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

Depersin, J.D.P.; Lecué, G.

DOI
10.1007/s00440-022-01127-y

Publication date
2022

Document Version
Final published version

Published in
Probability Theory and Related Fields

Probability Theory and Related Fields (2022) 183:997–1025
https://doi.org/10.1007/s00440-022-01127-y

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

Jules Depersin[1] · **Guillaume Lecué**[1]

Received: 2 February 2021 / Revised: 3 February 2022 / Accepted: 11 March 2022 /
Published online: 3 July 2022

Abstract
We consider the problem of robust mean and location estimation with respect to any pseudo-norm of the form $x \in \mathbb{R}^d \mapsto \|x\|_S = \sup_{v\in S}\langle v, x\rangle$ where S is any symmetric subset of $\mathbb{R}^d$. We show that the deviation-optimal minimax sub-Gaussian 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 Σ the covariance of the data. This improves the entropic minimax lower bound from Lugosi and Mendelson (Probab Theory Relat Fields 175(3–4):957–973, 2019) 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.

Keywords Robustness · Entropy · Gaussian mean widths · Heavy-tailed data · Location parameter · Median-of-means · Fenchel–Legendre transform

Mathematics Subject Classification 62F35

✉ Guillaume Lecué
guillaume.lecue@ensae.fr

Jules Depersin
jules.depersin@ensae.fr

[1] CREST, ENSAE, IP Paris, 5, avenue Henry Le Chatelier, 91120 Palaiseau, France

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 $\nu \in \mathbb{R}^d \mapsto \|\nu\|_S = \sup_{\mu \in S} \langle \mu, \nu \rangle$ 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 [8–12, 14–16, 18–20, 32, 33, 37, 43, 44]. Only little is known for general symmetric sets S and we will mainly refer to [36] where this problem has been handled for S which is the unit dual ball B° of a norm $\|\cdot\|$ (so that $\|\cdot\|_S = \|\cdot\|$).

In [36], the authors introduced the problem of robust to heavy-tailed data estimation of a mean vector with respect to any norm. The problem can be stated as follow: given N i.i.d. random vectors $X_1, \dots, X_N$ in $\mathbb{R}^d$ with mean μ^* and covariance matrix Σ, 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 [36], the authors use the median-of-means principle [1, 22, 46] to construct an estimator satisfying the following result.

Theorem 1 (Theorem 2 in [36]) *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 $1 - \delta$*

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

where B° is the unit dual ball associated with $\|\cdot\|$, (ϵ_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 it. 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 – an existence result – while robust multivariate mean estimation can also be interesting from a practical point of view [17].

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 [36], 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 Sect. 3 and that this holds even in the adversarial corruption model (see Assumption 1 in Sect. 3 below for a formal definition) and even in some situations

where there is not even a first moment; in that case, μ^* is a *location* parameter and Σ a *scatter* parameter.

The optimality of the rate in Theorem 1 has been raised in [36]. 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 [36] to obtain the following deviation-minimax lower bound result.[1]

Theorem 2 (Theorem 3 and first paragraph in p.962 in [36]) *There exists an absolute constant $c > 0$ such that the following holds. If $\hat{\mu} : \mathbb{R}^{Nd} \mapsto \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4)$,*

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

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 translations of ηB_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 [36] from Proposition 6.1 in [8] 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].

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 [31] which has also been used to obtain minimax lower bounds based on the entropy in [28] and [41].

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 [31] or Theorem 5.6 in [47]):

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

[1] The result from [36] is proved for $\Sigma = I_d$, it is however straightforward to extend it to the general case.

[2] The argument used in [36] goes from the one dimensional case studied in [8] to the d-dimensional case. It is given in a nonformal way and may require some extra argument to hold. Indeed the estimator $x^*(\hat{\Psi}_N)$ in [36] 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 [8] 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 [8] 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.

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 [36], 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 sub-Gaussian rate for the standard Euclidean norm case (that is for $S = B_2^d$) which is given in [37] by

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

Indeed, when $\|\cdot\|$ is the ℓ_2^d Euclidean norm then $\mathbb{E}\left\|\Sigma^{1/2}G\right\| = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \sim \sqrt{\mathrm{Tr}(\Sigma)}$ (see, for instance, Proposition 2.5.1 in [48]). Whereas, for the entropy of $\Sigma^{1/2}B^\circ = \Sigma^{1/2}B_2^d$ with respect to ηB_2^d, it follows from equation (5.45) in [47] 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 $\Sigma^{1/2} : \ell_2^d \mapsto \ell_2^d$ (see page 62 in [47] for a definition) and $\lambda_1 \geq \cdots \geq \lambda_d$ are the singular values of Σ. 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 Sect. 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 with respect to any $\|\cdot\|_S$.

The paper is organized as follows. In the next section, we obtain the deviation-minimax optimal rate in the benchmark i.i.d. Gaussian case. In Sect. 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 Sect. 4.

2 Deviation minimax rates in the Gaussian case: benchmark sub-Gaussian rates for the mean estimation with respect to $\|\cdot\|_S$

In this section, we obtain the optimal deviation-minimax rates of estimation of a mean vector μ^* 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 Σ repeated N times and 0 outside of these diagonal blocks.

Unlike classical minimax results holding in expectation or with constant probability (see Chapter 2 in [49]) we want, in this section, the deviation parameter δ to appear explicitly in the minimax lower bound. Moreover, this dependency of the convergence rate with respect to δ should be of the right order given by the sub-Gaussian $\sqrt{\log(1/\delta)}$ 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 [8]). This subtle behavior of the rate in terms of δ 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 [49].

Fortunately, in [28], a minimax lower bound has been proved thanks to the Gaussian shift theorem which makes the deviation parameter δ 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 μ^* with respect to $\|\cdot\|_S$. Let $S \subset \mathbb{R}^d$ be a symmetric set. We first obtain an upper bound result revealing the sub-Gaussian rate. We use the empirical mean $\bar{X}_N = N^{-1} \sum_i X_i$ as an estimator of μ^*. Using Borell TIS's inequality (Theorem 7.1 in [30] or pages 56-57 in [48]) we get: for all $0 < \delta < 1$, with probability at least $1 - \delta$,

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

where $\sigma_S = \sup_{v \in S} \sqrt{\mathbb{E} \langle v, \bar{X}_N - \mu \rangle^2}$ 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 $\ell^*(\Sigma^{1/2} S) = \sup \left(\langle G, x \rangle : x \in \Sigma^{1/2} S \right) = \mathbb{E} \left\| \Sigma^{1/2} G \right\|_S$, 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 sub-Gaussian rate (2) in (4). Our aim is now to show that the rate in (4) is deviation-minimax optimal.

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

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

then

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

It follows from the upper bound (4) and the deviation-minimax lower bound from Theorem 3 the subgaussian rate for the problem of mean estimation in $\mathbb{R}^d$ with respect to $\|\cdot\|_S$ is (up to absolute constants)

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

We may identify the two complexity and deviation terms in the rate above. 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 from Theorem 2. Theorem 3 together with (4) show that the right way to measure the statistical complexity in the problem of mean estimation in $\mathbb{R}^d$ with respect 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 [28, 41]. Following the later results in the regression model, Theorem 3 is a bit unexpected because one may think that by taking an ERM over an epsilon net of $\mathbb{R}^d$ (for instance $\hat{\mu} \in \operatorname{argmin}_{\mu\in\Lambda}\sum_{i=1}^N \|X_i - \mu\|^p$ for some $p > 0$ and Λ an ϵ-net of $\mathbb{R}^d$ with respect to $\|\cdot\|_S$) for the right choice of ϵ 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 complexity of the ϵ-net that is of the (smaller) entropy instead of the (larger) Gaussian mean width of some localized model. Theorem 3 shows that this is not the case: even discrete ERM cannot achieve a better rate than the one driven by the Gaussian mean width for the mean estimation problem with respect to any pseudo-norm.

