error along the iterations, (b) the properties related to the theoretical breakdown point and (c) the performance of the estimator obtained from Algorithm 1 as compared to some simple estimators of the mean and to the oracle. In this section, the error of estimation is understood as the Euclidean distance between the estimated mean and its true value.
Notice that, due to the equivariance stated in Fact 2, it is sufficient to take as the true target mean vector $\boldsymbol{\mu}^*$ the zero vector $\mathbf{0}_p$ and as $\boldsymbol{\Sigma}$ any diagonal matrix with nonnegative entries. We consider the following two schemes of outlier generation:

- **Contamination by "smallest" eigenvector:** Sample n i.i.d. observations $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ drawn from $\mathcal{N}(\mathbf{0}_p,\mathbf{I}_p)$ and compute the smallest eigenvalue λ_p and the corresponding eigenvector $\boldsymbol{v}_p$ of the sample covariance matrix, defined as
$$\widehat{\boldsymbol{\Sigma}}_n=\frac{1}{n}\sum_{i=1}^n(\boldsymbol{Y}_i-\overline{\boldsymbol{Y}}_n)^{\otimes 2},$$
where $\overline{\boldsymbol{Y}}_n\stackrel{\mathrm{def}}{=}(\boldsymbol{Y}_1+\cdots+\boldsymbol{Y}_n)/n$. Choose the $n\varepsilon$ observations from $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ that have the highest (in absolute value) correlation coefficient with $\boldsymbol{v}_p$ and replace them by a vector proportional to $\boldsymbol{v}_p$ with proportionality coefficient equal to $\sqrt{p}$.
- **Uniform outliers:** Sample n observations according to this model
$$\boldsymbol{Y}_i=\boldsymbol{\theta}_i+\boldsymbol{\xi}_i,\ \text{with}\ \boldsymbol{\xi}_i\stackrel{\text{i.i.d.}}{\sim}\mathcal{N}(\mathbf{0}_p,\mathbf{I}_p)\ \text{for}\ \ i=1,\dots,n,$$
where $\boldsymbol{\theta}_i$s are all-zero vectors for $n(1-\varepsilon)$ observations $i\in\mathcal{I}=\{1,\dots,n(1-\varepsilon)\}$, while for the indices $i\notin\mathcal{I}$ we have $\|\boldsymbol{\theta}_i\|_2\neq 0$. We take the values of $\{\boldsymbol{\theta}_i\}_{i\in\mathcal{O}}$ to be i.i.d. from uniform distribution, *i.e.*, for $i\notin\mathcal{I}$ we take $\{\boldsymbol{\theta}_i^j\}_{j=1}^p\stackrel{\text{i.i.d.}}{\sim}\mathcal{U}[a,b]$. We took different values of a and b in different experiments reported below.

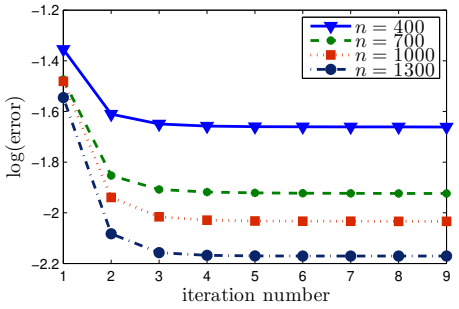

FIG 2. *The decay of the error along the iterations for $p = 9$, $\varepsilon = 0.2$ and different values of n. The contamination scheme is "uniform outliers" with $(a, b) = (0.5, 2)$.*

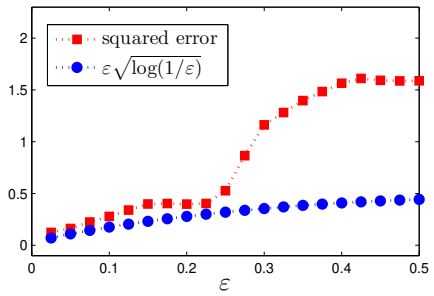

FIG 3. *The error as a function of the contamination rate ε for $n = 500$ and $p = 5$. The contamination scheme is "smallest" eigenvector. The number of iterations for $\varepsilon > \varepsilon^*$ was set to 30.*

7.1. *Improvement along iterations.* In this experiment we show the improvement of the estimation error along the iterations. The data for this experiment were generated according to the second contamination scheme mentioned above with $a = 0.5$ and $b = 2$. In Figure 2, we drew the logarithm of the estimation error (i.e., the risk between $\widehat{\boldsymbol{\mu}}_n$ and $\boldsymbol{\mu}^*$) as a function of the iteration number. The results are obtained by averaging over 50 independent repetitions. We observe that the error decreases very fast during the first iteration and remains almost constant during the rest of time. As a matter of fact, in order to speed-up the procedure, we can stop the iterations if the current weights $\boldsymbol{w}$ satisfy $\lambda_{\max}(\sum_i w_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2} - \boldsymbol{\Sigma}) \leq \varepsilon$. It is easy to check that this modified estimator still possesses all the desirable properties described in previous sections.

7.2. *Breakdown point.* The goal of this experiment is to check empirically the validity of the breakdown point $\varepsilon^* \stackrel{\text{def}}{=} (5 - \sqrt{5})/10 \approx 0.28$. To this end, we chose to contaminate the standard normal vectors by *"smallest" eigenvector* scheme. Note that if the outliers are well separated from the inliers, like in the uniform outliers scheme, then Algorithm 1 detects well these outliers by pushing their weights to 0 even when ε is large (larger than 0.28).
For $n = 500$ and $p = 5$, we conducted 50 independent repetitions of the experiment and plotted the error averaged over these 50 repetitions in Figure 3. For $\varepsilon > 0.28$, we manually set the number of iterations to[7] 30. It is interesting to observe that there is a clear change in the mean error occurring at a value close to 0.28. This empirical result allows us to conjecture that the breakdown point of the presented estimator is indeed close to 0.28.

