# Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms

Jules Depersin and Guillaume Lecué

email: jules.depersin@ensae.fr, email: guillaume.lecue@ensae.fr
CREST, ENSAE, IPParis. 5, avenue Henry Le Chatelier, 91120 Palaiseau, France.

February 2, 2021

### Abstract

We consider the problem of robust mean and location estimation w.r.t. any pseudo-norm of the form $x \in \mathbb{R}^d \to \|x\|_S = \sup_{v \in S}\langle v, x\rangle$ where $S$ is any symmetric subset of $\mathbb{R}^d$. We show that the deviation-optimal minimax subgaussian rate for confidence $1 - \delta$ is

$$\max\left(\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \sup_{v \in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{\log(1/\delta)}{N}}\right)$$

where $\ell^*(\Sigma^{1/2}S)$ is the Gaussian mean width of $\Sigma^{1/2}S$ and $\Sigma$ the covariance of the data (in the benchmark i.i.d. Gaussian case). This improves the entropic minimax lower bound from [32] and closes the gap characterized by Sudakov's inequality between the entropy and the Gaussian mean width for this problem. This shows that the right statistical complexity measure for the mean estimation problem is the Gaussian mean width. We also show that this rate can be achieved by a solution to a convex optimization problem in the adversarial and $L_2$ heavy-tailed setup by considering minimum of some Fenchel-Legendre transforms constructed using the Median-of-means principle. We finally show that this rate may also be achieved in situations where there is not even a first moment but a location parameter exists.

**AMS subject classification:** 62F35
**Keywords:** Robustness, entropy, Gaussian mean widths, heavy-tailde data, location parameter.

## 1 Introduction

We consider the problem of robust (to adversarial corruption and heavy-tailed data) multivariate mean and location estimation with respect to any pseudo-norm $\nu \in \mathbb{R}^d \to \|\nu\|_S = \sup_{\mu \in S}\langle\mu, \nu\rangle$ where $S$ is any symmetric subset of $\mathbb{R}^d$ (i.e. if $x \in S$ then $-x \in S$). This problem has been extensively studied during the last decade for $S = B_2^d$ the unit euclidean ball [36, 8, 17, 7, 13, 33, 6, 15, 30, 16, 9, 10, 12, 28, 11]. Only little is known for general symmetric sets $S$ and we will mainly refer to [32] where this problem has been handled for $S$ which is the unit dual ball $B^\circ$ of a norm $\|\cdot\|$ (so that $\|\cdot\|_S = \|\cdot\|$).

In [32], the authors introduced the problem of robust to heavy-tailed data estimation of a mean vector w.r.t. any norm. The problem can be stated as follow: given $N$ i.i.d. random vectors $X_1, \ldots, X_N$ in $\mathbb{R}^d$ with mean $\mu^*$ and covariance matrix $\Sigma$, a norm $\|\cdot\|$ on $\mathbb{R}^d$ and a confidence parameter $\delta \in (0, 1)$ find an estimator $\tilde{\mu}_N(\delta)$ and the best possible accuracy $r^*(N, \delta)$ such that with probability at least $1 - \delta, \|\tilde{\mu}_N(\delta) - \mu^*\| \leq r^*(N, \delta)$. In [32], the authors use the median-of-means principle [37, 19, 1] to construct an estimator satisfying the following result.

**Theorem 1.** *[Theorem 2 in [32]] There exist an absolute constant c such that the following holds. Given a norm $\|\cdot\|$ on $\mathbb{R}^d$ and a confidence $\delta \in (0,1)$, one can construct $\tilde{\mu}_N(\delta)$ such that with probability at least $1 - \delta$*

$$\|\tilde{\mu}_N(\delta) - \mu^*\| \leq \frac{c}{\sqrt{N}} \left( \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i (X_i - \mu^*) \right\| + \mathbb{E} \left\| \Sigma^{1/2} G \right\| + \sup_{v \in B^\circ} \left\| \Sigma^{1/2} v \right\|_2 \sqrt{\log(1/\delta)} \right)$$

*where $B^\circ$ is the unit dual ball associated with $\|\cdot\|$, $(\epsilon_i)$ are i.i.d. Rademacher variables independent of the $X_i$'s and $G \sim \mathcal{N}(0, I_d)$.*

The construction of $\tilde{\mu}_N(\delta)$ is pretty involved and it seems hard to design an algorithm out of this procedure. In particular, $\tilde{\mu}_N(\delta)$ has not been proved to be solution to a convex optimization problem. Theorem 1's main interest is thus from a theoretical point of view, while robust multivariate mean estimation can also be interesting from a practical point of view [14].

The rate obtained in Theorem 1 can be decomposed into two terms: a deviation term

$$\sup_{v \in B^\circ} \left\| \Sigma^{1/2} v \right\|_2 \sqrt{\log(1/\delta)}$$

where $\sup_{v \in B^\circ} \left\| \Sigma^{1/2} v \right\|_2$ is a weak variance term and a complexity term which is the sum of a Rademacher complexity $\mathbb{E} \left\| N^{-1/2} \sum_{i=1}^{N} \epsilon_i (X_i - \mu^*) \right\|$ and a Gaussian mean width $\mathbb{E} \left\| \Sigma^{1/2} G \right\|$. The intuition behind this rate is explained in [32], in particular, in Question 1. We will however show that this rate is not the right one and that the Gaussian mean width term is actually not necessary. Moreover, we will show that the improved rate can be achieved by an estimator solution to a convex optimization problem in Section 3 and that this holds even in the adversarial corruption model (see Assumption 1 in Section 3 below for a formal definition) and even in some situations where there is not even a first moment; in that case, $\mu^*$ is a *location* parameter and $\Sigma$ a *scatter* parameter.

The optimality of the rate in Theorem 1 has been raised in [32]. The classical approach to answer this type of question is to consider the Gaussian case that is when the data $X_i, i \in [N]$ are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$. This is also the strategy used in [32] to obtain the following deviation-minimax lower bound result[1].

**Theorem 2.** *[Theorem 3 and first paragraph in p.962 in [32]] There exists an absolute constant $c > 0$ such that the following holds. If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4)$,*

$$\mathbb{P}_{\mu^*}^N \left[ \|\hat{\mu} - \mu^*\| \leq r^* \right] \geq 1 - \delta$$

*where $\mathbb{P}_{\mu^*}^N$ is the probability distribution of $(X_i)_{i \in [N]}$ when the $X_i$ are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$ then*

$$r^* \geq \frac{c}{\sqrt{N}} \left( \sup_{\eta > 0} \eta \sqrt{\log N(\Sigma^{1/2} B^\circ, \eta B_2^d)} + \sup_{v \in B^\circ} \left\| \Sigma^{1/2} v \right\|_2 \sqrt{\log(1/\delta)} \right)$$

*where $N(\Sigma^{1/2} B^\circ, \eta B_2^d)$ is the minimal number of translated of $\eta B_2^d$ needed to cover $\Sigma^{1/2} B^\circ$.*

The term $\sup_{v \in S} \left\| \Sigma^{1/2} v \right\|_2 \sqrt{\log(1/\delta)}$ in the lower bound from Theorem 2 is obtained in [32] from Proposition 6.1 in [6] which is a deviation-minimax lower bound result holding in the one dimensional case which relies on the fact that the empirical mean is a sufficient statistics in the Gaussian shift theorem[2].

---

[1] the result from [32] is proved for $\Sigma = I_d$, it is however straightforward to extend it to the general case.

[2] The argument used in [32] goes from the one dimensional case studied in [6] to the $d$-dimensional case. It is given in a none formal way and may require some extra argument to hold. Indeed the estimator $x^*(\hat{\Psi}_N)$ in [32] is constructed using the $d$-dimensional data $X_1, \ldots, X_N$ and not one-dimensional data such as $x^*(X_1), \ldots, x^*(X_N)$. However, the result from [6] holds for estimators of a one dimensional mean using one-dimensional data and not $d$-dimensional ones. Nevertheless, Olivier Catoni showed us how to adapt the proof of Proposition 6.1 in [6] by using the sufficiency of the empirical mean in the Gaussian shift model in $\mathbb{R}^d$ to get this deviation dependent lower bound term.

The complexity term $\sup_{\eta>0} \eta \sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)}$ obtained in Theorem 2 follows from the duality theorem of metric entropy from [2] and a volumetric argument in the Gauss space similar to the one used to prove dual Sudakov's inequality in p.82-83 in [27] which has also been used to obtain minimax lower bounds based on the entropy in [24] and [34].

In general, there is a gap between the upper bound from Theorem 1 and the lower bound from Theorem 2 even in the Gaussian case. This gap is characterized by Sudakov's inequality (see Theorem 3.18 in [27] or Theorem 5.6 in [38]):

$$\sup_{\eta>0} \eta \sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)} \leq c\mathbb{E}\left\|\Sigma^{1/2}G\right\| \tag{1}$$

where $G \sim \mathcal{N}(0, I_d)$. Indeed, in the Gaussian case the complexity term of the rate obtained in Theorem 1 is the Gaussian mean width, that is the right-hand term from (1) whereas the complexity term from Theorem 2 is the entropy, that is the left-hand term in (1).