An important consequence of Theorem 3 is obtained when $S = B_2^d$ that is for the classical problem of multivariate mean estimation with respect to the ℓ_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 sub-Gaussian 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 $\hat{\mu} : \mathbb{R}^{Nd} \mapsto \mathbb{R}^d$ is an estimator such that $\mathbb{P}_{\mu^*}^N\left[\left\|\hat{\mu} - \mu^*\right\|_2 \leq r^*\right] \geq 1-\delta$ for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$, then*

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

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 with respect to ℓ_2^d and that it is achieved by the empirical mean. In particular, there are no statistical procedures that can do better than the empirical mean uniformly over all mean vectors $\mu^* \in \mathbb{R}^d$ up to absolute constants, this includes in particular all discretized versions of $\bar{X}_N$.

Remark 1 (Exact deviation-minimax rate) An inspection of the proof of Theorem 3 reveals that it is possible to identify the exact deviation-minimax rate of convergence for the mean estimation problem with respect to any pseudo-norm for Gaussian data: for $\delta \in (0,1)$, this exact rate is given by $q^S_{1-\delta}/\sqrt{N}$ where

$$q^S_{1-\delta} := q_{1-\delta}\left(\left\|\Sigma^{1/2}G\right\|_S\right) = \inf\left(q \in \mathbb{R} : \mathbb{P}\left[\left\|\Sigma^{1/2}G\right\|_S \leq q\right] \geq 1-\delta\right) \quad (6)$$

and $G \sim \mathcal{N}(0, I_d)$. Indeed, it is clear that the empirical mean $\bar{X}_N$ achieves this rate since $\bar{X}_N - \mu^* \sim \Sigma^{1/2}G/\sqrt{N}$ and the minimax lower bound follows from the proof of Theorem 3 (more details are provided in Sect. 4).

An interesting consequence of (6) and the deviation-minimax optimality of the empirical mean and of the sub-Gaussian rate from Theorem 4 is the following computation of a quantile: for $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ and $g_1, \ldots, g_d$ i.i.d. $\mathcal{N}(0,1)$, the quantile of order $1-\delta$ of $\left(\sum_{j=1}^d \sigma_j g_j^2\right)^{1/2}$ is such that

$$\inf\left(q \in \mathbb{R} : \mathbb{P}\left[\left(\sum_{j=1}^d \sigma_j g_j^2\right)^{1/2} \leq q\right] \geq 1-\delta\right) \sim \sqrt{\sum_{j=1}^d \sigma_j} + \sqrt{\sigma_1 \log(1/\delta)}.$$

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 $f : \mathbb{R}^d \mapsto \mathbb{R}$. The Fenchel–Legendre transform of f on S is the function f_S^* defined for all $\mu \in \mathbb{R}^d$ by $f_S^*(\mu) = \sup_{v\in S} \big(\langle \mu, v\rangle - f(v)\big)$.

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 $(\mu \in \mathbb{R}^d \mapsto \langle \mu, v\rangle - f(v) : v \in S)$ of affine 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 μ^* with respect to $\|\cdot\|_S$. The construction of these two functions are based on the median-of-means principle: the dataset $\{X_1, \dots, X_N\}$ 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

$$f(v) = \frac{1}{|I_K|} \sum_{k\in I_K} \langle \bar{X}_k, v\rangle_{(k)}^* \text{ and } g(v) = \mathrm{Med}(\langle \bar{X}_k, v\rangle) = \langle \bar{X}_k, v\rangle_{\left(\frac{K+1}{2}\right)}^* \tag{7}$$

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 $a_{(1)}^* \le \cdots \le a_{(K)}^*$ (this is the rearrangement of the values a_k's themselves and not of their absolute values) and

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

is the inter-quartiles interval—without loss of generality 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:

$$f_S^* : \mu \in \mathbb{R}^d \mapsto \sup_{v\in S} \big(\langle \mu, v\rangle - f(v)\big) \text{ and } g_S^* : \mu \in \mathbb{R}^d \mapsto \sup_{v\in S} \big(\langle \mu, v\rangle - g(v)\big). \tag{8}$$

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 sub-Gaussian estimators of the mean / location vector μ^* with respect to $\|\cdot\|_S$:

$$\hat{\mu}_S^f \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} f_S^*(\mu) \quad \text{and} \quad \hat{\mu}_S^g \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} g_S^*(\mu). \tag{9}$$

In the one dimensional case, one can check that for $S = [-1, 1]$, μ_S^g is the median of the bucketed means whereas $\hat{\mu}_S^f$ is the average of the interquartile values of the bucketed means.

Remark 2 (Comparison with the trimmed mean from [40]) Authors of [40] propose a multivariate trimmed mean for robust estimation. They do not use bucketed mean and take the trimmed mean of the original data points, removing a varying fraction of the sample, that fraction depends on δ and $|\mathcal{O}|$. In contrast, $\hat{\mu}_S^f$ always use the same fraction (half) of the bucketed mean: what is varying here is rather the size of the blocks.

For some special choices of S, the Fenchel–Legendre minimization estimator $\hat{\mu}_S^g$ coincides with some classical procedures. This is for instance the case when $S = B_1^d$ (the unit ball of the ℓ_1^d-norm) or $S = B_2^d$. Indeed, when $S = B_1^d$, $\hat{\mu}_S^g$ is the coordinate-wise median-of-means:

$$\hat{\mu}_S^g = \underset{\mu=(\mu_j)\in\mathbb{R}^d}{\operatorname{argmin}} \max_{j\in[d]} \left|\mu_j - \operatorname{Med}\left(\left\langle \bar{X}_k, e_j\right\rangle\right)\right| = \left(\operatorname{Med}\left(\left\langle \bar{X}_k, e_j\right\rangle\right) : j \in [d]\right) \quad (10)$$

where $(e_j)_{j=1}^d$ 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 with respect to the ℓ_∞^d-norm from general upper bounds on $\hat{\mu}_S^g$ since in that case $\|\cdot\|_S = \|\cdot\|_\infty$.

In the case $S = B_2^d$ (that is for the mean/location estimation problem with respect to ℓ_2^d), the Fenchel–Legendre minimum estimator $\hat{\mu}_S^g$ is a minmax MOM estimator [29]. This connection allows to write $\hat{\mu}_S^g$ (as well as $\hat{\mu}_S^f$) 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 sub-Gaussian estimators in [29]. They have been proved to be deviation-minimax optimal for the mean estimation problem in [33] with respect to $\|\cdot\|_2$. Their definition only requires to consider a loss function; here we take for all $\mu \in \mathbb{R}^d$, $\ell_\mu : x \in \mathbb{R}^d \mapsto \|x - \mu\|_2^2$ and the minmax MOM estimator is then defined as

$$\tilde{\mu} \in \underset{\mu\in\mathbb{R}^d}{\operatorname{argmin}} \sup_{\nu\in\mathbb{R}^d} \operatorname{Med}\left(P_{B_k}(\ell_\mu - \ell_\nu) : k \in [K]\right) \quad (11)$$

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 sub-Gaussian rate in (2) with confidence $1 - \delta$ when the number of blocks is $K \sim \log(1/\delta)$ and K is larger than the number of adversarial outliers (i.e. $K \gtrsim |\mathcal{O}|$ where $|\mathcal{O}|$ will denote later the number of outliers) in [33]. An adaptive to K version of this estimator via the Lepski's method may also be constructed at the price of knowing $\operatorname{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ [15, 16].