7.3. *Comparison with other estimators.* In this last experiment we wanted to provide a visual illustration of the performance of the proposed estimator $\widehat{\boldsymbol{\mu}}_n^{\text{IR}}$ as compared to some simple competitors: the sample mean, the coordinatewise median, the geometric median and the oracle obtained by averaging all the inliers. The obtained errors, averaged over 50 independent repetitions, are depicted in Figure 4 and Figure 5. The former corresponds to $n = 500$, $p = 20$ and varying ε, while the parameters of the latter are $p = 10$, $\varepsilon = 0.1$ and $n \in [10, 1000]$. In the legend of these figures, `Estimated Mu` refers to $\widehat{\boldsymbol{\mu}}_n^{\text{IR}}$, the output of Algorithm 1, `Sample Mean` and `Sample Median` correspond respectively to the sample mean and sample coordinatewise median, `Geometric Median` refers to the estimator defined in (4) and `Oracle` refers to the sample mean of inliers only.

[7] Since the number of iteration K is well-defined for $\varepsilon \leq \varepsilon_r^* \approx 0.28$.

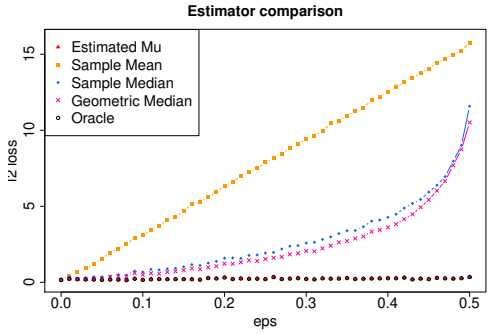

FIG 4. *ℓ_2-loss for different estimators when $n = 500, p = 20$ and the **uniform outliers** scheme with $a = 4$ and $b = 10$.*

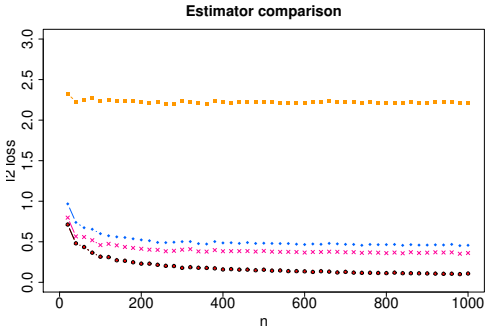

FIG 5. *ℓ_2-loss for different estimators when $p = 10$, $\varepsilon = 0.1$ and the **uniform outliers** scheme with $a = 4$ and $b = 10$.*

The plots show that the iteratively reweighted mean estimators has an error which is nearly as small as the error of the oracle. These errors are way smaller than the errors of the other estimators included in this experiment, which is in line with all the existing theoretical results.

8. Postponed proofs. We collected in this section all the technical proofs postponed from the Section 3 and Section 5. Throughout this section, we will always assume that $\lambda_{\max}(\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$. As we already mentioned above, the general case can be reduced to this one by dividing all the data vectors by $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$. The proofs in this section are presented according to the order of the appearance of the corresponding claims in the previous sections. Since the proof of Theorem 1 relies on several lemmas and propositions, we provide in Figure 6 a diagram showing the relations between these results.

8.1. *Additional details on (8).* One can check that

$$\sum_{i=1}^n w_i^*(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})^{\otimes 2} = \sum_{i=1}^n w_i^*(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}^*})^{\otimes 2} + (\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1})^{\otimes 2}$$
$$\preceq \sum_{i=1}^n w_i^*(\boldsymbol{X}_i - \boldsymbol{\mu}^*)^{\otimes 2} + \big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\big\|_2^2 \mathbf{I}_p.$$

This relation, combined with $\big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\big\|_2 \le \big\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}^*}\big\|_2 + \big\|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{k-1}\big\|_2$, yields

$$G(\boldsymbol{w}^*, \widehat{\boldsymbol{\mu}}^{k-1}) \le G(\boldsymbol{w}^*, \boldsymbol{\mu}^*) + \big(\big\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}^*}\big\|_2 + \big\|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{k-1}\big\|_2\big)^2.$$

To get the last line of (8), it suffices to apply the elementary consequence of the Minkowskii inequality $\sqrt{a + (b+c)^2} \le \sqrt{a+b^2} + c$, for every $a, b, c > 0$.

8.2. *Rough bound on the error of the geometric median.* This subsection is devoted to the proof of an error estimate of the geometric median. This estimate is rather crude, in terms of its dependence on the sample size, but it is sufficient for our purposes. As a matter of fact, it also shows that the breakdown point of the geometric median is equal to $1/2$.

LEMMA 2. *For every $\varepsilon \le 1/2$, the geometric median satisfies the inequality $\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} - \boldsymbol{\mu}^*\|_2 \le \frac{2}{n(1-2\varepsilon)} \sum_{i=1}^n \|\boldsymbol{\zeta}_i\|_2$. Furthermore, $\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} - \boldsymbol{\mu}^*\|_{\mathbb{L}_2} \le 2\mathbf{r}_{\boldsymbol{\Sigma}}^{1/2}/(1-2\varepsilon)$.*