As mentioned in Remark 3 from [32], when Sudakov's inequality (1) is sharp then upper and lower bounds from Theorem 1 and 2 match in the Gaussian case (in that case the Rademacher complexity is equal to the Gaussian mean width in Theorem 1). Sharpness in Sudakov's inequality is however not a typical situation. In particular, for ellipsoids, Sudakov's bound (1) is not sharp in general and therefore the lower bound from Theorem 2 fails to recover the classical subgaussian rate for the standard Euclidean norm case (that is for $S = B_2^d$) which is given in [33] by

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

Indeed, when $\|\cdot\|$ is the $\ell_2^d$ Euclidean norm then $\mathbb{E}\left\|\Sigma^{1/2}G\right\| = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \sim \sqrt{\operatorname{Tr}(\Sigma)}$ (see, for instance, Proposition 2.5.1 in [39]). Whereas, for the entropy of $\Sigma^{1/2}B^\circ = \Sigma^{1/2}B_2^d$ w.r.t. $\eta B_2^d$, it follows from equation (5.45) in [38] that

$$\sup_{\eta>0} \eta \sqrt{\log_2 N(\Sigma^{1/2}B_2^d, \eta B_2^d)} = \sup_{n\geq 1} e_{n+1}(\Sigma^{1/2})\sqrt{n+1} \sim \sup_{n\geq 1, k\in[d]} \frac{\sqrt{n}}{2^{n/k}} \left|\prod_{j=1}^k \sqrt{\lambda_j}\right|^{1/k} \sim \sqrt{\sup_{k\in[d]} k \left|\prod_{j=1}^k \lambda_j\right|^{1/k}} \tag{3}$$

where $(e_{n+1}(\Sigma^{1/2}))_n$ are the entropy numbers of $\Sigma^{1/2} : \ell_2^d \to \ell_2^d$ (see page 62 in [38] for a definition) and $\lambda_1 \geq \ldots \geq \lambda_d$ are the singular values of $\Sigma$. In particular, when $\lambda_j = 1/j$, the entropy bound (3) is of the order of a constant whereas the Gaussian mean width is of the order of $\sqrt{\log d}$. We will fill this gap in Section 2 by showing a lower bound where the entropy is replaced by the (larger) Gaussian mean width. We will therefore obtain matching upper and lower bounds revealing that Gaussian mean width is the right way to measure the statistical complexity for the mean estimation problem w.r.t. any $\|\cdot\|_S$.

The paper is organized as follows. In the next section, we obtain the deviation-minimax optimal rate in the i.i.d. Gaussian case. In Section 3 we show that the rate from Theorem 1 can be improved and that it can be achieved by a solution to a convex program in the adversarial contamination model and in under weak or no moment assumptions. All the proofs have been gathered in Section 4.

## 2 Deviation minimax rates in the Gaussian case: benchmark subgaussian rates for the mean estimation w.r.t. $\|\cdot\|_S$

In this section, we obtain the optimal deviation-minimax rates of estimation of a mean vector $\mu^*$ when we are given $N$ i.i.d. $X_1, \ldots, X_N$ distributed like $\mathcal{N}(\mu^*, \Sigma)$ when $\Sigma \succeq 0$ is some unknown covariance matrix. In the following, $\mathbb{P}_{\mu^*}^N$ denotes the probability distribution of $(X_1, \ldots, X_N)$; it is a Gaussian measure on $\mathbb{R}^{Nd}$ with mean $((\mu^*)^\top, \ldots, (\mu^*)^\top)$ and a block $(Nd) \times (Nd)$ covariance matrix with $d \times d$ diagonal blocks given by $\Sigma$ repeated $N$ times and 0 outside of these blocks.

Unlike classical minimax results holding in expectation or with constant probability (see Chapter 2 in [40]) we want, in this section, the deviation parameter $\delta$ to appear explicitly in the minimax lower bound. Moreover, this dependency of the convergence rate with respect to $\delta$ should be of the right order given by the subgaussian $\sqrt{\log(1/\delta)}$ rate and not other polynomial dependency such as $\sqrt{1/\delta}$ as one gets for the empirical mean for $L_2$ variables (see Proposition 6.2 in [6]). This subtle behavior of the rate in terms of $\delta$ cannot be seen in expectation or constant deviation minimax lower bounds. In particular, this makes such results (like Theorem 3 or 4 below) unachievable via classical information theoretic arguments as in Chapter 2 in [40].

Fortunately, in [24], a minimax lower bound has been proved thanks to the Gaussian shift theorem which makes the deviation parameter $\delta$ appearing explicitly in the minimax lower bound. We use the same strategy here to prove our main result Theorem 3 below and its corollary Theorem 4 in the classical Euclidean $S = B_2^d$ case.

We consider the general problem of estimating $\mu^*$ w.r.t. $\|\cdot\|_S$. Let $S \subset \mathbb{R}^d$ be a symmetric set. We first obtain an upper bound result revealing the subgaussian rate. We use the empirical mean $\bar{X}_N = N^{-1} \sum_i X_i$ as an estimator of $\mu^*$. Using Borell TIS's inequality (Theorem 7.1 in [26] or pages 56-57 in [39]) we get: for all $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\left\|\bar{X}_N - \mu\right\|_S = \sup_{v \in S}\left\langle v, \bar{X}_N - \mu\right\rangle \leq \mathbb{E}\sup_{v \in S}\left\langle v, \bar{X}_N - \mu\right\rangle + \sigma_S\sqrt{2\log(1/\delta)}$$

where $\sigma_S = \sup_{v \in S}\sqrt{\mathbb{E}\left\langle v, \bar{X}_N - \mu\right\rangle^2}$ is called the weak variance. It follows that with probability at least $1 - \delta$,

$$\left\|\bar{X}_N - \mu\right\|_S \leq \frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}} + \frac{\sup_{v \in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}}{\sqrt{N}} \tag{4}$$

where $\ell^*(\Sigma^{1/2}S) = \sup\left(\left\langle G, x\right\rangle : x \in \Sigma^{1/2}S\right) = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_S$, for $G \sim \mathcal{N}(0, I_d)$, is the Gaussian mean width of the set $\Sigma^{1/2}S$. In particular, in the case where $S = B_2^d$, we recover the subgaussian rate (2) in (4). Our aim is now to show that the rate in (4) is deviation-minimax optimal. This is what is obtained in the next result.

**Theorem 3.** *Let $S$ be a symmetric subset of $\mathbb{R}^d$ such that $\mathrm{span}(S) = \mathbb{R}^d$. If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$,*

$$\mathbb{P}_{\mu^*}^N\left[\|\hat{\mu} - \mu^*\|_S \leq r^*\right] \geq 1 - \delta$$

*then*

$$r^* \geq \max\left(\frac{1}{24}\sqrt{\frac{\log 2}{\log(5/4)}}\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v \in S}\left\|\Sigma^{1/2}v\right\|_2}{12}\sqrt{\frac{\log(1/\delta)}{\sqrt{N}}}\right).$$

It follows from the upper bound (4) and the deviation-minimax lower bound from Theorem 3 that it is now possible to know exactly (up to absolute constants) the subgaussian rate for the problem of mean estimation in $\mathbb{R}^d$ w.r.t. $\|\cdot\|_S$, it is given by

$$\max\left(\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v \in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}}{\sqrt{N}}\right). \tag{5}$$

We may identify the two complexity and deviation terms in this rate. In particular, the complexity term is measured here via the Gaussian mean width of the set $\Sigma^{1/2}S$ and not its entropy as it was previously known following Theorem 2. Theorem 3 together with (4) show that the right way to measure the statistical complexity in the problem of mean estimation in $\mathbb{R}^d$ w.r.t. to any $\|\cdot\|_S$ is via the Gaussian mean width. This differs from other statistical problems such as the regression model with random design where the entropy has been proved to be the right statistical complexity in several examples [34, 24]. Following the later results in the regression model, Theorem 3 is a bit unexpected because one may though that by

taking an ERM over an epsilon net of $\mathbb{R}^d$ for the right choice of $\epsilon$ one could obtain a better rate than the one driven by the Gaussian mean width in (5); indeed, for this type of procedure, one may expect a rate depending on the (smaller) entropy instead of the (larger) Gaussian mean width. Theorem 3 shows that this is not the case: even discretized ERM cannot achieve a better rate than the one driven by the Gaussian mean width in the mean estimation problem.

An important consequence of Theorem 3 is obtained when $S = B_2^d$ that is for the problem of multivariate mean estimation w.r.t. the $\ell_2^d$-norm which is the problem that has been extensively considered during the last decade. In the following result, we recover the well-known subgaussian rate (2) showing that all the upper bound results where this rate has been proved to be achieved are actually deviation-minimax optimal and therefore could not have been improved uniformly over all $\mu^* \in \mathbb{R}^d$.

**Theorem 4.** *If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that $\mathbb{P}_{\mu^*}^N \left[ \|\hat{\mu} - \mu^*\|_2 \leq r^* \right] \geq 1 - \delta$ for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$, then*

$$r^* \geq \max \left( \frac{1}{24} \sqrt{\frac{\log 2}{2 \log(5/4)}} \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}}, \frac{1}{12} \sqrt{\frac{\|\Sigma\|_{op} \log(1/\delta)}{N}} \right).$$

Given that the empirical mean $\bar{X}_N$ is such that for all $\mu \in \mathbb{R}^d$ with $\mathbb{P}_\mu^N$-probability at least $1 - \delta$,

$$\left\| \bar{X}_N - \mu \right\|_2 \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{2 \|\Sigma\|_{op} \log(1/\delta)}{N}}$$

we conclude from Theorem 4 that the sub-gaussian rate (2) is the deviation-minimax rate of convergence for the multivariate mean estimation problem w.r.t. $\ell_2^d$ and that it is achieved by the empirical mean. In particular, there are no statistical procedure that can do better than the empirical mean uniformly over all mean vectors $\mu^* \in \mathbb{R}^d$ up to constant, this includes in particular all discretized versions of $\bar{X}_N$.

## 3 Convex programs

In this section, we introduce statistical procedures which are solutions to convex programs and which can achieve the rate from Theorem 1 without the unnecessary Gaussian mean width term $\mathbb{E} \left\| \Sigma^{1/2} G \right\|$. We also show that these procedures handle adversarial corruption and may still perform optimally in some situations where there is not even a first moment.