Even though the minmax formulation of $\tilde{\mu}$ suggests a robust version of a descent/ascent gradient method over the median block (see [29, 33] 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 (11) is in fact equal to $\hat{\mu}_S^g$ for $S = B_2^d$ and it is therefore solution to a convex optimization problem.

Proposition 1 *The minmax MOM estimator $\tilde{\mu}$ defined in (11) satisfies $\tilde{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} g^*_{B_2^d}(\mu)$. The minmax MOM estimator (11) 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

$$\begin{aligned}
&\sup_{\nu\in\mathbb{R}^d} \operatorname{Med}\left(P_{B_k}(\ell_\mu - \ell_\nu)\right) = \sup_{\nu\in\mathbb{R}^d} \left(-2\operatorname{Med}(\langle \bar{X}_k - \mu, \mu-\nu\rangle) - \|\mu-\nu\|_2^2\right) \\
&= \sup_{\|v\|_2=1} \sup_{\theta\geq 0} \left(2\theta \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle) - \theta^2\right) = \sup_{\|v\|_2=1} \left(\operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)\right)^2 \\
&= \left(\sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)\right)^2.
\end{aligned}$$

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 not obvious given the definition of $\tilde{\mu}$ in (11).

Proposition 1 suggests a new formulation for $\hat{\mu}_S^g$ and $\hat{\mu}_S^f$. It is indeed possible to write these estimators as regularized estimators instead of their original constraint formulation (note that the Fenchel–Legendre transforms in (8) 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 (7): for all $\nu \in \mathbb{R}^d$,

$$F_S(\nu) = f(\nu) + \frac{\|\nu\|_{S^\circ}^2}{4} \text{ and } G_S(\nu) = g(\nu) + \frac{\|\nu\|_{S^\circ}^2}{4} \tag{12}$$

where $v \mapsto \|v\|_{S^\circ} = \sup(\langle x, v\rangle : \|v\|_S \leq 1)$ is the dual norm of $\|\cdot\|_S$. We also consider their Fenchel–Legendre transforms over the entire set $\mathbb{R}^d$: for all $\mu \in \mathbb{R}^d$,

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

The next result shows that the later two Fenchel–Legendre transforms can be used to define the two estimators $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$.

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

Proof We prove the result only for $\hat{\mu}_S^g$ since it is almost the same for $\hat{\mu}_S^f$. The proof of Proposition 2 for $\hat{\mu}_S^g$ is similar to the one of Proposition 1 where the ℓ_2-norm is replaced by $\|\cdot\|_{S^\circ}$. We have for all $\mu \in \mathbb{R}^d$

$$G_S^*(\mu) = \sup_{v\in\mathbb{R}^d} \left(-\operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle) - \frac{\|v\|_{S^\circ}^2}{4}\right) = \sup_{\|v\|_{S^\circ}=1} \sup_{\theta>0} \left(-\operatorname{Med}(\langle \bar{X}_k - \mu\rangle, \theta v) - \frac{\theta^2}{4}\right)$$
$$= \sup_{\|v\|_{S^\circ}=1} \left(\operatorname{Med}\left(\langle \mu - \bar{X}_k, v\rangle\right)\right)^2 = \sup_{v\in\operatorname{conv}(S)} \left(\operatorname{Med}\left(\langle \mu - \bar{X}_k, v\rangle\right)\right)^2$$
$$= \left(\sup_{v\in\operatorname{conv}(S)} \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)\right)^2$$

where we used that $\operatorname{conv}(S)$ is the unit ball of $\|\cdot\|_{S^\circ}$ i.e. $\operatorname{conv}(S) = \{v \in \mathbb{R}^d : \|v\|_{S^\circ} \leq 1\}$ and the symmetry of S. We conclude since

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

□

As a consequence of Proposition 2, one can write the two estimators $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ as solutions to unconstrained minmax optimization problems like the minmax MOM estimator (11) and in particular, one may design an alternating ascent/descent subgradient algorithm similar to the one from [29]—we expect the one associated with $\hat{\mu}_S^f$ which uses half of the dataset at each iteration to be more efficient than the one associated with $\hat{\mu}_S^g$ which uses only the N/K data in the median block at each iteration. That is the reason why we provide in Algorithm 1 this algorithm only for

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

We also recall that S° is the dual body of $\operatorname{conv}(S)$ and that by the Danskin's theorem the subdifferential of $\|\cdot\|_{S^\circ}$ at $v \in \mathbb{R}^d$ when S° is a compact and non empty set is given by the convex hull of all $x \in S^\circ$ such that $\|v\|_{S^\circ} = \langle x, v\rangle$.

input : the data $X_1, \dots, X_N$, 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 μ

Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \dots, N\}$ at random
Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
Compute $\tilde{\mu}^{(0)}$ the coordinate-wise median-of-means and put $\mu^{(0)} = \tilde{\mu}^{(0)}$ and $\nu^{(0)} = \tilde{\mu}^{(0)}$
while $\|\mu^{(t)} - \mu^{(t+1)}\|_S \geq \epsilon$ **do**
    Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \dots, N\}$ at random
    Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
    Find the inter-quartile block numbers $k_1, \dots, k_{(K+1)/2} \in [K]$ such that

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

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

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

    Update $\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \nabla_\nu^{(t+1)}$.
    Make one descent step: $\mu^{(t+1)} \leftarrow \mu^{(t)} - \theta_t \nu^{(t+1)}$.
end
Return $\mu^{(t+1)}$

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

In a recent work [3], the author introduces a minmax estimator in a general separable Banach space having the following form in the finite dimensional case

$$\hat{\mu} \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} \sup_{\nu \in r_N \mathrm{conv}(S)} r_N^{-1} \left| \langle \mu, \nu \rangle - \frac{1}{N} \sum_{i=1}^N \sin(\langle \nu, X_i \rangle) \right| \tag{13}$$

where r_N has to depend on δ, Σ as well as $\|\mu^*\|_S$ but can be chosen in a data-driven way using a Lepski's method. This estimator is proved to achieve the same rate as in Theorem 5 below up to a mean-dependent term of higher order incurred due the lack of shift-equivariance. The author of [3] also underlines that $\hat{\mu}$ is solution to a convex optimization. However, and this is also the case for estimators $\hat{\mu}_S^g$ and $\hat{\mu}_S^f$, convexity of the objective function is not enough to insure convergence guarantees of

a related algorithm such as the one in Algorithm 1 or other saddle point algorithms. As in [3], we also leave the problem of finding a polynomial time algorithm for robust mean estimation with respect to any norm opened, we refer to [45] where robust mean estimation in a broader algorithmic context is considered.