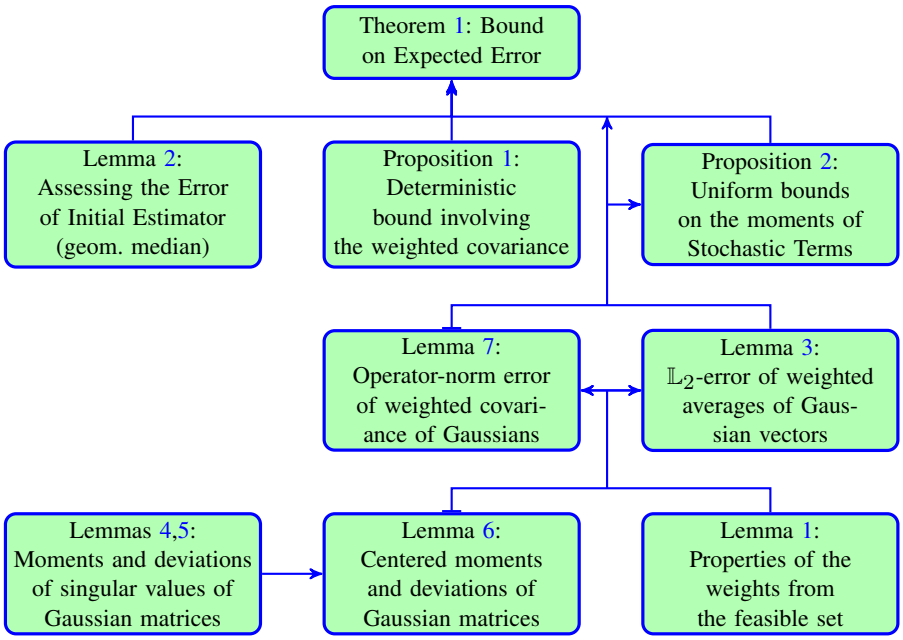

FIG 6. *A diagram showing the relation between different lemmas and propositions used in the proof of Theorem 1.*

PROOF. Without loss of generality, we assume that $\boldsymbol{\mu}^* = \mathbf{0}$. We also assume that $n\varepsilon$ is an integer. The definition of the geometric median implies that $\sum_{i=1}^n \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 \le \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}^*\|_2 = \sum_{i=1}^n \|\boldsymbol{X}_i\|_2$. Therefore, we have the simple bound

$$\begin{aligned}
n(1-\varepsilon)\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 &\le \sum_{i\in\mathcal{I}} \big(\|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 + \|\boldsymbol{\zeta}_i\|_2\big) \\
&\le \sum_{i=1}^n \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 + \sum_{i\in\mathcal{I}} \|\boldsymbol{\zeta}_i\|_2 - \sum_{i\in\mathcal{O}} \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 \\
&\le \sum_{i=1}^n \|\boldsymbol{X}_i\|_2 + \sum_{i\in\mathcal{I}} \|\boldsymbol{\zeta}_i\|_2 - \sum_{i\in\mathcal{O}} \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2 \\
&\le 2\sum_{i\in\mathcal{I}} \|\boldsymbol{\zeta}_i\|_2 + \sum_{i\in\mathcal{O}} \big(\|\boldsymbol{X}_i\|_2 - \|\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2\big) \\
&\le 2\sum_{i\in\mathcal{I}} \|\boldsymbol{\zeta}_i\|_2 + n\varepsilon\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_2,
\end{aligned}$$

where the first and the last inequalities follow from triangle inequality. We infer that

$$\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\|_{\mathbb{L}_2} \le \frac{2}{n(1-2\varepsilon)} \sum_{i=1}^n \|\boldsymbol{\zeta}_i\|_{\mathbb{L}_2} \le \frac{2}{1-2\varepsilon}\|\boldsymbol{\zeta}_1\|_{\mathbb{L}_2} \le \frac{2\mathbf{r}_{\boldsymbol{\Sigma}}^{1/2}}{1-2\varepsilon}$$

and we get the claim of the lemma. □

8.3. *Proof of Proposition 1.* To ease notation throughout this proof, we write $\bar{\varepsilon}$ instead of ε_w. Simple algebra yields

$$\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{\mathcal{I}}} = \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}}} + \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{\mathcal{I}}}$$

$$= (1-\bar\varepsilon)\boldsymbol{\mu}^* + \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_\mathcal{I}} + \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_\mathcal{I}}$$
$$= \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \bar\varepsilon\boldsymbol{\mu}^* = \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \bar\varepsilon\bar{\boldsymbol{X}}_{\boldsymbol{w}} + \bar\varepsilon(\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*).$$

This implies that $(1-\bar\varepsilon)(\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*) - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_\mathcal{I}} = \bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \bar\varepsilon\bar{\boldsymbol{X}}_{\boldsymbol{w}}$, which is equivalent to $\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}} = (\bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \bar\varepsilon\bar{\boldsymbol{X}}_{\boldsymbol{w}})/(1-\bar\varepsilon)$. Therefore, we have

$$\big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\big\|_2 = \frac{\big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}_{\mathcal{I}^c}} - \bar\varepsilon\bar{\boldsymbol{X}}_{\boldsymbol{w}}\big\|_2}{1-\bar\varepsilon} = \frac{1}{1-\bar\varepsilon}\Big\|\sum_{i\in\mathcal{I}^c} w_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})\Big\|_2$$
$$= \frac{1}{1-\bar\varepsilon}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sum_{i\in\mathcal{I}^c} w_i\,\boldsymbol{v}^\top(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}}). \tag{17}$$

Using the Cauchy-Schwarz inequality as well as the notation $\mathbf{M}_i = (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2}$ and $\mathbf{M} = \sum_{i=1}^n w_i\mathbf{M}_i$, we get

$$\Big\{\sum_{i\in\mathcal{I}^c} w_i\,\boldsymbol{v}^\top(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})\Big\}^2 \le \sum_{i\in\mathcal{I}^c} w_i \sum_{i\in\mathcal{I}^c} w_i|\boldsymbol{v}^\top(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})|^2 = \bar\varepsilon\,\boldsymbol{v}^\top\sum_{i\in\mathcal{I}^c} w_i\mathbf{M}_i\boldsymbol{v}$$
$$= \bar\varepsilon\,\boldsymbol{v}^\top\sum_{i=1}^n w_i\mathbf{M}_i\boldsymbol{v} - \boldsymbol{v}^\top\sum_{i\in\mathcal{I}} w_i\mathbf{M}_i\boldsymbol{v}$$
$$= \bar\varepsilon\,\boldsymbol{v}^\top\big(\mathbf{M}-\boldsymbol{\Sigma}\big)\boldsymbol{v} + \bar\varepsilon\boldsymbol{v}^\top\Big(\boldsymbol{\Sigma} - \sum_{i\in\mathcal{I}} w_i\mathbf{M}_i\Big)\boldsymbol{v}$$
$$\le \bar\varepsilon\,\lambda_{\max}\big(\mathbf{M}-\boldsymbol{\Sigma}\big) + \bar\varepsilon\Big\{\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v} - \boldsymbol{v}^\top\sum_{i\in\mathcal{I}} w_i\mathbf{M}_i\boldsymbol{v}\Big\}. \tag{18}$$