### 3.1 Construction of the Fenchel-Legendre minimum estimators.

**Definition 1.** *Let $S$ be a subset of $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$. The Fenchel-Legendre transform of $f$ on $S$ is the function $f_S^*$ defined for all $\mu \in \mathbb{R}^d$ by $f_S^*(\mu) = \sup_{v \in S} \left( \langle \mu, v \rangle - f(v) \right)$.*

For our purpose, the main property of a Fenchel-Legendre transform we will use is that it is a convex function as it is the maximal function of the family $\left( \mu \in \mathbb{R}^d \to \langle \mu, v \rangle - f(v) : v \in S \right)$ of linear functions.

We are now defining two examples of functions such that by taking the minimum of their Fenchel-Legendre transform over $S$ will lead to optimal estimators of $\mu^*$ w.r.t. $\|\cdot\|_S$. The construction of these two functions are based on the median-of-means principle: the dataset $\{X_1, \ldots, X_N\}$ is split into $K$ equal size blocks of data indexed by $(B_k)_k$ forming an equipartition of $[N]$. On each block, an empirical mean is constructed $\bar{X}_k = |B_k|^{-1} \sum_{i \in B_k} X_i$. The two functions we are considering are using the $K$ bucketed means $(\bar{X}_k)_k$ and are defined, for all $v \in \mathbb{R}^d$, by

$$f(v) = \frac{1}{|I_K|} \sum_{k \in I_K} \left\langle \bar{X}_k, v \right\rangle_{(k)}^* \text{ and } g(v) = \mathrm{Med}(\langle \bar{X}_k, v \rangle) = \left\langle \bar{X}_k, v \right\rangle_{\left( \frac{K+1}{2} \right)}^* \tag{6}$$

where if $a_k = \langle \bar{X}_k, v \rangle, k \in [K]$ then $\langle \bar{X}_k, v \rangle^*_{(k)}, k \in [K]$ are the rearrangement of $(a_k)_k$ such that $a^*_{(1)} \leq \ldots \leq a^*_{(K)}$ (this is the rearrangement of the values $a_k$'s themselves and not of their absolute values) and

$$I_K = \left[ \frac{K+1}{4}, \frac{3(K+1)}{4} \right] = \left\{ \frac{K+1}{2} \pm k : k = 0, 1, \cdots, \frac{K+1}{4} \right\}$$

is the inter-quartiles interval – w.l.o.g. we assume that $K + 1$ can be divided by 4. In other words, $f(v)$ is the average sum over all inter-quartile values of the vector $(\langle \bar{X}_k, v \rangle)_{k \in [K]}$ and $g(v)$ is the median of this vector. Note that both functions $f$ and $g$ are homogeneous i.e. $f(\theta v) = \theta f(v)$ and $g(\theta v) = \theta g(v)$ for every $v \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$ and in particular they are odd functions; two facts we will use later.

We are now considering the Fenchel-Legendre transform of the functions $f$ and $g$ over a symmetric set $S$:

$$f^*_S : \mu \in \mathbb{R}^d \to \sup_{v \in S} \left( \langle \mu, v \rangle - f(v) \right) \text{ and } g^*_S : \mu \in \mathbb{R}^d \to \sup_{v \in S} \left( \langle \mu, v \rangle - g(v) \right). \tag{7}$$

As mentioned previously the two functions $f^*_S$ and $g^*_S$ are convex functions. We are now using them to define convex programs whose solutions will be proved to be robust and subgaussian estimators of the mean / location vector $\mu^*$ w.r.t. $\|\cdot\|_S$:

$$\hat{\mu}^f_S \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} f^*_S(\mu) \text{ and } \hat{\mu}^g_S \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} g^*_S(\mu). \tag{8}$$

For some special choices of $S$, the Fenchel-Legendre minimization estimator $\hat{\mu}^g_S$ coincides with some classical procedures. This is for instance the case when $S = B^d_1$ (the unit ball of the $\ell^d_1$-norm) or $S = B^d_2$. Indeed, when $S = B^d_1$, $\hat{\mu}^g_S$ is the coordinate-wise Median of Means:

$$\hat{\mu}^g_S = \operatorname*{argmin}_{\mu = (\mu_j) \in \mathbb{R}^d} \max_{j \in [d]} \left| \mu_j - \operatorname{Med} \left( \langle \bar{X}_k, e_j \rangle \right) \right| = \left( \operatorname{Med} \left( \langle \bar{X}_k, e_j \rangle \right) : j \in [d] \right) \tag{9}$$

where $(e_j)^d_{j=1}$ is the canonical basis of $\mathbb{R}^d$, because $\|\cdot\|_S = \|\cdot\|_{\operatorname{conv}(S)}$ where $\operatorname{conv}(S)$ is the convex hull of $S$ and so one may just take $S = \{\pm e_j : j \in [d]\}$. It is therefore possible to derive deviation-minimax optimal bounds for the coordinate-wise Median of Means w.r.t. the $\ell^d_\infty$-norm from general upper bounds on $\hat{\mu}^g_S$ since in that case $\|\cdot\|_S = \|\cdot\|_\infty$.

In the case $S = B^d_2$ (that is for the mean/location estimation problem w.r.t. $\ell^d_2$), the Fenchel-Legendre minimum estimator $\hat{\mu}^g_S$ is a minmax MOM estimator [25]. This connection allows to write $\hat{\mu}^g_S$ (as well as $\hat{\mu}^f_S$) as a non-constraint estimator, it also shows that this minmax MOM estimator is actually solution to a convex optimization problem and how minmax MOM estimator can be generalized to other estimation risks.

Minmax MOM estimators have been introduced as a systematic way to construct robust and subgaussian estimators in [25]. They have been proved to be deviation-minimax optimal for the mean estimation problem in [30] w.r.t. $\|\cdot\|_2$. Their definition only requires to consider a loss function; here we take for all $\mu \in \mathbb{R}^d$, $\ell_\mu : x \in \mathbb{R}^d \to \|x - \mu\|^2_2$ and the minmax MOM estimator is then defined as

$$\tilde{\mu} \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} \sup_{\nu \in \mathbb{R}^d} \operatorname{Med} \left( P_{B_k}(\ell_\mu - \ell_\nu) : k \in [K] \right) \tag{10}$$

where $P_{B_k}$ is the empirical measure on the data in block $B_k$. The minmax MOM estimator $\tilde{\mu}$ was proved to achieve the subgaussian rate in (2) with confidence $1 - \delta$ when the number of blocks is $K \sim \log(1/\delta)$ and $K \gtrsim |\mathcal{O}|$ in [30].

Even though the minmax formulation of $\tilde{\mu}$ suggests a robust version of a descent/ascent gradient method over the median block (see [25, 30] for more details), no proof of convergence of this algorithm is known so far. Moreover, the main drawback of the minmax MOM estimator seems to be that it is solution of a non-convex optimization problem and may therefore be likely to be rather difficult to compute in practice. In the next result, we show that this is not the case since the minmax MOM estimator (10) is in fact equal to $\hat{\mu}^g_S$ for $S = B^d_2$ and it is therefore solution to a convex optimization problem.

**Proposition 1.** *The minmax MOM estimator $\tilde{\mu}$ defined in* (10) *satisfies $\tilde{\mu} \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} g^*_{B_2^d}(\mu)$. The minmax MOM estimator is therefore solution to a convex optimization problem.*

**Proof.** We show that $\tilde{\mu} \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} \sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v \rangle)$. We consider the quadratic/multiplier decomposition of the difference of loss functions: for all $\mu, \nu \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we have $(\ell_\mu - \ell_\nu)(x) = \|x - \mu\|_2^2 - \|x - \nu\|_2^2 = -2\langle x - \mu, \mu - \nu \rangle - \|\mu - \nu\|_2^2$. Hence, for all $\mu \in \mathbb{R}^d$, we have

$$\sup_{\nu \in \mathbb{R}^d} \operatorname{Med}\left(P_{B_k}(\ell_\mu - \ell_\nu)\right) = \sup_{\nu \in \mathbb{R}^d} \left(-2\operatorname{Med}(\langle \bar{X}_k - \mu, \mu - \nu \rangle) - \|\mu - \nu\|_2^2\right)$$

$$= \sup_{\|v\|_2=1} \sup_{\theta \geq 0} \left(2\theta \operatorname{Med}(\langle \bar{X}_k - \mu, v \rangle) - \theta^2\right) = \sup_{\|v\|_2=1} \left(\operatorname{Med}(\langle \bar{X}_k - \mu, v \rangle)\right)^2 = \left(\sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v \rangle)\right)^2.$$

We conclude since

$$\operatorname{argmin}_{\mu \in \mathbb{R}^d} \left(\sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v \rangle)\right)^2 = \operatorname{argmin}_{\mu \in \mathbb{R}^d} \sup_{\|v\|_2=1} \operatorname{Med}\left(\langle \bar{X}_k - \mu, v \rangle\right).$$

∎

It follows from Proposition 1 that the minmax MOM estimator $\tilde{\mu}$ is solution to a convex optimization problem. This fact is far from being obvious given the definition of $\tilde{\mu}$ in (10).

Proposition 1 suggests a new formulation for $\hat{\mu}^g_S$ and $\hat{\mu}^f_S$. It is indeed possible to write these estimators as regularized estimators instead of their original constraint formulation (note that the Fenchel-Legendre transforms in (7) are suprema over $S$ and are therefore constraint optimization problems). We now show that we may write them as suprema over all $\mathbb{R}^d$ if we add an ad hoc regularization function.