Remark 3 Tournament estimators from [38, 39] and the minmax MOM estimator from [29, 34] share similar ideas: tournament estimators are like Le Cam test estimators and, following the work of [4], we can derive a ρ-aggregation method out of it which appears to be a minmax MOM estimator (see [29]) in the setup of MOM estimators. While those three estimators lead to the same optimal results, the different formulations used to define is key to the construction of algorithms. To give an example from a different field, in compressed sensing, the minimum ℓ_0-norm and minimum ℓ_1-norm estimators are equal with high probability for Gaussian measurements. However, the minimum ℓ_0-norm estimator does not suggest any helpful algorithm (a descent algorithm for the min-ℓ_0 estimator is likely to be with an exponential time complexity time) whereas meaningful algorithms can be used to approach a minimum ℓ_1-norm estimator; for instance, it can be written as Linear Programming. The same way, the minmax MOM suggests algorithms one of them was tested in [29, 34], this is something that had not been suggested by previous formulations such as the tournament or the Le Cam test estimators. What we found here brings a new brick to the picture by showing that the minmax MOM is in fact solution to a convex program (see Proposition 1), a fact which was not obvious given the original definition of this estimator. This new convex formulation of the minmax MOM estimator also suggests a generalization to any norm: this is what we did here with the Fenchel–Legendre transform estimators, moreover, given that this estimator is also a minmax MOM estimator it thereby suggests algorithms such as the one in Algorithm 1.

3.2 The adversarial corruption model and two models for inliers.

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 inliers which will be either the existence of a second moment or a regularity assumption on a family of cumulative distribution functions 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 inliers $(X_i)_{i\in\mathcal{I}}$ such that $\forall i \in \mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \dots, N\}$ is unknown to the statistician.

In the adversarial contamination model from Assumption 1, the set $\mathcal{O} \subset [N]$ 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 ℓ_2^d-norm are modified by the adversary). The adversarial corruption model covers the Huber ϵ-contamination model [21] and also the $\mathcal{O} \cup \mathcal{I}$ framework from [27, 29, 34].

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

The two assumptions on the inliers 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 [50]). 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 μ^* 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 in cases where a first moment may not exist. In that case, μ^* is called a *location parameter*, Σ is called a *scale parameter* and 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 of the normalized bucketed means at / around (in the non-asymptotic version) the median adapted to the multidimensional and non-asymptotic setup. We are now introducing such an assumption.

Assumption 3 The inliers 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

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}}\sum_{i=1}^{N/K}\left\langle \tilde{X}_i - \mu^*, v\right\rangle > r\right]. \tag{14}$$

When S is a symmetric set, if we let r tends to zero in Assumption 3, we see that for all $v \in S, \langle \mu^*, v \rangle$ is a median of $\sqrt{K/N} \sum_{i=1}^{N/K} \langle \tilde{X}_i, v \rangle$. Hence, μ^* should be considered as a multivariate location / median parameter of $\sqrt{K/N} \sum_{i=1}^{N/K} \tilde{X}_i$ and not as a mean since we do not assume the existence of any moment in Assumption 3.

A typical example where Assumption 3 holds is when $S = \mathcal{S}_2^{d-1}$ (that is for the location estimation problem with respect to the Euclidean ℓ_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 $r \in \mathbb{R}$,

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

$\mathbb{P}_R$ 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 [7]). 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 $f(c_0') \geq c_1'$, which in turn holds when there exists constants $c_0, c_1 > 0$ such that $H(c_0) \leq 1/2 - c_1$.

Furthermore, we have, for all $t > 0$

$$\mathbb{P}[R\langle U, e_1 \rangle \geq c_0] \leq \mathbb{P}[\langle U, e_1 \rangle \geq t/\sqrt{d}] + \frac{1}{2}\mathbb{P}[R \geq c_0\sqrt{d}/t] \leq e^{-t^2/2} + \frac{1}{2}\mathbb{P}[R \geq c_0\sqrt{d}/t],$$

where the classical second inequality can be found for instance in [53], Chapter 5. So if for some constants $\tilde{c}_0, \tilde{c}_1 > 0$, $\mathbb{P}[R \geq \tilde{c}_0\sqrt{d}] \leq 1 - \tilde{c}_1$, then Assumption 3 holds. This is for instance the case, when R is distributed like $\|G\|_2$ for $G \sim \mathcal{N}(0, I_d)$ by Borell-TIS inequality, but as well when R is the positive part of a standard Cauchy variable for instance. As a consequence, Assumption 3 has nothing to do with the existence of moments 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 μ^* 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.

Let us provide a final example where averaging is necessary in Assumption 3. We consider the case $d = 1$ and the density function (with respect to the Lebesgue measure on $\mathbb{R}$) $t \in \mathbb{R} \mapsto f(t) = (3/2)t^2 I(|t| \leq 1)$. We assume that $\tilde{X}_1$ is distributed according to f. In that case, we have for $K = N$, $v = 1$ and all $0 < r < 1$, $H_{N,K=N,1}(r) - 1/2 = -\int_0^r f(t)dt = -r^3/2$ and there are no absolute constants $c_0 > 0$ and $c_1 > 0$ such that for all $0 < r \leq c_0$, $H_{N,K=N,1}(r) \leq 1/2 - c_1 r$. In other words, Assumption 3 does not hold for $K = N$. However, $g := t \mapsto \sqrt{2}(f * f)(\sqrt{2}t)$ is a density function of $(\tilde{X}_1 + \tilde{X}_2)/\sqrt{2}$ which is such that $g(t) = 27/5 + O(t)$ when $t \sim 0$. As a consequence, one can find an absolute constant $c_0 > 0$ such that for all $0 < r \leq c_0$, $H_{N,K=N/2,1}(r) \leq 1/2 - (27/10)r$ and so Assumption 3 holds for $K = N/2$. We can see in this simple one-dimensional example that averaging may be needed in order to satisfy Assumption 3. The reason behind this observation is that in this example the density at 0 of $\tilde{X}_1$ is zero (and so the classical assumption of asymptotic normality of empirical median does not hold) whereas it is equal to $27/5$ (an absolute non zero constant) for $(\tilde{X}_1 + \tilde{X}_2)/\sqrt{2}$ (and so normality asymptotic of the empirical median holds). It would be interesting to see if one can find an example extending the previous one for any value of K; that is to find a density function for $\tilde{X}_1$ so that the density function at zero of $(\tilde{X}_1 + \cdots + \tilde{X}_k)/\sqrt{k}$ is zero for all $1 \leq k \leq n$ and the one of $(\tilde{X}_1 + \cdots + \tilde{X}_{n+1})/\sqrt{n+1}$ is positive. We leave this problem for future research. However, note that if such an example exists then the $\sqrt{N/K}$ normalization used in Assumption 3 may not be the correct one, in particular under a $L_{1+\gamma}$ moment assumption for some $0 < \gamma < 1$, the right normalization should be like $(N/K)^{1/(1+\gamma)}$ and the resulting rates may not be anymore sub-Gaussian rates.

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

In this section, we obtain estimation bounds with respect to $\|\cdot\|_S$ 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 2.

In this section, we obtain high probability estimation upper bounds satisfied by $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ with respect to $\|\cdot\|_S$ in the adversarial contamination and heavy-tailed inliers model. The rate of convergence is given by the quantity

$$r_S^* = \max\left(\frac{64}{\sqrt{N}}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu^*)\right\|_S, \sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{64K}{N}}\right). \quad (15)$$

The key metric property satisfied by the two Fenchel–Legendre transforms f_S^* and g_S^* in the adversarial contamination and heavy-tailed inliers 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 $\mu \in \mathbb{R}^d$, $\left|g_S^*(\mu) - \|\mu - \mu^*\|_S\right| \leq g_S^*(\mu^*) \leq r_S^*$ and $\left|f_S^*(\mu) - \|\mu - \mu^*\|_S\right| \leq f_S^*(\mu) \leq r_S^*$.*

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 $\mu \mapsto \|\mu-\mu^*\|_S$ 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 $\mu \mapsto \|\mu-\mu^*\|_S$ which is μ^*. 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 (15). For all $K > 16|\mathcal{O}|$, with probability at least $1-\exp(-K/512)$,*

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

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 [36], this result shows that the Gaussian mean width term appearing in Theorem 1 is actually not necessary and may be responsible of a suboptimal rate since one can construct examples where the Gaussian mean width is strictly larger than the Rademacher complexity. One such example can be seen when $N=1$ (because of the CLT, the gap is better seen for $N=1$), $S=B_1^d$, $\mu^*=0$ and $\Sigma=I_d$ for which the Gaussian mean width is $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_S = \mathbb{E}\sup_{\|x\|_1\leq 1}\langle x, G\rangle \sim \sqrt{\log d}$ whereas the Rademacher complexity is

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i(X_i-\mu^*)\right\|_S = \mathbb{E}\sup_{\|x\|_1\leq 1}\epsilon_1\langle X_1, x\rangle = \mathbb{E}\|X_1\|_\infty \sim 1$$

for instance when $X_1=(\eta_j)_{j=1}^d$ and $\eta_1,\dots\eta_d$ are independent variance one and bounded random variables.