Finally, for any unit vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top\sum_{i\in\mathcal{I}} w_i\mathbf{M}_i\boldsymbol{v} = (1-\bar\varepsilon)\,\boldsymbol{v}^\top\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}})^{\otimes 2}\boldsymbol{v}$$
$$\ge (1-\bar\varepsilon)\,\boldsymbol{v}^\top\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\boldsymbol{w}_{|\mathcal{I}}})^{\otimes 2}\boldsymbol{v}$$
$$= (1-\bar\varepsilon)\,\boldsymbol{v}^\top\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{\zeta}_i - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}})^{\otimes 2}\boldsymbol{v}$$
$$= (1-\bar\varepsilon)\,\boldsymbol{v}^\top\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\boldsymbol{v} - (1-\bar\varepsilon)\,(\boldsymbol{v}^\top\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}})^2$$
$$\ge (1-\bar\varepsilon)\Big(\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v} - \lambda_{\max}\Big(\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big) - \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\|_2^2\Big). \tag{19}$$

Combining (18) and (19), we get

$$\Big\{\sum_{i\in\mathcal{I}^c} w_i\,\boldsymbol{v}^\top(\boldsymbol{X}_i - \boldsymbol{\mu})\Big\}^2 \le \bar\varepsilon\,\lambda_{\max}\big(\mathbf{M}-\boldsymbol{\Sigma}\big) + \bar\varepsilon^2\,\lambda_{\max}\big(\boldsymbol{\Sigma}\big)$$
$$+ \bar\varepsilon(1-\bar\varepsilon)\Big\{\lambda_{\max}\Big(\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big) + \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\|_2^2\Big\}.$$

In conjunction with (17), this yields

$$\big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^* - \bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\big\|_2 \le \frac{\sqrt{\bar\varepsilon}}{1-\bar\varepsilon}\Big(\lambda_{\max}\big(\mathbf{M}-\boldsymbol{\Sigma}\big) + \bar\varepsilon\,\lambda_{\max}\big(\boldsymbol{\Sigma}\big)$$

$$+ (1-\bar\varepsilon)\bigg\{\lambda_{\max}\bigg(\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{\Sigma}-\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\bigg)+\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\|_2^2\bigg\}\bigg)^{1/2}.$$

From this relation, using the triangle inequality and the inequality $\sqrt{a_1+\ldots+a_n}\le\sqrt{a_1}+\ldots+\sqrt{a_n}$, we get

$$\big\|\bar{\boldsymbol{X}}_{\boldsymbol{w}}-\boldsymbol{\mu}^*\|_2\le\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\big\|_2+\frac{\sqrt{\bar\varepsilon}}{1-\bar\varepsilon}\lambda_{\max,+}^{1/2}\big(\mathbf{M}-\boldsymbol{\Sigma}\big)+\frac{\bar\varepsilon\|\boldsymbol{\Sigma}\|_{\rm op}^{1/2}}{1-\bar\varepsilon}$$
$$+\sqrt{\frac{\bar\varepsilon}{1-\bar\varepsilon}}\lambda_{\max,+}^{1/2}\bigg(\sum_{i\in\mathcal{I}}(\boldsymbol{w}_{|\mathcal{I}})_i(\boldsymbol{\Sigma}-\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\bigg)+\sqrt{\frac{\bar\varepsilon}{1-\bar\varepsilon}}\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|\mathcal{I}}}\|_2$$

and, after rearranging the terms, the claim of the proposition follows.

8.4. *Proof of Lemma 1.* Claim i) of this lemma is straightforward. For ii), we use the fact that the compact convex polytope $\mathcal{W}_{n,\ell}$ is the convex hull of its extreme points. The fact that each uniform weight vector $\boldsymbol{u}^J$ is an extreme point of $\mathcal{W}_{n,\ell}$ is easy to check. Indeed, if for two points $\boldsymbol{w}$ and $\boldsymbol{w}'$ from $\mathcal{W}_{n,\ell}$ we have $\boldsymbol{u}^J=0.5(\boldsymbol{w}+\boldsymbol{w}')$, then we necessarily have $\boldsymbol{w}_{J^c}=\boldsymbol{w}'_{J^c}=0$. Therefore, for any $j\in J$,

$$\frac{1}{\ell}\ge w_j=1-\sum_{i\in J\setminus\{j\}}w_i\ge 1-\big(\ell-1\big)\times\frac{1}{\ell}=\frac{1}{\ell}.$$

This implies that $w_j=\frac{1}{\ell}\mathbb{1}(j\in J)$. Hence, $\boldsymbol{w}=\boldsymbol{u}^J$ and the same is true for $\boldsymbol{w}'$. Hence, $\boldsymbol{u}^J$ is an extreme point. Let us prove now that all the extreme points of $\mathcal{W}_{n,\ell}$ are of the form $\boldsymbol{u}^J$ with $|J|=\ell$. Let $\boldsymbol{w}\in\mathcal{W}_{n,\ell}$ be such that one of its coordinates is strictly positive and strictly smaller than $1/\ell$. Without loss of generality, we assume that the two smallest nonzero entries of $\boldsymbol{w}$ are w_1 and w_2. We have $0<w_1<1/\ell$ and $0<w_2<1/\ell$. Set $\rho=w_1\wedge w_2\wedge\{1/\ell-w_1\}\wedge\{1/\ell-w_2\}$. For $\boldsymbol{w}^+=\boldsymbol{w}+(\rho,-\rho,0,\ldots,0)$ and $\boldsymbol{w}^-=\boldsymbol{w}-(\rho,-\rho,0,\ldots,0)$, we have $\boldsymbol{w}^+,\boldsymbol{w}^-\in\mathcal{W}_{n,\ell}$ and $\boldsymbol{w}=0.5(\boldsymbol{w}^++\boldsymbol{w}^-)$. Therefore, $\boldsymbol{w}$ is not an extreme point of $\mathcal{W}_{n,\ell}$. This completes the proof of ii).
Claim ii) implies that $\mathcal{W}_{n,\ell}=\{\sum_{|J|=\ell}\alpha_J\boldsymbol{u}^J:\boldsymbol{\alpha}\in\boldsymbol{\Delta}^{K-1}\}$ with $K=\binom{n}{\ell}$. Hence,