Let us introduce the two following functions which may be seen as regularized versions of the two $f$ and $g$ functions from (6): for all $\nu \in \mathbb{R}^d$,

$$F_S(\nu) = f(\nu) + \frac{\|\nu\|_S^2}{4} \text{ and } G_S(v) = g(\nu) + \frac{\|\nu\|_S^2}{4}. \tag{11}$$

We also consider their Fenchel-Legendre transforms over the entire set $\mathbb{R}^d$: for all $\mu \in \mathbb{R}^d$,

$$F_S^*(\mu) = \sup_{\nu \in \mathbb{R}^d} \left(\langle \mu, \nu \rangle - F_S(\nu)\right) \text{ and } G_S^*(\mu) = \sup_{\nu \in \mathbb{R}^d} \left(\langle \mu, \nu \rangle - G_S(\nu)\right).$$

The next result shows that the later two Fenchel-Legendre transforms can be used to define the two estimators $\hat{\mu}^f_S$ and $\hat{\mu}^g_S$. The proof of Proposition 2 is similar to the one of Proposition 1 where the $\ell_2$-norm is replaced by $\|\cdot\|_S$ and is therefore omitted.

**Proposition 2.** *Let $S$ be a symmetric subset of $\mathbb{R}^d$ such that $\operatorname{span}(S) = \mathbb{R}^d$. We have $\hat{\mu}^f_S \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} F_S^*(\mu)$ and $\hat{\mu}^g_S \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} G_S^*(\mu)$.*

As a consequence of Proposition 2, one can write the two estimators $\hat{\mu}^f_S$ and $\hat{\mu}^g_S$ as solutions to unconstrained minmax optimization problems like the minmax MOM estimator (10) and in particular, one may design an alternating ascent/descent sub-gradient algorithm similar to the one from [25] – we expect the one associated with $\hat{\mu}^f_S$ which uses half of the dataset at each iteration to be more efficient than the one associated with $\hat{\mu}^g_S$ which uses only the $N/K$ data in the median block at each iteration. That is the reason why we provide in Figure 1 this algorithm only for

$$\hat{\mu}^f_S \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} \sup_{\nu \in \mathbb{R}^d} \left(\langle \mu, \nu \rangle - \frac{1}{|I_K|} \sum_{k \in I_K} \langle \bar{X}_k, v \rangle^*_{(k)} - \frac{\|\nu\|_S^2}{4}\right).$$

We also recall that by the Danskin-Bertekas theorem the subgradient of $\|\cdot\|_S$ at $\nu \in \mathbb{R}^d$ when $S$ is a compact and non empty set is given by the convex hull of all $x \in S$ such that $\|\nu\|_S = \langle x, \nu \rangle$.

---

**input** : the data $X_1, \ldots, X_N$, a number $K$ of blocks, two decreasing steps size sequences $(\eta_t)_t, (\theta_t)_t \subset \mathbb{R}_+^*$ and $\epsilon > 0$ a stopping parameter

**output:** A robust estimator of the mean $\mu$

**1** Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random

**2** Construct the $K$ empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$

**3** Compute $\tilde{\mu}^{(0)}$ the coordinate-wise median-of-means and put $\mu^{(0)} = \tilde{\mu}^{(0)}$ and $\nu^{(0)} = \tilde{\mu}^{(0)}$

**4 while** $\left\| \mu^{(t)} - \mu^{(t+1)} \right\|_S \geq \epsilon$ **do**

**5** $\quad$ Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random

**6** $\quad$ Construct the $K$ empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$

**7** $\quad$ Find the inter-quartile block numbers $k_1, \ldots, k_{(K+1)/2} \in [K]$ such that

$$f(\nu^{(t)}) = \frac{1}{|I_K|} \sum_{j=1}^{(K+1)/2} \left\langle \bar{X}_{k_j}, \nu^{(t)} \right\rangle.$$

$\quad$ Construct $g^{(t)}$ a subgradient of $\|\cdot\|_S$ at $\nu^{(t)}$ and the ascent direction

$$\nabla_\nu^{(t+1)} = \mu^{(t)} - \frac{1}{|I_K|} \sum_{j=1}^{(K+1)/2} \bar{X}_{k_j} - \frac{\left\| \nu^{(t)} \right\|_S g^{(t)}}{2}.$$

$\quad$ Update $\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \nabla_\nu^{(t+1)}$.

**8** $\quad$ Make one descent step: $\mu^{(t+1)} \leftarrow \mu^{(t)} - \theta_t \nu^{(t+1)}$.

**9 end**

**10 Return** $\mu^{(t+1)}$

**Algorithm 1:** An alternating ascent/descent algorithm for the robust mean estimation problem w.r.t. $\|\cdot\|_S$ with randomly chosen blocks of data at each step.

## 3.2 The adversarial corruption model and two models for inlier.

In this section, we introduce the assumptions under which we will obtain some statistical upper bounds for the Fenchel-Legendre minimum estimators introduced above. We are considering two types of assumptions: one for the outliers which will be the adversarial corruption model and one for the inlier which will be either the existence of a second moment or a regularity assumption on a family of cdf around 0. We start with the adversarial corruption model.

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

In the adversarial contamination model from Assumption 1, the set $\mathcal{O} \subset [N]$ can depend arbitrarily on the initial data $(\tilde{X}_i)_{i=1}^N$; the corrupted data $(X_i)_{i \in \mathcal{O}}$ can have any arbitrary dependence structure; and

the informative data $(X_i)_{i \in \mathcal{I}}$ may also be correlated (for instance, it is, in general, the case when the $|\mathcal{O}|$ data $\tilde{X}_i$ with largest $\ell_2^d$-norm are modified by the adversary). The adversarial corruption model covers the Huber $\epsilon$-contamination model [18] and also the $\mathcal{O} \cup \mathcal{I}$ framework from [23, 25, 29].

Assumption 1 does not grant any property of the inlier data $(\tilde{X}_i)_{i \in [N]}$ except that they are independent. We will obtain a general result under only Assumption 1 in Section 4. However, to recover convergence rates similar to the one in Theorem 1 or the subgaussian rate in (5), we will grant some assumptions on the inlier as well. We are now considering two assumptions on the inlier which are of different nature.

The two assumptions on the inlier we are now considering are related to a subtle property of the Median-of-Means (MOM) principle which somehow benefits from its two components: the empirical median and the empirical mean. Indeed, MOM is en empirical median of empirical means and so if we refer to the classical asymptotic normality (a.n.) results of the empirical mean and the empirical median, the first one holds under the existence of a second moment and the second one holds under the assumption that the cdf is differentiable at the median with positive derivative at the median (see Corollary 21.5 in [41]). We therefore recover these two types of assumptions when we work with estimators using the MOM principle. A nice feature of MOM based estimators is that their estimation results hold under either one of the two conditions and do not require the two assumptions to hold simultaneously. We can therefore consider the two assumptions independently and get two estimation results for the Fenchel-Legendre minimum estimators introduced above (which are based on the MOM principle). We start with the moment assumption.

**Assumption 2.** *The $N$ independent random vectors $(\tilde{X}_i)_{i=1}^N$ have mean $\mu^*$ and there exists a SDP matrix $\Sigma \in \mathbb{R}^{d \times d}$ such that $\mathbb{E}(\tilde{X}_i - \mu^*)(\tilde{X}_i - \mu^*)^\top \preceq \Sigma$.*

Most of the statistical bounds obtained on MOM based estimators have focused on the heavy-tailed setup and have therefore consider Assumption 2 as their main assumption. This is the 'empirical mean component' of the MOM principle which has been the most exploited so far. It is however also possible to use the 'empirical median component' of the MOM principle to get statistical bounds even in cases where a first moment does not even exist. In that case, $\mu^*$ is called a *location parameter* and $\Sigma$ a *scale parameter*. Also, a natural assumption is similar to the one used to get the a.n. of the empirical median, that is an assumption on the cdf at the median adapted to the multidimensional and non-asymptotic setup. We are now introducing such an assumption.

**Assumption 3.** *The inlier data $(\tilde{X}_i)_{i=1}^N$ are i.i.d.. There exists $\mu^* \in \mathbb{R}^d$ and two absolute constants $c_0 > 0$ and $c_1 > 0$ such that the following holds: for all $v \in S$ and all $0 < r \leq c_0$, $H_{N,K,v}(r) \leq 1/2 - c_1 r$ where*

$$H_{N,K,v}(r) = \mathbb{P}\left[ \frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \tilde{X}_i - \mu^*, v \rangle > r \right]. \tag{12}$$

A typical example where Assumption 3 holds is when $S = \mathcal{S}_2^{d-1}$ (that is for the location estimation problem w.r.t. the Euclidean $\ell_2^d$ norm) and the $\tilde{X}_i$'s are rotational invariant that is when for all $v \in \mathcal{S}_2^{d-1}$, $\langle \tilde{X}_1 - \mu^*, v \rangle$ has the same distribution as $\langle \tilde{X}_1 - \mu^*, e_1 \rangle$ where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$. In that case, $\tilde{X}_1$ has the same distribution as $\mu^* + RU$ where $R$ is a real-valued random variable on $\mathbb{R}_+$ independent of $U$ a random vector uniformly distributed over $\mathcal{S}_2^{d-1}$. In that case and for $K = N$, for all $v \in \mathcal{S}_2^{d-1}$ and all $r \in \mathbb{R}$,

$$H_{N,K=N,v}(r) = H(r) := \mathbb{P}[R\langle U, e_1 \rangle \geq r] = \int_r^{+\infty} f(x) dx \text{ where } f : x \in \mathbb{R} \to C_d \int_{|x|}^{+\infty} \left(1 - \frac{x^2}{u^2}\right)^{\frac{d-3}{2}} d\mathbb{P}_R(u),$$