Theorem 5 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 sub-Gaussian rate because in that case the Rademacher complexity term in r_S^* is less or equal to $\sqrt{\mathrm{Tr}(\Sigma)}$ [24]. In particular, since $\hat{\mu}_S^g$ is the minmax MOM estimator in that case, we recover the main result from [33].

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, μ^* is a location parameter so that Assumption 3 holds. 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 [51, 52] of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0,1\}$. The rate of convergence we obtain in that case is given by

$$r^\diamond = \frac{C_0}{c_1}\left(\sqrt{\frac{VC(S_0^*)}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{|\mathcal{O}|}{c_1\sqrt{KN}} \tag{16}$$

where S_0^* is the set of extreme points of S (that is any point in S which cannot be written as the mid point of any two different points in S), c_1 is the absolute constant from Assumption 3, C_0 the absolute constant from (32) and $u > 0$ a confidence parameter. We abusively call VC-dimension of a set $C \subset \mathbb{R}^d$ the VC-dimension of the set of half-spaces generated by the vectors of C.

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 $\mathbb{R}^d$. Let S_0^* denote the set of extreme points of S. Grant Assumptions 1 and 3 for some $K \in [N]$. Let $u > 0$. Assume that $C_0\left(\sqrt{VC(S_0^*)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \le c_0c_1$ where c_0 is the constant from Assumption 3. With probability at least $1-\exp(-u)$, for all $\mu \in \mathbb{R}^d$, $\left|g_S^*(\mu) - \|\mu - \mu^*\|_S\right| \le r^\diamond$.*

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 μ^*, a minimum of $\mu \mapsto \|\mu - \mu^*\|_S$. This yields 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 $C_0\left(\sqrt{VC(S_0^*)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \le c_0c_1$ where c_0 is the constant from Assumption 3. With probability at least $1-\exp(-u)$, $\left\|\hat{\mu}_S^g - \mu^*\right\|_S \le 2r^\diamond$ where $r^\diamond$ is defined in (16).*

Unlike Theorem 5, Theorem 6 may hold even when a first moment does not exist. The result from Theorem 6 holds 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 choose adaptively K in that case, [15, 16]). The price for adversarial corruption in (16) 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 Assumption 3 holds and Theorem 6 applies, 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.

Sub-Gaussian rates under weak or no moment assumption It is possible to recover (up to absolute constants) the sub-Gaussian rate (5) in Theorem 5 for $K \sim \log(1/\delta)$ when the Rademacher complexity term from (15) and the Gaussian mean width from (5) satisfy

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\right\|_S \lesssim \ell^*\left(\Sigma^{1/2}S\right). \tag{17}$$

Such a result (i.e. Rademacher complexity is smaller than the Gaussian mean width up to absolute constants) 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 sub-Gaussian rate (5) in that case. It is also true when the $\tilde{X}_i$'s are sub-Gaussian vectors. There are other situations under weaker moment assumption where (17) holds.

For instance, when $S = B_2^d$, (17) holds under only a L_2-moment assumption (see [24]). It also holds for $S = B_1^d$ when the $\tilde{X}_i$'s are isotropic with coordinates having $\log d$ sub-Gaussian 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 $j \in [d]$) and $N \gtrsim \log d$. Together with (10) and Theorem 5, this implies that the coordinate-wise MOM is a sub-Gaussian estimator of the mean with respect to the ℓ_∞^d-norm under a $\log d$ sub-Gaussian moment assumption. Upper bounds such as (17) have been extended in [42] to general unconditional norms.

It is also possible to recover the sub-Gaussian 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 independent 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 (10). It follows from Theorem 6, the fact that $S = B_1^d$ has $2d$ extreme points given by the vectors of the canonical basis and their opposite and the VC dimension of a set of $2d$ points is less than $\log(2d)$ that, when $\log(d) \lesssim N$ and $|\mathcal{O}| \lesssim N$ then for all $\log(d) \lesssim u \lesssim N$, with probability at least $1 - \exp(-u)$,

$$\left\|\hat{\mu}_S^g - \mu^*\right\|_\infty \leq 2C_0\left(\sqrt{\frac{\log(d)}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{2\pi|\mathcal{O}|}{N} \tag{18}$$

which is the deviation-minimiax optimal sub-Gaussian rate (5) we would have gotten if the $\tilde{X}_i$ were i.i.d. isotropic Gaussian vectors centered at μ^* corrupted by $|\mathcal{O}|$ adversarial outliers (up to absolute constants). But here, (18) is obtained without the existence of a first moment. Moreover, in (18), 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 inliers has a Gaussian distribution – this differs from the $\sqrt{|\mathcal{O}|/N}$ information-theoretic lower bound that has been obtained for some non-symmetric inliers. 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 $\mathcal{O}(K)$ [5], 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.

We also observe that without adversarial corruption (that is in the i.i.d. case) the rate obtained in (18) is minimax in the optimal range $N \gtrsim \log(d)$ as proved in the following result.

Proposition 3 *Let $\tilde{X} = \mu^* + Z$ where $Z = (z_j)_{j=1}^d$ has symmetric around 0 Cauchy distributed coordinates and let $\tilde{X}_1, \dots \tilde{X}_N$ be N i.i.d. copies of $\tilde{X}$. We denote by $\mathbb{P}_{\mu^*}^{\otimes N}$ the joint distribution of $(\tilde{X}_1, \dots \tilde{X}_N)$. We have*

$$\inf_{\hat{\mu}} \sup_{\mu^* \in \{\pm e_1, \dots, \pm e_d\} \cup \{0\}} \mathbb{P}_{\mu^*}^{\otimes N} \left[\left\| \hat{\mu} - \mu^* \right\|_\infty \geq r^*(N,d) \right] \geq \frac{1}{10} \quad \text{where} \quad r^*(N,d) := 2\sqrt{\exp\left(\frac{\log(2d)}{16N}\right) - 1} \tag{19}$$

where $\inf_{\hat{\mu}}$ denotes the infimum over all estimators and $(e_1, \dots, e_d)$ is the canonical basis of $\mathbb{R}^d$.

It follows from Proposition 3 that when $N \gtrsim \log(d)$ then $r^*(N,d) \gtrsim \sqrt{\log(d)/N}$ and so the estimation rate achieved by the coordinate-wise median in (18) is minimax (when $|\mathcal{O}| = 0$). It is the regime where we recover the sub-Gaussian rate. However, the sub-Gaussian rate still keeps (up to absolute constants) the value $\sqrt{\log(d)/N}$ even for $N \lesssim \log(d)$ but this is not the case for the model with a Cauchy noise studied in (18). For instance, when $N = 1$, the minimax lower bound rate $r^*(N,d)$ is of the order of $d^{1/32}$ which is different from the $\sqrt{\log d}$ sub-Gaussian rate. As a consequence, the sub-Gaussian rate is indeed achieved up to $N \gtrsim \log d$ by $\hat{\mu}_S^g$ but it cannot be the case (even for any other estimator) for a number of data less than an order of $\log d$ as proved by Proposition 3.