$$\sup_{\boldsymbol{w}\in\mathcal{W}_{n,\ell}}G(\boldsymbol{w})=\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}^{K-1}}G\bigg(\sum_{|J|=\ell}\alpha_J\boldsymbol{u}^J\bigg)\le\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Delta}^{K-1}}\sum_{|J|=\ell}\alpha_JG\big(\boldsymbol{u}^J\big)\le\max_{|J|=\ell}G(\boldsymbol{u}^J)$$

and claim iii) follows.
To prove iv), we check that

$$\|\boldsymbol{w}_I\|_1=\sum_{i\in I}w_i=1-\sum_{i\notin I}w_i\ge 1-\frac{|I^c|}{\ell}\ge 1-\frac{n-\ell'}{\ell}=\frac{\ell+\ell'-n}{\ell}.$$

This readily yields $(\boldsymbol{w}_{|I})_i\le 1/(\ell+\ell'-n)$, which leads to the claim of item iv).

8.5. *Moments of suprema over $\mathcal{W}_{n,\ell}$ of weighted averages of Gaussian vectors.* We recall that $\boldsymbol{\xi}_i$'s, for $i=1,\ldots,n$ are i.i.d. Gaussian vectors with zero mean and identity covariance matrix, and $\boldsymbol{\zeta}_i=\boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi}_i$. In addition, the covariance matrix $\boldsymbol{\Sigma}$ satisfies $\|\boldsymbol{\Sigma}\|_{\rm op}=1$.

LEMMA 3. *Let $p\ge 1$, $n\ge 1$, $m\in[2,0.562n]$ and $o\in[0,m]$ be four positive integers. It holds that*

$$\mathbf{E}^{1/2}\bigg[\sup_{\substack{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}\\|I|\ge n-o}}\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2^2\bigg]\le\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\left(1+7\sqrt{\frac{m}{2n}}\right)+\frac{4.6m}{n}\sqrt{\log(^{ne}/_m)}.$$

PROOF OF LEMMA 3. We have

$$\sup_{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}} \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2 \overset{(a)}{\le} \sup_{\boldsymbol{w}\in\mathcal{W}_{n,n-m}} \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}}\|_2 \overset{(b)}{\le} \max_{|J|=n-m} \big\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{u}^J}\big\|_2,$$

where (a) follows from claim iv) of Lemma 1 and (b) is a direct consequence of claim iii) of Lemma 1. Thus, we get

$$\begin{aligned}\sup_{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}} \|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2 &\le \max_{|J|=n-m} \big\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{u}^J}\big\|_2 = \frac{1}{n-m}\max_{|J|=n-m}\Big\|\sum_{i\in J}\boldsymbol{\zeta}_i\Big\|_2 \\ &\le \frac{1}{n-m}\Big\|\sum_{i=1}^n \boldsymbol{\zeta}_i\Big\|_2 + \frac{1}{n-m}\max_{|\bar J|=m}\Big\|\sum_{i\in\bar J}\boldsymbol{\zeta}_i\Big\|_2. \qquad (20)\end{aligned}$$

On the one hand, one readily checks that

$$\mathbf{E}\bigg[\Big\|\sum_{i=1}^n \boldsymbol{\zeta}_i\Big\|_2^2\bigg] = n\mathbf{r}_{\boldsymbol{\Sigma}}. \qquad (21)$$

On the other hand, it is clear that for every $\bar J$ of cardinality m, the random variable $\big\|\sum_{i\in\bar J}\boldsymbol{\zeta}_i\big\|_2^2$ has the same distribution as $m\sum_{j=1}^p \lambda_j(\boldsymbol{\Sigma})\xi_j^2$, where $\xi_1,\dots,\xi_p$ are i.i.d. standard Gaussian. Therefore, by the union bound, for every $t\ge 0$, we have

$$\begin{aligned}\mathbf{P}\bigg(\max_{|\bar J|=m}\Big\|\sum_{i\in\bar J}\boldsymbol{\zeta}_i\Big\|_2^2 > m\big(2\mathbf{r}_{\boldsymbol{\Sigma}}+t\big)\bigg) &\le \binom{n}{m}\mathbf{P}\bigg(\Big\|\sum_{i=1}^m\boldsymbol{\zeta}_i\Big\|_2^2 > m\big(2\mathbf{r}_{\boldsymbol{\Sigma}}+t\big)\bigg) \\ &= \binom{n}{m}\mathbf{P}\bigg(\sum_{j=1}^p\lambda_j(\boldsymbol{\Sigma})\xi_j^2 > 2\mathbf{r}_{\boldsymbol{\Sigma}}+t\bigg) \le \binom{n}{m}e^{-t/3},\end{aligned}$$

where the last line follows from a well-known bound on the tails of the generalized χ^2-distribution, see for instance [Comminges and Dalalyan, 2012, Lemma 8].