$\mathbb{P}_R$ is the probability distribution of $R$ and $C_d$ is a normalization constant which can be proved to satisfy $\sqrt{d} \leq C_d \leq 6\sqrt{d}$ (see for instance, Chapter 4 in [5]). In particular, it follows from the mean value theorem

that for all $r \geq 0$, $H(r) \leq H(0) - \min_{0 \leq x \leq r} f(x) r = 1/2 - f(r) r$. Therefore, Assumption 3 holds in that case when there exists constants $c_0, c_1 > 0$ such that $f(c_0) \geq c_1$. Furthermore, we have

$$f(c_0) \geq C_d \int_{c_0 \sqrt{d}}^{+\infty} \left( 1 - \frac{c_0^2}{u^2} \right)^{\frac{d-3}{2}} d\mathbb{P}_R(u) \geq \frac{\sqrt{d}}{2} \mathbb{P}[R \geq c_0 \sqrt{d}]$$

because $C_d \geq \sqrt{d}$ and for all $u \geq c_0 \sqrt{d}$, $(1 - (c_0/u)^2)^{(d-3)/2} \geq 1/2$. As a consequence, Assumption 3 holds if there are some constants $c_0 > 0$ and $c_2 > 0$ such that $\mathbb{P}[R \geq c_0 \sqrt{d}] \geq c_2/\sqrt{d}$. This is for instance the case, when $R$ is distributed like $\|G\|_2$ for $G \sim \mathcal{N}(0, I_d)$ (in that case $\tilde{X}_1 \sim \mathcal{N}(\mu^*, I_d)$) because $\mathbb{P}[\|G\|_2 \geq \mathbb{E}\|G\|_2 / 2] \geq 1/2$ and $\mathbb{E}\|G\|_2 \geq \sqrt{d}/2$ by Borell-TIS inequality but as well when $R$ is the positive part of a Cauchy variable because $\int_{\sqrt{d}}^{+\infty} (1/(1 + x^2)) dx \geq 1/(2\sqrt{d})$. As a consequence, Assumption 3 has nothing to do with the existence of any moment and it may hold even when there is not a first moment and even for $K = N$.

Another example where Assumption 3 holds, that we will use in the following to obtain statistical bounds for the coordinate-wise median of means for the location problem is when $S = \{\pm e_j : j \in [d]\}$ and $\tilde{X}_1 = \mu^* + Z$ where $Z = (z_j)_{j=1}^d$ is random vector in $\mathbb{R}^d$ with coordinates $z_1, \ldots, z_d$ having a symmetric around 0 Cauchy distribution. In that case, $\tilde{X}_1$ does not have a first moment and $\mu^*$ is a location parameter as the center of symmetry of the distribution of $\tilde{X}_1$. We have for all $j \in [d]$,

$$H_{N, K=N, \pm e_j}(r) = \mathbb{P}\left[ \langle \tilde{X}_1 - \mu^*, \pm e_j \rangle \geq r \right] = \mathbb{P}[z_j \geq r] = \int_r^{+\infty} \frac{dx}{\pi(1 + x^2)} \leq \frac{1}{2} - \frac{r}{\pi(1 + r^2)} \leq \frac{1}{2} - \frac{r}{2\pi}$$

for all $0 < r \leq 1$. Therefore, Assumption 3 holds in that case as well.

## 3.3 Statistical bounds for $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$

In this section, we obtain estimation bounds w.r.t. $\|\cdot\|_S$ for $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ in the adversarial contamination model with either the $L_2$ moment Assumption 1 or the regularity at 0 Assumption 3.

**Estimation properties of $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ under Assumption 1.** In this section, we obtain high probability estimation upper bounds satisfied by $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ w.r.t. $\|\cdot\|_S$ in the adversarial contamination and heavy-tailed inlier model. The rate of convergence is given by the quantity

$$r_S^* = \max \left( \frac{64}{\sqrt{N}} \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i \in [N]} \epsilon_i (\tilde{X}_i - \mu^*) \right\|_S, \sup_{v \in S} \left\| \Sigma^{1/2} v \right\|_2 \sqrt{\frac{64K}{N}} \right). \tag{13}$$

The key metric property satisfied by the two Fenchel-Legendre transforms $f_S^*$ and $g_S^*$ in the adversarial contamination and heavy-tailed inlier model is the following isomorphic result.

**Lemma 1.** *Grant Assumption 1 and Assumption 2. Let $S$ be a symmetric subset of $\mathbb{R}^d$. Assume that $|\mathcal{O}| < K/16$. With probability at least $1 - \exp(-K/512)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq g_S^*(\mu^*) \leq r_S^*$ and $|f_S^*(\mu) - \|\mu - \mu^*\|_S| \leq f_S^*(\mu) \leq r_S^*$.*

Lemma 1 shows that if $\|\mu - \mu^*\|_S \geq 2r_S^*$ then $\|\mu - \mu^*\|_S \leq g_S^*(\mu) \leq 2\|\mu - \mu^*\|_S$ and the same holds for $f_S^*$. It means that both $g_S^*$ and $f_S^*$ are two convex functions equivalent (up to absolute constants) to $\mu \to \|\mu - \mu^*\|_S$ on $\mathbb{R}^d \backslash (2r_S^*) B_S$, where $B_S$ is the unit ball associated with $\|\cdot\|_S$ and, on $(2r_S^*) B_S$, they are both smaller than $2r_S^*$. Hence, both $g_S^*(\cdot - \mu^*)$ and $f_S^*(\cdot - \mu^*)$ provide a good approximation of the metric space $(\mathbb{R}^d, \|\cdot\|_S)$. In particular, any minimum of $g_S^*$ and $f_S^*$ will be close (up to $r_S^*$) to a minimum of $\mu \to \|\mu - \mu^*\|_S$ which is $\mu^*$. This explains the statistical properties of $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$: from Lemma 1,

$$\left\| \hat{\mu}_S^f - \mu^* \right\|_S \leq f_S^*(\hat{\mu}_S^f) + f_S^*(\mu^*) \leq 2f_S^*(\mu^*) \leq 2r_S^*$$

and the same holds for $\hat{\mu}_S^g$. This leads to the following result.

**Theorem 5.** *Grant Assumption 1 and Assumption 2. Let $S$ be a symmetric subset of $\mathbb{R}^d$ and $r_S^*$ be defined in (13). For all $K > 16|\mathcal{O}|$, with probability at least $1 - \exp(-K/512)$,*

$$\left\|\hat{\mu}_S^f - \mu^*\right\|_S \leq 2r_S^* \ and \ \left\|\hat{\mu}_S^g - \mu^*\right\|_S \leq 2r_S^*.$$

The rate $r_S^*$ obtained in Theorem 5 can be split into two terms: the complexity term given by the Rademacher complexity and a deviation term exhibiting the weak variance term as in the Gaussian case. Compare with Theorem 1 from [32], this result shows that the Gaussian mean width term appearing in Theorem 1 is actually not necessary, it also shows that this improved rate can be obtained by a procedure solution to a convex program and that it can also handle adversarial corruption. When $S = B_2^d$, we recover the classical subgaussian rate because in that case the Rademacher complexity term in $r_S^*$ is less or equal to $\sqrt{\text{Tr}(\Sigma)}$ [21]. In particular, since $\hat{\mu}_S^g$ is the minmax MOM estimator in that case, we recover the main result from [30].

**Estimation properties of $\hat{\mu}_S^g$ under Assumption 3.** In this section, we consider some cases where a first moment may not exist; in that case, $\mu^*$ is a location parameter so that Assumption 3 holds. The rate of convergence we obtain in that case is given by

$$r^\diamond = \frac{C_0}{c_1}\left(\sqrt{\frac{d+1}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{|\mathcal{O}|}{c_1\sqrt{KN}} \tag{14}$$

where $c_1$ is the absolute constant from Assumption 3, $C_0$ the absolute constant from (28) and $u > 0$ a confidence parameter.

The following result is an isomorphic result satisfied by the Fenchel-Legendre transforms $g_S^*$ under Assumption 3. It is similar to the one of Lemma 1 but with the rate $r^\diamond$.

**Lemma 2.** *Let $S$ be a symmetric subset of $\mathbb{R}^d$. Grant Assumption 1 and Assumption 3 for some $K \in [N]$. Let $u > 0$. Assume that $C_0\left(\sqrt{(d+1)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \leq c_0 c_1$. With probability at least $1 - \exp(-u)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq r^\diamond$.*

As explained below Lemma 1, a result such as Lemma 2 may be used to upper bound the $\|\cdot\|_S$ distance between $\hat{\mu}_S^g$, a minimum of $g_S^*$, and $\mu^*$, a minimum of $\mu \to \|\mu - \mu^*\|_S$. This yields to the following result.

**Theorem 6.** *Let $S$ be a symmetric subset of $\mathbb{R}^d$. Grant Assumption 1 and Assumption 3 for some $K \in [N]$. Let $u > 0$ and assume that $C_0\left(\sqrt{(d+1)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \leq c_0 c_1$. With probability at least $1 - \exp(-u)$, $\left\|\hat{\mu}_S^g - \mu^*\right\|_S \leq 2r^\diamond$ where $r^\diamond$ is defined in (14).*

Unlike Theorem 5, Theorem 6 may hold even when there is not a first moment. The result from Theorem 6 hold for all $0 < u \lesssim K$ whereas Theorem 5 holds only for $u = K$ (even though one may use a Lepski's adaptive scheme to chose adaptively $K$). The price for adversarial corruption in (14) is between $|\mathcal{O}|/N$ (for $K \sim N$) and $\sqrt{|\mathcal{O}|/N}$ (for $K \sim |\mathcal{O}|$). It therefore depends on the choice of $K$ for which Assumption 3 holds. As shown after Assumption 3 for spherically symmetric random variables one can take $K = N$ and so the best possible price $|\mathcal{O}|/N$ for adversarial corruption may be achieved even when a first moment does not exist. If one needs some averaging effect so that Theorem 6 holds, then one should take $K$ as small as possible that is $K \sim |\mathcal{O}|$ and then $\sqrt{|\mathcal{O}|/N}$ will be the price for adversarial corruption as in the $L_2$ case described in Theorem 6.