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 δ. 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 Σ and the first one with mean 0 and the second one with mean $3r^* v_1$. We set $A_0 = (\hat{\mu})^{-1}(B_S(0, r^*)) = \{(x_1, \dots, x_N) \in \mathbb{R}^{Nd} : \left\| \hat{\mu}(x_1, \dots, x_N) \right\|_S \leq r^*\}$ 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 slight generalization of Lemma 3.3 in [28] which is based on a version of the Gaussian shift Theorem from [35]. □

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

$$\nu_v(A) \geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_u(A)) + \|\Sigma_0^{-1/2}(u - v)\|_2\big) \tag{20}$$

where $\Sigma_0^{-1/2}$ is the square root of the pseudo-inverse of Σ_0.

Proof of Lemma 3 When $\Sigma_0 = I_{Nd}$, Lemma 3 is exactly Lemma 3.3 in [28] 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 [28] 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 (20). □

It follows from Lemma 3 that

$$\mathbb{P}_1[A_0] \geq 1 - \Phi\left[\Phi^{-1}(1 - \mathbb{P}_0[A_0]) + \left\|\Sigma_0^{-1/2}(0 - (3r^* v_1, \ldots, 3r^* v_1))\right\|_2\right]. \tag{21}$$

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

$$\left\|\Sigma_0^{-1/2}(0 - (3r^* v_1, \ldots, 3r^* v_1))\right\|_2 = 3r^* \sqrt{N} \left\|\Sigma^{-1/2} v_1\right\|_2. \tag{22}$$

As a consequence, if $3r^* \sqrt{N} \left\|\Sigma^{-1/2} v_1\right\|_2 \leq -\Phi^{-1}(\delta)$ then, in (21), 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 [25] (here we use that for all $x \geq 0$, $\Phi(-x) \geq 2\varphi(x)/\sqrt{4 + x^2} + x$ where φ 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

$$r^* \geq \frac{1}{12} \sqrt{\frac{\log(1/\delta)}{N}} \frac{1}{\inf_{\|v\|_S = 1} \left\|\Sigma^{-1/2} v\right\|_2}. \tag{23}$$

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*

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

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 $|\langle A^{-1}v/\left\|A^{-1}v\right\|, Aw\rangle| \leq 1/\left\|A^{-1}v\right\|$. The later holds for all v such that $\|v\|_S = 1$ and $\{A^{-1}v/\left\|A^{-1}v\right\| : \|v\|_S = 1\}$ 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 (23) and Lemma 4 for $\|\cdot\| = \|\cdot\|_2$ and $A = \Sigma^{1/2}$ that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log(1/\delta)}{N}} \sup_{w\in S} \left\|\Sigma^{1/2}w\right\|_2. \tag{24}$$

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 $\mu \in \mathbb{R}^d$, $\mathbb{P}_\mu^N\left[\left\|\hat{\mu}-\mu\right\|_S \leq r^*\right] \geq 1-\delta$ which is equivalent to say that

$$\delta \geq \sup_{\mu\in\mathbb{R}^d} \mathbb{E}_\mu^N \phi\left(\frac{\left\|\hat{\mu}-\mu\right\|_S}{r^*}\right) \tag{25}$$

where we set $\phi : t \in \mathbb{R} \mapsto I(t > 1)$ 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 γ 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 (25) that

$$\begin{aligned}\delta &\geq \int_{\mu\in\mathbb{R}^d} \mathbb{E}_\mu^N \phi\left(\frac{\left\|\hat{\mu}-\mu\right\|_S}{r^*}\right)\gamma(\mu)d\mu \\ &= \mathbb{E}\left[\mathbb{E}\left[\phi\left(\frac{\left\|\hat{\mu}(X_1,\ldots,X_N)-\mu\right\|_S}{r^*}\right)|X_1,\ldots,X_N\right]\right].\end{aligned} \tag{26}$$

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 (26). We first recall Anderson's Lemma. □

Theorem 7 (Anderson's Lemma, see p. 70 in [26]) *Let Γ be a semi-definite $d \times d$ matrix and $Z \sim \mathcal{N}(0, \Gamma)$. Let $w : \mathbb{R}^d \mapsto \mathbb{R}$ 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 on $X_1, \ldots, X_N$. Therefore, applying Anderson's Lemma conditionally on $X_1, \ldots, X_N$, we obtain in (26) 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

$$\delta \geq \mathbb{P}\left[\left\|\Sigma^{1/2}G\right\|_S \geq \sqrt{N}r^*\right]. \tag{27}$$

Using Borell-TIS's inequality (Theorem 7.1 in [30] or pages 56-57 in [48]), 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 (24) for $\delta = 1/4$ that

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

□

Proof of the exact minimax rate from Remark 1 It follows from (27) that if $\hat{\mu}$ is an estimator such that for all $\mu \in \mathbb{R}^d$, $\mathbb{P}_\mu^N\left[\left\|\hat{\mu} - \mu\right\|_S \leq r^*\right] \geq 1 - \delta$ then necessarily $\delta \geq \mathbb{P}\left[\left\|\Sigma^{1/2}G\right\|_S \geq \sqrt{N}r^*\right]$ which is equivalent to say that $r^* \geq q^S_{1-\delta}/\sqrt{N}$. This lower bound holds for any value of $\delta \in (0, 1)$.

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

$$\begin{aligned}\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2^2 - \left(\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 &= \mathbb{E}\left(\left\|\Sigma^{1/2}G\right\|_2 - \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 \\ &= \int_0^\infty \mathbb{P}\left[\left|\left\|\Sigma^{1/2}G\right\|_2 - \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right| \geq \sqrt{t}\right]dt \leq 2\sigma^2_{B_2^d}\end{aligned}$$

where $\sigma^2_{B_2^d} = \sup_{\|v\|_2=1}\left\|\Sigma^{1/2}v\right\|_2^2 = \|\Sigma\|_{op}$. Since $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2^2 = \mathrm{Tr}(\Sigma)$, we have $\left(\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 \geq \mathrm{Tr}(\Sigma) - 2\|\Sigma\|_{op}$. Therefore, $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \geq \sqrt{\mathrm{Tr}(\Sigma)/2}$ when $\mathrm{Tr}(\Sigma) \geq 4\|\Sigma\|_{op}$ and when $\mathrm{Tr}(\Sigma) < 4\|\Sigma\|_{op}$, we use the lower bound from (24) and an argument similar to the one appearing in the end of the proof of Theorem 3 to get the result. □

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

$$\frac{4}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu)\right\|_S + \sqrt{\frac{2u}{K}} + \sup_{v\in S}H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}. \tag{28}$$

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

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

$$\begin{aligned}\left|g_S^*(\mu) - \left\|\mu - \mu^*\right\|_S\right| &= \left|\sup_{v\in S}\left(\langle \mu, v\rangle - g(v)\right) - \sup_{v\in S}\langle v, \mu - \mu^*\rangle\right| \\ &\leq \sup_{v\in S}\left|\langle \mu^*, v\rangle - g(v)\right| = g_S^*(\mu^*) \end{aligned} \tag{29}$$