Therefore, setting $Z = \frac{1}{m}\max_{|J|=m}\big\|\sum_{i\in J}\boldsymbol{\zeta}_i\big\|_2^2 - 2\mathbf{r}_{\boldsymbol{\Sigma}}$ and using the well-known identity $\mathbf{E}[Z]\le\mathbf{E}[Z_+] = \int_0^\infty \mathbf{P}(Z\ge t)\,dt$, we get

$$\begin{aligned}\mathbf{E}\bigg[\max_{|\bar J|=m}\frac{1}{m}\Big\|\sum_{i\in\bar J}\boldsymbol{\zeta}_i\Big\|_2^2\bigg] &\le 2\mathbf{r}_{\boldsymbol{\Sigma}} + \int_0^\infty 1\wedge\binom{n}{m}e^{-t/3}\,dt \\ &= 2\mathbf{r}_{\boldsymbol{\Sigma}} + 3\log\binom{n}{m} + \binom{n}{m}\int_{3\log\binom{n}{m}}^\infty e^{-t/3}\,dt \\ &\le 2\mathbf{r}_{\boldsymbol{\Sigma}} + 3m\log(ne/m) + 3 \le 2\mathbf{r}_{\boldsymbol{\Sigma}} + 3.9m\log(ne/m), \qquad (22)\end{aligned}$$

where the last two steps follow from the inequality $\log\binom{n}{m}\le m\log(ne/m)$ and the fact that $m\ge 2$, $m\log(ne/m)\ge n\inf_{x\in[2/n,1]}x(1-\log x)\ge 2(1-\log(2/n))\ge 10/3$. Combining (20), (21) and (22), we arrive at

$$\bigg(\mathbf{E}\bigg[\sup_{\substack{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}\\ |I|\ge n-o}}\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2^2\bigg]\bigg)^{1/2} \le \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\Big(1+\frac{m}{n-m}+\frac{\sqrt{2mn}}{n-m}\Big)+\frac{4.6m}{n}\sqrt{\log(ne/m)}.$$

Finally, note that for $\alpha = m/n\le 0.562$, we have

$$\frac{m}{n-m}+\frac{\sqrt{2mn}}{n-m} = \sqrt{\frac{m}{2n}}\bigg(\frac{\sqrt{2mn}}{n-m}+\frac{2n}{n-m}\bigg) = \sqrt{\frac{m}{2n}}\bigg(\frac{\sqrt{2\alpha}}{1-\alpha}+\frac{2}{1-\alpha}\bigg)\le 7\sqrt{\frac{m}{2n}}.$$

This completes the proof of the lemma. □

8.6. *Moments and deviations of singular values of Gaussian matrices.* Let $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n$ be i.i.d. random vectors drawn from $\mathcal{N}_p(0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix. We denote by $\boldsymbol{\zeta}_{1:n}$ the $p \times n$ random matrix obtained by concatenating the vectors $\boldsymbol{\zeta}_i$. Recall that $s_{\min}(\boldsymbol{\zeta}_{1:n}) = \lambda_{\min}^{1/2}(\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top)$ and $s_{\max}(\boldsymbol{\zeta}_{1:n}) = \lambda_{\max}^{1/2}(\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top)$ are the smallest and the largest singular values of the matrix $\boldsymbol{\zeta}_{1:n}$.

LEMMA 4 (Vershynin [2012], Theorem 5.32 and Corollary 5.35). *Let $\lambda_{\max}(\boldsymbol{\Sigma}) = 1$. For every $t > 0$ and for every pair of positive integers n and p, we have*

$$\mathbf{E}\big[s_{\max}(\boldsymbol{\zeta}_{1:n})\big] \le \sqrt{n} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}, \qquad \mathbf{P}\big(s_{\max}(\boldsymbol{\zeta}_{1:n}) \ge \sqrt{n} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + t\big) \le e^{-t^2/2}.$$

If, in addition, $\boldsymbol{\Sigma}$ is the identity matrix, then

$$\mathbf{E}\big[s_{\min}(\boldsymbol{\zeta}_{1:n})\big] \ge \big(\sqrt{n} - \sqrt{p}\big)_+, \qquad \mathbf{P}\big(s_{\min}(\boldsymbol{\zeta}_{1:n}) \le \sqrt{n} - \sqrt{p} - t\big) \le e^{-t^2/2}.$$

The corresponding results in [Vershynin, 2012] treat only the case of identity covariance matrix, however the proof presented therein carries with almost no change over the case of an arbitrary covariance matrix. These bounds allow us to establish the following inequalities.

LEMMA 5. *For a subset $\bar{J}$ of $\{1, \dots, n\}$, we denote by $\boldsymbol{\zeta}_{\bar{J}}$ the $p \times |\bar{J}|$ matrix obtained by concatenating the vectors $\{\boldsymbol{\zeta}_i : i \in \bar{J}\}$. Let the covariance matrix $\boldsymbol{\Sigma}$ be such that $\lambda_{\max}(\boldsymbol{\Sigma}) = 1$. For every pair of integers $n, p \ge 1$ and for every integer $m \in [1, n]$, we have*

$$\mathbf{E}\big[s_{\max}^2(\boldsymbol{\zeta}_{1:n})\big] \le \big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{n}\big)^2 + 4,$$

$$\mathbf{E}\big[(s_{\max}(\boldsymbol{\zeta}_{1:n})^2 - n)_+\big] \le 6\sqrt{n\mathbf{r}_{\boldsymbol{\Sigma}}} + 4\mathbf{r}_{\boldsymbol{\Sigma}}, \qquad \forall n \ge 8, \tag{23}$$

$$\mathbf{E}\big[(n - s_{\min}(\boldsymbol{\xi}_{1:n})^2)_+\big] \le 6\sqrt{np}, \qquad \forall n \ge 8, \tag{24}$$

$$\mathbf{E}\Big[\max_{|\bar{J}|=m} s_{\max}^2(\boldsymbol{\zeta}_{\bar{J}})\Big] \le \Big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} + \sqrt{m} + 1.81\sqrt{m\log(ne/m)}\Big)^2 + 4.$$

The proof of this lemma is presented in the supplementary material [Dalalyan and Minasyan, 2021].