**Subgaussian rates under weak or no moment assumption.** It is possible to recover (up to absolute constants) the subgaussian rate (5) in Theorem 5 for $K \sim \log(1/\delta)$ when the Rademacher complexity term from (13) and the Gaussian mean width from (5) satisfy

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu^*)\right\|_S \lesssim \ell^*\left(\Sigma^{1/2}S\right). \tag{15}$$

Such a result (i.e. Rademacher complexity is smaller than the Gaussian mean width up to constant) depends on the set $S$ and the number of moments granted on the $\tilde{X}_i$'s as well as the sample size. It obviously holds when the $\tilde{X}_i$'s are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$, so that we recover the deviation-minimax optimal subgaussian rate (5) in that case. It is also true when the $\tilde{X}_i$'s are subgaussian vectors. There are other situations under weaker moment assumption where (15) holds.

For instance, when $S = B_2^d$, (15) holds under only a $L_2$-moment assumption (see [21]). It also holds for $S = B_1^d$ when the $\tilde{X}_i$'s are isotropic with coordinates having $\log d$ subgaussian moments (i.e. $\left\|\langle \tilde{X}_i, e_j\rangle\right\|_{L_p} \leq L\sqrt{p}$ for all $1 \leq p \leq \log d$ and coordinate $j \in [d]$) and $N \gtrsim \log d$. Together with (9) and Theorem 5, this implies that the coordinate-wise MOM is a subgaussian estimator of the mean under a $\log d$ subgaussian moment assumption. Upper bounds such as (15) have been extended in [35] to general unconditional norms.

It is also possible to recover the subgaussian rate (5) in situations where there is not even a first moment thanks to Theorem 6. Indeed, for the case $S = B_1^d$ and $\tilde{X}_1 = \mu^* + Z$ where $Z = (z_j)_{j=1}^d$ has symmetric around 0 Cauchy distributed coordinates, we showed that Assumption 3 holds for $K = N$ and that $\hat{\mu}_S^g$ is the coordinate-wise median (here $K = N$) in (9). It follows from Theorem 6 that, when $d \lesssim N$ and $|\mathcal{O}| \lesssim N$ then for all $d \leq u \lesssim N$, with probability at least $1 - \exp(-u)$,

$$\left\|\hat{\mu}_S^g - \mu^*\right\|_\infty \leq 2C_0\left(\sqrt{\frac{d+1}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{2\pi|\mathcal{O}|}{N} \tag{16}$$

which is the deviation-minimiax optimal subgaussian rate (5) we would have gotten if the $\tilde{X}_i$ were i.i.d. isotropic Gaussian vectors centered in $\mu^*$ corrupted by $|\mathcal{O}|$ adversarial outliers (up to absolute constants). But here, (16) is obtained without the existence of a first moment. Moreover, in (16), the number of outliers is allowed to be proportional to $N$ and the price for adversarial corruption is of the order of $|\mathcal{O}|/N$ which is the same price we have to pay when inlier have a Gaussian distribution – this differs from the $\sqrt{|\mathcal{O}|/N}$ information theoretical lower bound that has been obtained for some non-symmetric inlier. Furthermore, the computational cost of the coordinate-wise MOM is $\mathcal{O}(Nd)$ since the cost for computing the bucketed means is $\mathcal{O}(Nd)$, the one of finding the median of $K$ numbers is $\mathcal{O}(K)$ [3], it is therefore the same computational cost as the one of the empirical mean. It is therefore possible to achieve the same computational and statistical properties as the empirical mean in a setup where a first moment does not even exist.

## Acknowledgements

## 4   Proofs

**Proof of Theorem 3.** The minimax lower bound rate $r^*$ exhibits two quantities: one which is a *complexity term* depending on the Gaussian mean width of $\Sigma^{1/2}S$ and a *deviation term* depending on $\delta$. The two terms come from two arguments. We start with the deviation term.

Let $v_1 \in \mathbb{R}^d$ be such that $\|v_1\|_S = 1$. We consider two Gaussian measures on $\mathbb{R}^{dN}$: $\mathbb{P}_0 = \mathcal{N}(0, \Sigma)^{\otimes N}$ and $\mathbb{P}_1 = \mathcal{N}(3r^*v_1, \Sigma)^{\otimes N}$. They are the distributions of a sample of $N$ i.i.d. Gaussian vectors in $\mathbb{R}^d$ with the same covariance matrix $\Sigma$ and the first one with mean 0 and the second one with mean $3r^*v_1$. We set $A_0 = (\hat{\mu})^{-1}(B_S(0, r^*)) = \{(x_1, \ldots, x_N) \in \mathbb{R}^{Nd} : \|\hat{\mu}(x_1, \ldots, x_N)\|_S \leq r^*\}$ and $A_1 = (\hat{\mu})^{-1}(B_S(3r^*v_1, r^*))$. It follows from the statistical properties of $\hat{\mu}$ that $\mathbb{P}_0[A_0] \geq 1 - \delta$ and $\mathbb{P}_1[A_1] \geq 1 - \delta$.

The key ingredient for the deviation lower bound term is a slightly generalization of Lemma 3.3 in [24] which is based on a version of the Gaussian shift Theorem from [31].

**Lemma 3.** *Let $t \mapsto \Phi(t) = \mathbb{P}(g \leq t)$ be the cumulative distribution function of a standard gaussian random variable on $\mathbb{R}$. Let $\Sigma_0 \succeq 0$ be in $\mathbb{R}^{(Nd) \times (Nd)}$ and $u, v \in \mathbb{R}^{dN}$. Let two gaussian measures $\nu_u \sim \mathcal{N}(u, \Sigma_0)$ and $\nu_v \sim \mathcal{N}(v, \Sigma_0)$ on $\mathbb{R}^{Nd}$. If $A \subset \mathbb{R}^{dN}$ is measurable, then*

$$\nu_v(A) \geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_u(A)) + \|\Sigma_0^{-1/2}(u - v)\|_2\big) \tag{17}$$

*where $\Sigma_0^{-1/2}$ is the square root of the pseudo-inverse of $\Sigma_0$.*

**Proof of Lemma 3.** When $\Sigma_0 = I_{Nd}$, Lemma 3 is exactly Lemma 3.3 in [24] for $\sigma = 1$. To prove Lemma 3, we observe that $\nu_v(A) = \mathbb{P}[G + \Sigma_0^{-1/2}v \in B]$ where $B = \Sigma_0^{-1/2}A$ and $G$ is a standard Gaussian variable in $\mathrm{Im}(\Sigma_0)$. Hence, it follows from Lemma 3.3 in [24] that

$$\mathbb{P}[G + \Sigma_0^{-1/2}v \in B] \geq 1 - \Phi\big(\Phi^{-1}(1 - \mathbb{P}[G + \Sigma_0^{-1/2}u \in B]) + \|\Sigma_0^{-1/2}(u - v)\|_{\ell_2^N}\big)$$

which is exactly (17). ∎

It follows from Lemma 3 that

$$\mathbb{P}_1[A_0] \geq 1 - \Phi\left[\Phi^{-1}(1 - \mathbb{P}_0[A_0]) + \left\|\Sigma_0^{-1/2}(0 - (3r^*v_1, \ldots, 3r^*v_1))\right\|_2\right]. \tag{18}$$

Moreover, we have $\Phi^{-1}(1 - \mathbb{P}_0[A_0]) \leq \Phi^{-1}(\delta)$ (because $1 - \mathbb{P}_0[A_0] \leq \delta$) and

$$\left\|\Sigma_0^{-1/2}(0 - (3r^*v_1, \ldots, 3r^*v_1))\right\|_2 = 3r^*\sqrt{N}\left\|\Sigma^{-1/2}v_1\right\|_2. \tag{19}$$

As a consequence, if $3r^*\sqrt{N}\left\|\Sigma^{-1/2}v_1\right\|_2 \leq -\Phi^{-1}(\delta)$ then, in (18), we get $\mathbb{P}_1[A_0] \geq 1 - \Phi[0] \geq 1/2$ which is not possible because $\mathbb{P}_1[A_1] \geq 1 - \delta > 3/4$ and $A_1 \cap A_0 = \emptyset$. As a consequence, we necessarily have $3r^*\sqrt{N} \geq (-\Phi^{-1}(\delta))\left\|\Sigma^{-1/2}v_1\right\|_2^{-1}$. The later holds for any $v_1 \in \mathbb{R}^d$ such that $\|v_1\|_S = 1$ hence $3r^*\sqrt{N} \geq (-\Phi^{-1}(\delta))[1/\inf_{\|v\|_S=1}\left\|\Sigma^{-1/2}v\right\|_2]$. It also follows from the bound on the Mill's ratio from [22] (here we use that for all $x \geq 0$, $\Phi(-x) \geq 2\varphi(x)/\sqrt{4 + x^2} + x$ where $\varphi$ is the standard Gaussian density function) that for all $0 < \delta < 1/4$, $-\Phi^{-1}(\delta) \geq 1/4\sqrt{\log(1/\delta)}$. This shows that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log(1/\delta)}{N}}\frac{1}{\inf_{\|v\|_S=1}\left\|\Sigma^{-1/2}v\right\|_2}. \tag{20}$$

To conclude on the deviation term, we use the following duality argument.