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 $v \in S$,

$$\sum_{k\in[K]} I(\left\langle \bar{X}_k - \mu^*, v\right\rangle > R_S^*) < \frac{K}{2}. \tag{30}$$

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

$$\begin{aligned}\sum_{k\in[K]} I(\left\langle \bar{X}_k - \mu^*, v\right\rangle > R_S^*) &= \sum_{k\in\mathcal{K}} I(\left\langle \bar{X}_k - \mu^*, v\right\rangle > R_S^*) + \sum_{k\notin\mathcal{K}} I(\left\langle \bar{X}_k - \mu^*, v\right\rangle > R_S^*) \\ &\leq \sum_{k\in[K]} I(\left\langle \overline{\tilde{X}}_k - \mu^*, v\right\rangle > R_S^*) + |\mathcal{O}|. \end{aligned} \tag{31}$$

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

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

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

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

Next, we use several tools from empirical process theory and in particular, for a symmetrization argument, we consider a family of N independent Rademacher variables $(\epsilon_i)_{i=1}^N$ independent of the $(\tilde{X}_i)_{i=1}^N$. In *(bdi)* below, we use the bounded difference inequality (Theorem 6.2 in [6]). In *(sa-cp)*, we use the symmetrization argument and the contraction principle (Chapter 4 in [31]) – we refer to the supplementary material of [34] for more details. We have, with probability at least $1 - \exp(-u)$,

$$\begin{aligned}
&\sup_{v\in S}\left(\sum_{k\in[K]} \phi\left(\frac{\langle \overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle \overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right)\right) \\
&\overset{(bdi)}{\leq} \mathbb{E}\sup_{v\in S}\left(\sum_{k\in[K]} \phi\left(\frac{\langle \overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle \overline{\tilde{X}}_k - \mu^*, v\rangle}{R_S^*}\right)\right) + \sqrt{2uK} \\
&\overset{(sa-cp)}{\leq} \frac{4K}{NR_S^*}\mathbb{E}\sup_{v\in S}\Big\langle v, \sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu^*)\Big\rangle + \sqrt{2uK} \\
&= \frac{4K}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu^*)\right\|_S + \sqrt{2uK}.
\end{aligned}$$

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

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

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

and therefore (28) 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 (28) but where the right-hand side value $1/2$ is replaced by $1/4$ in (28). 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 [51, 52] of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0, 1\}$ in our case, and instead of taking the maximum over $v \in \mathbb{R}^d$, we only take a maximum over $v \in S_0^*$. 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 $x_1, \dots, x_n \in \mathbb{R}^d$ for which the set $\{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F})\}$ is of maximal cardinality, that is of size 2^n. We also know (see, for instance, Chapter 3 in [23]) the following concentration bound: let $Y_1, \dots, Y_n$ 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 $1 - \exp(-u)$,

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n f(Y_i) - \mathbb{E} f(Y_i) \right) \leq C_0 \left(\sqrt{\frac{VC(\mathcal{F})}{n}} + \sqrt{\frac{u}{n}} \right). \tag{32}$$

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 $\mu \in \mathbb{R}^d$, $\left| g_S^*(\mu) - \|\mu - \mu^*\|_S \right| \leq R^\diamond$ where $R^\diamond$ is any point such that

$$C_0 \left(\sqrt{\frac{VC(S_0^*)}{K}} + \sqrt{\frac{u}{K}} \right) + \sup_{\|v\|_2 = 1} H_{N,K,v} \left(R^\diamond \sqrt{\frac{N}{K}} \right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{33}$$

where C_0 is the constant from (32). In particular, when Assumption 3 holds then one can check that (33) 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 above. To that end we follow the same strategy as in the proof of Lemma 1 up to (31) except that R_S^* is replaced by $R^\diamond$ and that S is replaced by its set of extreme points S_0^* (the latter holds because of the Krein-Milman theorem $\mathrm{conv}(S) = \mathrm{conv}(S_0^*)$). From that point, we use (32) and the VC dimension of the set of affine half spaces to get that with probability at least $1 - \exp(-u)$, for all $v \in S_0^*$,

$$\sum_{k \in [K]} I(\langle \bar{\tilde{X}}_k - \mu^*, v \rangle > R^\diamond) \leq H_{N,K,v} \left(R^\diamond \sqrt{\frac{N}{K}} \right) + C_0 \left(\sqrt{\frac{VC(S_0^*)}{K}} + \sqrt{\frac{u}{K}} \right)$$

and so by definition of $R^\diamond$, on the same event, for all $v \in S_0^*$, $\sum_{k \in [K]} I(\langle \bar{X}_k - \mu^*, v \rangle > R^\diamond) < 1/2$. This concludes the proof. □

Proof of Proposition 3 We apply Theorem 2.5 from [49] for the ℓ_∞^d-distance and a set of hypothesis $\Theta := \{\pm r e_1, \dots, \pm r e_d\} \cup \{0\}$ where $r > 0$ will be chosen later. We

have for all $\mu_1 \neq \mu_2 \in \Theta$, $\|\mu_1 - \mu_2\|_\infty \geq r$. We let $\mu_0 = 0$. We have for all $\mu = (\mu_j)_{j=1}^d \in \Theta \backslash \{0\}$, $\mathbb{P}_\mu^{\otimes N} << \mathbb{P}_{\mu_0}^{\otimes N}$ and using [13] we get the following estimate of the Kullback-Leiber divergence between $\mathbb{P}_\mu^{\otimes N}$ and $\mathbb{P}_{\mu_0}^{\otimes N}$:

$$KL(\mathbb{P}_\mu^{\otimes N}, \mathbb{P}_{\mu_0}^{\otimes N}) = NKL(\mathbb{P}_\mu, \mathbb{P}_{\mu_0}) = N \sum_{j=1}^d KL(\mathbb{P}_{\mu_j}, \mathbb{P}_{\mu_{0j}})$$
$$= N \sum_{j=1}^d \log\left(1 + \frac{(\mu_j - \mu_{j0})^2}{4}\right) = N \log\left(1 + \frac{r^2}{4}\right) \tag{34}$$

where $\mathbb{P}_\mu$ is the probability distribution of $\mu + Z$ (where $Z = (z_j)_{j=1}^d$ has symmetric around 0 Cauchy distributed coordinates) and $\mathbb{P}_{\mu_j}$ is the probability distribution of $\mu_j + z_j$. Next, we choose r such that

$$N \log\left(1 + \frac{r^2}{4}\right) = \frac{1}{2d} \sum_{j=1}^{2d} KL(\mathbb{P}_{re_j}^{\otimes N}, \mathbb{P}_{\mu_0}^{\otimes N}) + KL(\mathbb{P}_{-re_j}^{\otimes N}, \mathbb{P}_{\mu_0}^{\otimes N}) \leq \frac{\log(2d)}{16}$$

that is $r = 2\left(\exp(\log(2d)/(16N)) - 1\right)^{1/2}$. In that case, Theorem 2.5 from [49] applies and we get the desired result. □