LEMMA 6. *There is a constant $A_1 > 0$ such that for every pair of integers $n \ge 8$ and $p \ge 1$ and for every covariance matrix $\boldsymbol{\Sigma}$ such that $\lambda_{\max}(\boldsymbol{\Sigma}) = 1$, we have*

$$\mathbf{E}\Big[\|\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma}\|_{\mathrm{op}}\Big] \le A_1\big(\sqrt{n} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}\big)\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}, \tag{25}$$

$$\mathbf{E}\Big[\lambda_{\max,+}(\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma})\Big] \le 6\sqrt{np} + 4p, \tag{26}$$

$$\mathbf{E}\Big[\lambda_{\max,+}(n\boldsymbol{\Sigma} - \boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top)\Big] \le 6\sqrt{np}, \tag{27}$$

where the last inequality is valid under the additional assumption $p \le n$. Furthermore, there is a constant $A_2 > 0$ such that

$$\mathbf{P}\bigg(\|\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma}\|_{\mathrm{op}} - \mathbf{E}\Big[\|\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma}\|_{\mathrm{op}}\Big] \ge A_2\Big(\sqrt{tn} + \sqrt{t\mathbf{r}_{\boldsymbol{\Sigma}}} + t\Big)\bigg) \le e^{-t}, \quad \forall t \ge 1.$$

PROOF. Inequality (25) and the last claim of the lemma are respectively Theorems 4 and 5 in [Koltchinskii and Lounici, 2017]. Let us prove the two other claims. Since $\boldsymbol{\zeta}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi}_i$ where $\boldsymbol{\xi}_i$'s are i.i.d. $\mathcal{N}(0, \mathbf{I}_p)$, we have

$$\lambda_{\max,+}(\boldsymbol{\zeta}_{1:n}\boldsymbol{\zeta}_{1:n}^\top - n\boldsymbol{\Sigma}) \le \lambda_{\max,+}(\boldsymbol{\xi}_{1:n}\boldsymbol{\xi}_{1:n}^\top - n\mathbf{I}_p) = \big(s_{\max}(\boldsymbol{\xi}_{1:n})^2 - n\big)_+.$$

Inequality (26) now follows from (23) applied in the case of an identity covariance matrix so that $\mathbf{r}_{\boldsymbol{\Sigma}} = p$. Similarly, (27) follows from (24) using the same argument. □

8.7. *Moments of suprema over $\mathcal{W}_{n,\ell}$ of weighted centered Wishart matrices.* The next lemma, the proof of which is deferred to the supplementary material, provides the in-expectation bounds of the stochastic error.

LEMMA 7. *Let $p \geq 2$, $n \geq 4 \vee p$, $m \in [2, 0.6n]$ and $o \leq m$ be four integers. It holds that*

$$\mathbf{E}\left[\sup_{\substack{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}\\ |I|\geq n-o}} \lambda_{\max,+}\Big(\boldsymbol{\Sigma} - \sum_{i=1}^{n}(\boldsymbol{w}_{|I})_i\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\Big)\right] \leq 25\sqrt{p/n} + 33(m/n)\log(n/m).$$

Furthermore, for any $p \geq 1$, $n \geq 1$, $m \in [2, 0.6n]$ and $o \leq m$,

$$\mathbf{E}\left[\sup_{\substack{\boldsymbol{w}\in\mathcal{W}_{n,n-m+o}\\ |I|\geq n-o}} \Big\|\boldsymbol{\Sigma} - \sum_{i=1}^{n}(\boldsymbol{w}_{|I})_i\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\Big\|_{\mathrm{op}}\right] \leq (5.1A_1 + 2.5A_2)\sqrt{\frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}} + 7.5A_2\frac{m\log(ne/m)}{n},$$

where A_1 and A_2 are the same constants as in Lemma 6.

8.8. *Proof of Proposition 2.* Throughout this proof, $\sup_{\boldsymbol{w},I}$ stands for the supremum over all $\boldsymbol{w} \in \mathcal{W}_n(\varepsilon)$ and over all $I \subset \{1,\dots,n\}$ of cardinality larger than or equal to $n(1-\varepsilon)$. We recall that, $\Xi = \sup_{\boldsymbol{w},I} R(\boldsymbol{\zeta}, \boldsymbol{w}, I)$ where for any subset I of$\{1,\dots,n\}$,

$$R(\boldsymbol{\zeta}, \boldsymbol{w}, I) = 2\varepsilon_w + \sqrt{2\varepsilon_w}\,\lambda_{\max,+}^{1/2}\Big(\sum_{i\in I}(\boldsymbol{w}_{|I})_i(\boldsymbol{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big) + (1+\sqrt{2\varepsilon_w})\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2.$$

Furthermore, as already mentioned earlier, for every $\boldsymbol{w} \in \mathcal{W}_n(\varepsilon)$, $\varepsilon_w \leq \varepsilon/(1-\varepsilon) \leq 1.5\varepsilon$. This implies that

$$\begin{aligned}\mathbf{E}^{1/2}[\Xi^2] \leq 3\varepsilon + (1+\sqrt{3\varepsilon})\mathbf{E}^{1/2}\Big[\sup_{\boldsymbol{w},I}\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2^2\Big]\\ + \sqrt{3\varepsilon}\,\mathbf{E}^{1/2}\Big[\sup_{\boldsymbol{w},I}\lambda_{\max,+}\Big(\sum_{i\in I}(\boldsymbol{w}_{|I})_i(\boldsymbol{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big)\Big].\end{aligned} \tag{28}$$

As proved in Lemma 3 (by taking $m = 2o$ and $o = n\varepsilon$),

$$\mathbf{E}^{1/2}\Big[\sup_{\boldsymbol{w},I}\|\bar{\boldsymbol{\zeta}}_{\boldsymbol{w}_{|I}}\|_2^2\Big] \leq \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\,\big(1+7\sqrt{\varepsilon}\big) + 9.2\varepsilon\sqrt{\log(2/\varepsilon)}. \tag{29}$$

In addition, in view of the first claim of Lemma 7 (with $m = 2o$ and $o = n\varepsilon$), stated and proved in the last section, we have

$$\mathbf{E}\Big[\sup_{\boldsymbol{w},I}\lambda_{\max,+}\Big(\sum_{i\in I}(\boldsymbol{w}_{|I})_i(\boldsymbol{\Sigma} - \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top)\Big)\Big] \leq 25\sqrt{p/n} + 66\varepsilon\log(1/2\varepsilon). \tag{30}$$