**Lemma 4.** *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric and invertible matrix. Let $\|\cdot\|$ be a norm and its dual norm $\|\cdot\|^*$ on $\mathbb{R}^d$. Let $S$ be a symmetric subset of $\mathbb{R}^d$ such that $\mathrm{span}(S) = \mathbb{R}^d$. We have*

$$\frac{1}{\inf_{\|v\|_S=1}\|A^{-1}v\|} \geq \sup_{w \in S}\|Aw\|^*.$$

**Proof of Lemma 4.** Let $v$ be such that $\|v\|_S = 1$ and $w \in S$. We have $|\langle v, w \rangle| \leq 1$ and so $|\langle A^{-1}v/\|A^{-1}v\|, Aw \rangle| \leq 1/\|A^{-1}v\|$. The later holds for all $v$ such that $\|v\|_S = 1$ and $\{A^{-1}v/\|A^{-1}v\| : \|v\|_S = 1\}$ is the unit sphere of $\|\cdot\|$. Hence, we conclude by taking the sup over $v$ such that $\|v\|_S = 1$ and $w \in S$. ∎

It follows from (20) and Lemma 4 for $\|\cdot\| = \|\cdot\|_2$ and $A = \Sigma^{1/2}$ that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log(1/\delta)}{N}}\sup_{w \in S}\left\|\Sigma^{1/2}w\right\|_2. \tag{21}$$

Let us now turn to the second part of the lower bound; the one coming from the complexity of the problem (here, it is the Gaussian mean width of $\Sigma^{1/2}S$). We know that $\hat{\mu}$ is an estimator such that for all $\mu \in \mathbb{R}^d$, $\mathbb{P}_\mu^N \left[ \|\hat{\mu} - \mu\|_S \leq r^* \right] \geq 1 - \delta$ which is equivalent to say that

$$\delta \geq \sup_{\mu \in \mathbb{R}^d} \mathbb{E}_\mu^N \phi \left( \frac{\|\hat{\mu} - \mu\|_S}{r^*} \right) \tag{22}$$

where we set $\phi : t \in \mathbb{R} \to I(t > 1)$ and $\mathbb{E}_\mu^N$ is the expectation with respect to $X_1, \ldots X_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$. Next, we consider a Gaussian distribution $\gamma$ over the set of parameters $\mu \in \mathbb{R}^d$: for $s > 0$, we assume that $\mu \sim \mathcal{N}(0, s\Sigma)$. It follows from (22) that

$$\delta \geq \int_{\mu \in \mathbb{R}^d} \mathbb{E}_\mu^N \phi \left( \frac{\|\hat{\mu} - \mu\|_S}{r^*} \right) \gamma(\mu) d\mu = \mathbb{E} \left[ \mathbb{E} \left[ \phi \left( \frac{\|\hat{\mu}(X_1, \ldots, X_N) - \mu\|_S}{r^*} \right) | X_1, \ldots, X_N \right] \right]. \tag{23}$$

In other words, we lower bound the minmax risk by a Bayesian risk. We now use Anderson's lemma to lower bound the Bayesian risk appearing in (23). We first recall Anderson's Lemma.

**Theorem 7** (Anderson's Lemma). *Let $\Gamma$ be a semi-definite $d \times d$ matrix and $Z \sim \mathcal{N}(0, \Gamma)$. Let $w : \mathbb{R}^d \to \mathbb{R}$ be such that all its level sets (i.e. $\{x \in \mathbb{R}^d : w(x) \leq c\}$ for $c \in \mathbb{R}$) are convex and symmetric around the origin. Then for all $x \in \mathbb{R}^d$, $\mathbb{E}w(Z + x) \geq \mathbb{E}w(Z)$.*

We remark that $\mu - \mathbb{E}[\mu|X_1, \ldots, X_N]$ is distributed according to $\mathcal{N}(0, (s/(1 + Ns)\Sigma))$ conditionally to $X_1, \ldots, X_N$. Therefore, applying Anderson's Lemma conditionally to $X_1, \ldots, X_N$, we obtain in (23) that

$$\delta \geq \mathbb{E} \left[ \phi \left( \frac{\|\mathbb{E}[\mu|X_1, \ldots, X_N] - \mu\|_S}{r^*} \right) \right] = \mathbb{P} \left[ \left\| \Sigma^{1/2}G \right\|_S \geq \sqrt{\frac{1 + Ns}{s}} r^* \right]$$

where $G \sim \mathcal{N}(0, I_d)$. This result is true for all $s > 0$ so taking $s \uparrow +\infty$, we obtain

$$\delta \geq \mathbb{P} \left[ \left\| \Sigma^{1/2}G \right\|_S \geq \sqrt{N} r^* \right].$$

Using Borell-TIS's inequality (Theorem 7.1 in [26] or pages 56-57 in [39]), we know that with probability at least $4/5$, $\left\| \Sigma^{1/2}G \right\|_S \geq \mathbb{E} \left\| \Sigma^{1/2}G \right\|_S - \sigma_S \sqrt{2 \log(5/4)}$ where we set $\sigma_S = \sup_{\|v\|_S=1} \left\| \Sigma^{1/2}v \right\|_2$. As a consequence, for $\delta = 1/4$, we necessarily have $\sqrt{N} r^* \geq \mathbb{E} \left\| \Sigma^{1/2}G \right\|_S - \sigma_S \sqrt{2 \log(5/4)}$ and so $\sqrt{N} r^* \geq (1/2)\mathbb{E} \left\| \Sigma^{1/2}G \right\|_S$ when $\mathbb{E} \left\| \Sigma^{1/2}G \right\|_S \geq 2\sigma_S \sqrt{2 \log(5/4)}$. Finally, when $\mathbb{E} \left\| \Sigma^{1/2}G \right\|_S < 2\sigma_S \sqrt{2 \log(5/4)}$, we know from (21) for $\delta = 1/4$ that

$$r^* \geq \frac{1}{12} \sqrt{\frac{\log 4}{N}} \sigma_S \geq \frac{1}{24} \sqrt{\frac{\log 2}{\log(5/4)}} \frac{\mathbb{E} \left\| \Sigma^{1/2}G \right\|_S}{\sqrt{N}}.$$

$\blacksquare$

**Proof of Theorem 4.** Theorem 4 follows from Theorem 3 and the following lower bound on $\mathbb{E} \left\| \Sigma^{1/2}G \right\|_{B_2^d}$. We have from Borell-TIS's inequality that

$$\mathbb{E} \left\| \Sigma^{1/2}G \right\|_2^2 - \left( \mathbb{E} \left\| \Sigma^{1/2}G \right\|_2 \right)^2 = \mathbb{E} \left( \left\| \Sigma^{1/2}G \right\|_2 - \mathbb{E} \left\| \Sigma^{1/2}G \right\|_2 \right)^2$$

$$= \int_0^\infty \mathbb{P} \left[ \left| \left\| \Sigma^{1/2}G \right\|_2 - \mathbb{E} \left\| \Sigma^{1/2}G \right\|_2 \right| \geq \sqrt{t} \right] dt \leq 2\sigma_{B_2^d}^2$$

where $\sigma_{B_2^d}^2 = \sup_{\|v\|_2=1} \left\| \Sigma^{1/2}v \right\|_2^2 = \|\Sigma\|_{op}$. Since $\mathbb{E} \left\| \Sigma^{1/2}G \right\|_2^2 = \text{Tr}(\Sigma)$, we have $\left( \mathbb{E} \left\| \Sigma^{1/2}G \right\|_2 \right)^2 \geq \text{Tr}(\Sigma) - 2 \|\Sigma\|_{op}$. Therefore, $\mathbb{E} \left\| \Sigma^{1/2}G \right\|_2 \geq \sqrt{\text{Tr}(\Sigma)/2}$ when $\text{Tr}(\Sigma) \geq 4 \|\Sigma\|_{op}$ and when $\text{Tr}(\Sigma) < 4 \|\Sigma\|_{op}$, we use the lower bound from (21) and an argument similar to the one appearing in the end of the proof of Theorem 3 to get the result. $\blacksquare$

**Proof of Lemma 1.** We first prove the result for the $g_S^*$ function. The one for the $f_S^*$ is similar up to constants and will be sketched after. The proof of Lemma 1 for the $g_S^*$ function is a corollary of the general fact which holds under only Assumption 1. Let $u > 0$ be a confidence parameter and define $R_S^*$ such that

$$\frac{4}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu)\right\|_S + \sqrt{\frac{2u}{K}} + \sup_{v\in S}H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}. \tag{24}$$

Let us show that with large probability for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq R_S^*$.