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

References

1. Alon, N., Matias, Y., Szegedy, M.: The space complexity of approximating the frequency moments. J. Comput. Syst. Sci. **58**(1, part 2):137–147 (1999). Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)
2. Artstein, S., Milman, V., Szarek, S.J.: Duality of metric entropy. Ann. Math. (2) **159**(3), 1313–1328 (2004)
3. Bahmani, S.: Nearly optimal robust mean estimation via empirical characteristic function (2021)
4. Baraud, Y., Birgé, L., Sart, M.: A new method for estimation and model selection: ρ-estimation. Invent. Math. **207**(2), 425–517 (2017)
5. Blum, M., Floyd, R.W., Pratt, V.R., Rivest, R.L., Tarjan, R.E.: Time bounds for selection. J. Comput. Syst. Sci. **7**(4), 448–461 (1973)
6. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities. Oxford University Press, Oxford (2013). A nonasymptotic theory of independence, With a foreword by Michel Ledoux
7. Bryc, W.: The Normal Distribution. Lecture Notes in Statistics, vol. 100. Springer, New York (1995). Characterizations with applications
8. Catoni, O.: Challenging the empirical mean and empirical variance: a deviation study. Ann. Inst. Henri Poincaré Probab. Stat. **48**(4), 1148–1185 (2012)
9. Catoni, O., Giulini, I.: Dimension-free Pac-Bayesian bounds for matrices, vectors, and linear least squares regression. arXiv preprint arXiv:1712.02747 (2017)
10. Catoni, O., Giulini, I.: Dimension-free Pac-Bayesian bounds for the estimation of the mean of a random vector. arXiv preprint arXiv:1802.04308 (2018)

11. Chen, M., Gao, C., Ren, Z.: Robust covariance and scatter matrix estimation under Huber's contamination model. Ann. Stat. **46**(5), 1932–1960 (2018)
12. Cherapanamjeri, Y., Flammarion, N., Bartlett, P.L.: Fast mean estimation with sub-Gaussian rates (2019)
13. Chyzak, F., Nielsen, F.: A closed-form formula for the Kullback–Leibler divergence between Cauchy distributions. CoRR, arXiv:abs/1905.10965 (2019)
14. Dalalyan, A.S., Minasyan, A.: All-in-one robust estimator of the Gaussian mean. arXiv preprint arXiv:2002.01432 (2020)
15. Depersin, J., Lecué, G.: Robust sub-Gaussian estimator of a mean vector in nearly linear time (2019)
16. Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R.I.: Sub-Gaussian mean estimators. Ann. Stat. **44**(6), 2695–2725 (2016)
17. Diakonikolas, I., Kamath, G., Kane, D.M., Li, J., Moitra, A., Stewart, A.: Being robust (in high dimensions) can be practical. arXiv preprint arXiv:1703.00893 (2017)
18. Holland, M.J.: Distribution-robust mean estimation via smoothed random perturbations. arXiv preprint arXiv:1906.10300 (2019)
19. Hopkins, S.B.: Mean estimation with sub-Gaussian rates in polynomial time. Ann. Stat. **48**(2), 1193–1213 (2020)
20. Hsu, D., Sabato, S.: Loss minimization and parameter estimation with heavy tails. J. Mach. Learn. Res. **17**, Paper No. 18, 40 (2016)
21. Huber, P.J., Ronchetti, E.M.: Robust Statistics. Wiley Series in Probability and Statistics, 2nd edn. Wiley, Hoboken (2009)
22. Jerrum, M.R., Valiant, L.G., Vazirani, V.V.: Random generation of combinatorial structures from a uniform distribution. Theoret. Comput. Sci. **43**(2–3), 169–188 (1986)
23. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Springer, Berlin (2011)
24. Koltchinskii, V.: Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Stat. **34**(6), 2593–2656 (2006)
25. Komatu, Y.: Elementary inequalities for Mills' ratio. Rep. Stat. Appl. Res. Un. Jap. Sci. Eng. **4**, 69–70 (1955)
26. Le Cam, L., Yang, G.L.: Asymptotics in Statistics. Springer Series in Statistics, 2nd edn. Springer, New York (2000). Some basic concepts
27. Lecué, G., Lerasle, M.: Learning from mom's principles: Le cam's approach. Stoch. Process. Appl. **129**(11), 4385–4410 (2019)
28. Lecué, G., Mendelson, S.: Learning sub-Gaussian classes: upper and minimax bounds. arXiv preprint arXiv:1305.4825 (2013)
29. Lecué, G., Lerasle, M.: Robust machine learning by median-of-means: theory and practice. Ann. Stat. **48**(2), 906–931 (2020)
30. Ledoux, M.: The concentration of measure phenomenon. Mathematical Surveys and Monographs, vol. 89. American Mathematical Society, Providence, RI (2001)
31. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Classics in Mathematics. Springer, Berlin (2011). Isoperimetry and processes, Reprint of the 1991 edition
32. Lei, Z., Luh, K., Venkat, P., Zhang, F.: A fast spectral algorithm for mean estimation with sub-Gaussian rates. In: Conference on Learning Theory, pp. 2598–2612 (2020)
33. Lerasle, M., Szabó, Z., Mathieu, T., Lecué, G.: Monk outlier-robust mean embedding estimation by median-of-means. In: International Conference on Machine Learning, pp. 3782–3793 (2019)
34. Lerasle, M., Szabó, Z., Mathieu, T., Lecué, G.: MONK outlier-robust mean embedding estimation by median-of-means. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pp. 3782–3793. PMLR (2019)
35. Li, W.V., Kuelbs, J.: Some shift inequalities for Gaussian measures. In: High dimensional probability (Oberwolfach, 1996), volume 43 of Progr. Probab., pp. 233–243. Birkhäuser, Basel (1998)
36. Lugosi, G., Mendelson, S.: Near-optimal mean estimators with respect to general norms. Probab. Theory Relat. Fields **175**(3–4), 957–973 (2019)
37. Lugosi, G., Mendelson, S.: Sub-gaussian estimators of the mean of a random vector. Ann. Stat. **47**(2), 783–794 (2019)
38. Lugosi, G., Mendelson, S.: Sub-Gaussian estimators of the mean of a random vector. Ann. Stat. **47**(2), 783–794 (2019)

39. Lugosi, G., Mendelson, S.: Risk minimization by median-of-means tournaments. J. Eur. Math. Soc. **22**(3), 925–965 (2020)
40. Lugosi, G., Mendelson, S.: Robust multivariate mean estimation: the optimality of trimmed mean (2020)
41. Mendelson, S.: "Local" vs. "global" parameters–breaking the Gaussian complexity barrier. Ann. Stat. **45**(5), 1835–1862 (2017)
42. Mendelson, S.: On multiplier processes under weak moment assumptions. In: Geometric Aspects of Functional Analysis, volume 2169 of Lecture Notes in Math., pp. 301–318. Springer, Cham (2017)
43. Minsker, S.: Uniform bounds for robust mean estimators. Technical report, Department of Mathematics, University of Southern California (2019)
44. Minsker, S.: Geometric median and robust estimation in Banach spaces. Bernoulli **21**(4), 2308–2335 (2015)
45. Nazin, A.V., Nemirovsky, A.S., Tsybakov, A.B., Juditsky, A.B.: Algorithms of robust stochastic optimization based on mirror descent method. Autom. Remote Control **80**(9), 1607–1627 (2019)
46. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication. Wiley, New York (1983). Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics
47. Pisier, G.: The Volume of Convex Bodies and Banach Space Geometry. Cambridge Tracts in Mathematics, vol. 94. Cambridge University Press, Cambridge (1989)
48. Talagrand, M.: 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
49. Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, New York (2009). Revised and extended from the 2004 French original, Translated by Vladimir Zaiats
50. van der Vaart, A.W.: Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3. Cambridge University Press, Cambridge (1998)
51. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. In Measures of complexity, pp. 11–30. Springer, Cham: Reprint of Theor. Probab. Appl. **16**(1971), 264–280 (2015)
52. Vapnik, V.N.: The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science, 2nd edn. Springer, New York (2000)
53. Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (2018)