Combining (28), (29) and (30), we get

$$\begin{aligned}\mathbf{E}^{1/2}[\Xi^2] &\leq 3\varepsilon + (1+\sqrt{3\varepsilon})\big(\sqrt{p/n}\,\big(1+7\sqrt{\varepsilon}\big) + 9.2\varepsilon\sqrt{\log(2/\varepsilon)}\big)\\ &\quad + \sqrt{3\varepsilon}\,\big(25\sqrt{p/n} + 66\varepsilon\log(1/2\varepsilon)\big)^{1/2}\\ &\leq \sqrt{p/n}(1+16\sqrt{\varepsilon}) + 18\varepsilon\sqrt{\log(2/\varepsilon)} + \sqrt{3\varepsilon}\,\big(25\sqrt{p/n} + 66\varepsilon\log(1/2\varepsilon)\big)^{1/2}.\end{aligned}$$

This leads to (16). To obtain (15), we repeat the same arguments but use the second claim of Lemma 7 instead of the first one.

Acknowledgments. The work of AD and AM was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and by the FAST Advance grant.

SUPPLEMENTARY MATERIAL

Proofs of some Lemmas and Theorems.
In this supplement we provide the proofs of Lemmas 5 and 7, Theorems 2-4, as well as formalize a precise lower bound for the Gaussian model with adversarial contamination.

REFERENCES

S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *COLT 2017*, pages 169–212, 2017.

A.-H. Bateni and A. S. Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Statist.*, 14(2):2653–2677, 2020.

T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.*, 43(3):1027–1059, 2015.

T. I. Cannings, Y. Fan, and R. J. Samworth. Classification with imperfect training labels. *Biometrika*, 107(2): 311–330, 2020.

M. Chen, C. Gao, and Z. Ren. A general decision theory for Huber's ϵ-contamination model. *Electron. J. Statist.*, 10(2):3752–3774, 2016.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 10 2018.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In T. M. Chan, editor, *SODA 2019, San Diego*, pages 2755–2771. SIAM, 2019.

Y. Cherapanamjeri, N. Flammarion, and P. L. Bartlett. Fast mean estimation with sub-gaussian rates. In *Proceedings of COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806, Phoenix, USA, 25–28 Jun 2019. PMLR.

G. Chinot. Erm and rerm are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electron. J. Statist.*, 14(2):3563–3605, 2020.

O. Collier and A. S. Dalalyan. Multidimensional linear functional estimation in sparse gaussian models and robust estimation of the mean. *Electron. J. Statist.*, 13(2):2830–2864, 2019.

L. Comminges and A. S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40(5):2667–2696, 2012.

L. Comminges, O. Collier, M. Ndaoud, and A. B. Tsybakov. Adaptive robust estimation in sparse vector model, 2020.

B. Cox, A. Juditsky, and A. Nemirovski. Dual subgradient algorithms for large-scale nonsmooth learning problems. *Mathematical Programming*, 148(1-2):143–180, 2014.

A. Dalalyan and P. Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *NeurIPS 32*, pages 13188–13198. 2019.

A. S. Dalalyan and A. Minasyan. Supplement to "All-in-one robust estimator of the Gaussian mean". 2021.

J. Depersin and G. Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv*, abs/1906.03058, 2019.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6): 2695–2725, 2016.

I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *FOCS 2016*, pages 655–664, 2016.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS 2017*, pages 73–84, 2017.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *SODA 2018*, pages 2683–2702, 2018.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019a.

I. Diakonikolas, D. Kane, S. Karmalkar, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *NeurIPS 2019*, pages 10688–10699, 2019b.

Y. Dong, S. B. Hopkins, and J. Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *NeurIPS 2019*, pages 6065–6075, 2019.

D. Donoho. *Breakdown properties of multivariate location estimators*. Phd thesis, Harvard University, 1982.

D. Donoho and P. J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA, 1983.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

A. Elsener and S. van de Geer. Robust low-rank matrix estimation. *Ann. Statist.*, 46(6B):3481–3509, 2018.

C. Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

F. R. Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California, Berkeley, 1968.

S. B. Hopkins. Sub-gaussian mean estimation in polynomial time. *CoRR*, abs/1809.07425, 2018.

P. J. Huber and E. M. Ronchetti. *Robust Statistics, Second Edition*. Wiley Series in Probability and Statistics. Wiley, 2009.

O. Klopp, K. Lounici, and A. B. Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, Oct 2017.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

K. A. Lai, A. B. Rao, and S. S. Vempala. Agnostic estimation of mean and covariance. In *FOCS 2016*, pages 665–674, 2016.

G. Lecué and M. Lerasle. Learning from MOM's principles: Le Cam's approach. *Stochastic Process. Appl.*, 129 (11):4385–4410, 2019.

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931, 2020.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

O. V. Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(4):682–697, 1992.

A. H. Li and J. Bradic. Boosting in the presence of outliers: adaptive classification with nonconvex loss functions. *J. Amer. Statist. Assoc.*, 113(522):660–674, 2018.

P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.*, 45(2):866–896, 2017.

H. P. Lopuhaä and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1):229–248, 1991.

G. Lugosi and S. Mendelson. Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields*, 175(3-4):957–973, 2019.

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS)*, 22(3):925–965, 2020.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.*, 49(1):393 – 410, 2021.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2006.

S. Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, 46 (6A):2871–2903, 2018.

J. Polzehl and V. G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):335–354, 2000.

P. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht, 1985.

P. Rousseeuw and M. Hubert. High-breakdown estimators of multivariate location and scatter. In *Robustness and complex data structures*, pages 49–66. Springer, Heidelberg, 2013.

P. J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880, 1984.

M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *Ann. Statist.*, 40 (4):2195–2238, 2012.

W. Stahel. *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. Phd thesis, ETH Zurich, 1981.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

B. Zhu, J. Jiao, and J. Steinhardt. When does the tukey median work? In *IEEE International Symposium on Information Theory (ISIT)*, 2020.