We have for all $\mu \in \mathbb{R}^d$,

$$|g_S^*(\mu) - \|\mu - \mu^*\|_S| = \left|\sup_{v\in S}\left(\langle\mu, v\rangle - g(v)\right) - \sup_{v\in S}\langle v, \mu - \mu^*\rangle\right| \leq \sup_{v\in S}\left|\langle\mu^*, v\rangle - g(v)\right| = g_S^*(\mu^*) \tag{25}$$

where we used that $S$ is symmetric and $g$ is odd. It only remains to show that $g_S^*(\mu^*) \leq R_S^*$ with large probability. To that end, it is enough to prove that, with large probability, for all $v \in S$,

$$\sum_{k\in[K]}I(\langle\bar{X}_k - \mu^*, v\rangle > R_S^*) < \frac{K}{2}. \tag{26}$$

We use the notation introduced in Assumption 1 and we consider $\overline{\tilde{X}}_k = |B_k|^{-1}\sum_{i\in B_k}\tilde{X}_i$ for $k \in [K]$ which are the $K$ bucketed means constructed on the $N$ independent vectors $\tilde{X}_i, i \in [N]$ before contamination (whereas $\bar{X}_k$ are the ones constructed after contamination). We also set $\mathcal{K} = \{k \in [K] : B_k \cap \mathcal{O} = \emptyset\}$ the indices of the non corrupted blocks. We have

$$\sum_{k\in[K]}I(\langle\bar{X}_k - \mu^*, v\rangle > R_S^*) = \sum_{k\in\mathcal{K}}I(\langle\bar{X}_k - \mu^*, v\rangle > R_S^*) + \sum_{k\notin\mathcal{K}}I(\langle\bar{X}_k - \mu^*, v\rangle > R_S^*)$$
$$\leq \sum_{k\in[K]}I(\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*) + |\mathcal{O}|. \tag{27}$$

It only remains to show that with probability at least $1 - \exp(-u)$, for all $v \in S$,

$$\sum_{k\in[K]}I(\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*) \leq \frac{4K}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu^*)\right\|_S + \sqrt{2uK} + K\sup_{v\in S}H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right).$$

We define $\phi(t) = 0$ if $t \leq 1/2$, $\phi(t) = 2(t - 1/2)$ if $1/2 \leq t \leq 1$ and $\phi(t) = 1$ if $t \geq 1$. We have $I(t \geq 1) \leq \phi(t) \leq I(t \geq 1/2)$ for all $t \in \mathbb{R}$ and so

$$\sum_{k\in[K]}I(\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*)$$
$$\leq \sum_{k\in[K]}I(\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*) - \mathbb{P}[\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*/2] + \mathbb{P}[\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*/2]$$
$$\leq \sum_{k\in[K]}\phi\left(\frac{\langle\overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right) + \mathbb{P}[\langle\overline{\tilde{X}}_k - \mu^*, v\rangle > R_S^*/2]$$
$$\leq \sup_{v\in S}\left(\sum_{k\in[K]}\phi\left(\frac{\langle\overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right)\right) + K\sup_{v\in S}H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right).$$

Next, we use several tools from empirical process theory and in particular, for a symmetrization argument, we consider a family of $N$ independent Rademacher variables $(\epsilon_i)_{i=1}^N$ independent of the $(\tilde{X}_i)_{i=1}^N$. In

15

*(bdi)* below, we use the bounded difference inequality (Theorem 6.2 in [4]). In *(sa-cp)*, we use the symmetrization argument and the contraction principle (Chapter 4 in [27]) – we refer to the supplementary material of [29] for more details. We have, with probability at least $1 - \exp(-u)$,

$$
\sup_{v \in S} \left( \sum_{k \in [K]} \phi \left( \frac{\langle \overline{\tilde{X}}_k - \mu^*, v \rangle}{R_S^*} \right) - \mathbb{E}\phi \left( \frac{\langle \overline{\tilde{X}}_k - \mu^*, v \rangle}{R_S^*} \right) \right)
$$

$$
\overset{(bdi)}{\leq} \mathbb{E} \sup_{v \in S} \left( \sum_{k \in [K]} \phi \left( \frac{\langle \overline{\tilde{X}}_k - \mu^*, v \rangle}{R_S^*} \right) - \mathbb{E}\phi \left( \frac{\langle \overline{\tilde{X}}_k - \mu^*, v \rangle}{R_S^*} \right) \right) + \sqrt{2uK}
$$

$$
\overset{(sa-cp)}{\leq} \frac{4K}{NR_S^*} \mathbb{E} \sup_{v \in S} \left\langle v, \sum_{i \in [N]} \epsilon_i (\tilde{X}_i - \mu^*) \right\rangle + \sqrt{2uK}
$$

$$
= \frac{4K}{\sqrt{N} R_S^*} \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i \in [N]} \epsilon_i (\tilde{X}_i - \mu^*) \right\|_S + \sqrt{2uK}.
$$

We therefore showed that under Assumption 1, with probability at least $1 - \exp(-u)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq R_S^*$.

Now, if Assumption 2 holds then for all $v \in S$, we have from Markov's inequality that

$$
H_{N,K,v}\left( \frac{R_S^*}{2} \sqrt{\frac{N}{K}} \right) \leq \frac{\mathbb{E}\langle \overline{\tilde{X}}_k - \mu, v \rangle^2}{(r_S^*/2)^2} = \frac{4K v^\top \Sigma v}{N(r_S^*)^2} \leq \frac{4K \sup_{v \in S} \left\| \Sigma^{1/2} v \right\|_2^2}{N(r_S^*)^2} \leq \frac{1}{8}
$$

and therefore (24) holds for $R_S^* = r_S^*$ when $|\mathcal{O}| < K/8$ and $u = K/128$. This proves the result of Lemma 1 for $g_S^*$ under Assumption 2.

Finally, for the function $f_S^*$ one needs to control the average of the $K/2$ inter-quartiles. One way to do it is to control the value of all elements $\langle \bar{X}_k - \mu^*, v \rangle$ in the inter-quartiles interval. This can be done by defining an $R_S^*$ similar to the one in (24) but where the right-hand side value $1/2$ is replaced by $1/4$ in (24). This only modifies the absolute constants which are the one used in Lemma 1. ∎

**Proof of Lemma 2.** Unlike in Lemma 1 where we used the Rademacher complexities as a complexity measure, in this proof, the complexity measure we are using is the Vapnik and Chervonenkis (VC) dimension [43, 44] of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0, 1\}$ in our case. We recall that the Vapnik and Chervonenkis dimension of $\mathcal{F}$, denoted by $VC(\mathcal{F})$, is the maximal integer $n$ such that there exists $x_1, \ldots, x_n \in \mathbb{R}^d$ for which the set $\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ is of maximal cardinality, that is of size $2^n$. The VC dimension of the set of all indicators of half affine spaces in $\mathbb{R}^d$ is $d + 1$ (see Example 2.6.1 in [42]). We also know (see, for instance, Chapter 3 in [20]) the following concentration bound: let $Y_1, \ldots, Y_n$ be independent random vectors in $\mathbb{R}^d$, there exists an absolute constant $C_0$ such that for all $u > 0$, with probability at least $1 - \exp(-u)$,

$$
\sup_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^{n} f(Y_i) - \mathbb{E}f(Y_i) \right) \leq C_0 \left( \sqrt{\frac{VC(\mathcal{F})}{n}} + \sqrt{\frac{u}{n}} \right). \tag{28}
$$

Lemma 2 is a corollary of a general result which holds under the only Assumption 1. This general result says that for all $u > 0$, with probability at least $1 - \exp(-u)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq R^\diamond$ where $R^\diamond$ is any point such that

$$
C_0 \left( \sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \sup_{\|v\|_2 = 1} H_{N,K,v}\left( R^\diamond \sqrt{\frac{N}{K}} \right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{29}
$$

where $C_0$ is the constant from (28). In particular, when Assumption 3 holds then one can check that (29) holds for $R^\diamond = r^\diamond$ when $r^\diamond \leq c_0$ proving the result of Lemma 2. It only remains to show the general result. To that end we follow the same strategy as in the proof of Lemma 1 up to (27) (and with $R_S^*$ replaced by $R^\diamond$). From that point, we use (28) and the VC dimension of the set of affine half spaces to get that with probability at least $1 - \exp(-u)$, for all $v \in S$,

$$\sum_{k \in [K]} I(\langle \overline{\tilde{X}}_k - \mu^*, v \rangle > R^\diamond) \leq H_{N,K,v}\left(R^\diamond \sqrt{\frac{N}{K}}\right) + C_0\left(\sqrt{\frac{d+1}{N/K}} + \sqrt{\frac{u}{N/K}}\right)$$

and so by definition of $R^\diamond$, on the same event, for all $v \in S$, $\sum_{k \in [K]} I(\langle \bar{X}_k - \mu^*, v \rangle > R^\diamond) < 1/2$. This concludes the proof. ∎

# References

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

[2] S. Artstein, V. Milman, and S. J. Szarek. Duality of metric entropy. *Ann. of Math. (2)*, 159(3):1313–1328, 2004.

[3] Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ronald L. Rivest, and Robert Endre Tarjan. Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461, 1973.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[5] Wł odzimierz Bryc. *The normal distribution*, volume 100 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995. Characterizations with applications.

[6] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[7] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[8] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

[9] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018.

[10] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates, 2019.

[11] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *arXiv preprint arXiv:2002.01432*, 2020.

[12] Jules Depersin and Guillaume Lecué. Fast algorithms for robust estimation of a mean vector. 2019.

[13] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 12 2016.

[14] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*, 2017.

[15] Matthew J Holland. Distribution-robust mean estimation via smoothed random perturbations. *arXiv preprint arXiv:1906.10300*, 2019.

[16] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.*, 48(2):1193–1213, 2020.

[17] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17:Paper No. 18, 40, 2016.

[18] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[19] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[20] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Berlin, 2011.

[21] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[22] Yûsaku Komatu. Elementary inequalities for Mills' ratio. *Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs.*, 4:69–70, 1955.

[23] Guillaume Lecué and Matthieu Lerasle. Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410, 2019.

[24] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

[25] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.*, 48(2):906–931, 04 2020.

[26] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[27] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[28] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612, 2020.

[29] Matthieu Lerasle, Zoltan Szabo, Tilmothée Mathieu, and Guillaume Lecué. Monk – outliers-robust mean embedding estimation by median-of-means. Technical report, CNRS, University of Paris 11, Ecole Polytechnique and CREST, 2017.

[30] Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, and Guillaume Lecué. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pages 3782–3793, 2019.

[31] Wenbo V. Li and James Kuelbs. Some shift inequalities for Gaussian measures. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 233–243. Birkhäuser, Basel, 1998.

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

[33] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[34] Shahar Mendelson. "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *Ann. Statist.*, 45(5):1835–1862, 2017.

[35] Shahar Mendelson. On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 301–318. Springer, Cham, 2017.

[36] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

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

[38] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[39] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems.

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

[41] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[42] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[43] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham, 2015. Reprint of Theor. Probability Appl. **16** (1971), 264–280.

[44] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.