Concentration study of M-estimators using the influence function

Timothée Mathieu

Electronic Journal of Statistics
Vol. 16 (2022) 3695–3750
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2030

Concentration study of M-estimators using the influence function

Timothée Mathieu*

INRIA, Scool team. Univ. Lille, CRIStAL, CNRS
e-mail: timothee.mathieu@inria.fr

Abstract: We present a new finite-sample analysis of M-estimators of locations in a Hilbert space using the tool of the influence function. In particular, we show that the deviations of an M-estimator can be controlled thanks to its influence function (or its score function) and then, we use concentration inequality on M-estimators to investigate the robust estimation of the mean in high dimension in a corrupted setting (adversarial corruption setting) for bounded and unbounded score functions. For a sample of size n and covariance matrix Σ, we attain the minimax speed $\sqrt{Tr(\Sigma)/n} + \sqrt{\|\Sigma\|_{op}\log(1/\delta)/n}$ with probability larger than $1-\delta$ in a heavy-tailed setting. One of the major advantages of our approach compared to others recently proposed is that our estimator is tractable and fast to compute even in very high dimension with a complexity of $O(nd\log(Tr(\Sigma)))$ where n is the sample size and Σ is the covariance matrix of the inliers and in the code that we make available for this article is tested to be very fast.

MSC2020 subject classifications: Primary 62F35; secondary 60G25.
Keywords and phrases: Robust Statistics, concentration inequalities, mean estimation.

Received October 2021.

1. Introduction

Recently, there have been an increase in the quantity and dimensionality of data one has to investigate in Machine Learning tasks. Big datasets are rather difficult to handle because it is not possible anymore to check by hand some gigabyte or terabyte of data to search for possibly abnormal data, and the task is even more difficult in high dimension. An answer to this problem is robust statistics and in particular, because mean estimators are used everywhere in Machine Learning (for empirical risk minimization, cross validation, data preprocessing, feature engineering...), it has become critical to understand in depth robust mean estimators.

In this article, we try to understand one type of mean estimators: M-estimators. We show that they are nearly optimal in the heavy-tailed setting and we exhibit an algorithm that allows us to compute these estimators in a linear number of steps. Furthermore, the algorithm is very fast in practice. See Section 7.3 for more illustrations of our algorithms.

*This work was done while the author was in Laboratoire Mathématiques d'Orsay for his Ph.D.

The analysis of M-estimators to estimate the mean is divided in two steps: first we analyze the deviations of M-estimators around $T(P)$, which is some location parameter meant to exhibit a central tendency of the data and then we bound the bias $\|T(P) - \mathbb{E}[X]\|$. Let $X \sim P$ for some P probability on a Hilbert space $\mathcal{H}$, let ρ be an increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, we are interested in estimating the location parameter $T(P)$ defined by

$$T(P) \in \arg\min_{\theta \in \mathcal{H}} \mathbb{E}\left[\rho(\|X - \theta\|)\right] = 0, \tag{1}$$

where $\|\cdot\|$ is a norm associated to a scalar product. Alternatively, if ρ is smooth enough (which will be the case in this article), we define $T(P)$ by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi(\|X - T(P)\|)\right] = 0, \tag{2}$$

where $\psi = \rho'$ is called the score function. The empirical estimator obtained by plugging the empirical density $\widehat{P}_n$ in equation (2) is called M-estimator associated with ψ, it is denoted $T(X_1^n)$ and computed from a sample $X_1, \ldots, X_n$ using the following equation:

$$\sum_{i=1}^{n} \frac{X_i - T(X_1^n)}{\|X_i - T(X_1^n)\|}\psi(\|X_i - T(X_1^n)\|) = 0. \tag{3}$$

A particular case of $T(P)$ is obtained when choosing $\psi(x) = x$ in which case $T(P) = \mathbb{E}[X]$ and $T(X_1^n) = \frac{1}{n}\sum_{i=1}^n X_i$, however it is well known that the empirical mean is not robust. A careful choice of the function ψ yield estimators that are more robust to outliers and to heavy-tailed data (see [5]).

The subsequent problem is to see how the properties of ψ impact the robustness and efficiency of $T(X_1^n)$ when estimating $T(P)$ or $\mathbb{E}[X]$. To study the robustness of $T(X_1^n)$ we use the influence function, a common tool in robust statistics. The influence function is used to quantify the robustness of an estimator, see for example [19, 20, 27, 36] in which are derived properties such as the asymptotic variance or the breakdown point of the estimator $T(X_1^n)$ using the influence function. The influence function is the Gâteaux derivative of T evaluated in the Dirac distribution in a point $x \in \mathcal{H}$ and in the case of M-estimators in $\mathbb{R}^d$, from [20, Eq 4.2.9 in Section 4.2C.], the influence function takes the following simple form:

$$\mathrm{IF}(x, T, P) = \lim_{t \to 0} \frac{T((1-t)P + t\delta_x) - T(P)}{t} = M_{P,T}^{-1} \frac{x - T(P)}{\|x - T(P)\|}\psi(\|x - T(P)\|), \tag{4}$$

where $M_{P,T}$ is a non-singular matrix whose explicit formula is not important for most of our application because of our choice of ψ function (an explicit formula can however be found in the proof of Theorem 3).

The general idea is that, if the estimator is smooth enough (for example if it is Fréchet or Hadamard differentiable, see [17]), then one can write the following

expansion

$$T(P) = T(Q) + \int_{\mathcal{H}} \mathrm{IF}(x, T, Q) d(P - Q)(x) + R(P, Q), \tag{5}$$

where the remainder term $R(P, Q)$ is controlled. For example, if we apply equation (5) to $Q = \widehat{P}_n$ the empirical distribution, the influence function provides a first order approximation for the difference between the estimator $T(\widehat{P}_n) = T(X_1^n)$ and its limit $T(P)$. This technique of approximating the estimator by its influence function is also linked to the Bahadur decomposition, see [2] and [22] for applications to M-estimators. The influence function of M-estimators is usually chosen bounded in robust statistics, in particular from [18, 27] we have that if ψ is bounded, then the influence function is bounded and T is qualitatively robust (i.e. the estimator $T(X_1^n)$ is equi-continuous, c.f. [27]) and have asymptotic breakdown point 1/2. On the other hand if ψ is unbounded, then $T(X_1^n)$ is not qualitatively robust, the influence function is not bounded and the asymptotic breakdown point is zero. From Hampel's Theorem [27, Theorem 2.21] we also have that ψ is bounded if and only if T is a continuous functional with respect to the Levy metric. More generally, the influence function has been used in a lot of works on asymptotic robustness, see [20, 27] or [19, 36].

The influence function has also been used recently in Machine Learning literature in order to have a model selection tool specialized in robustness, see for example [11], [28] and the closely related tool of leave one out error [16]. The field of Robustness in Machine Learning has been very active in the last few years, in particular after several works by Olivier Catoni and co-authors in [5, 6], the goal is to prove non-asymptotic deviation bounds when the data are more heavy-tailed than what is usually considered in classical Machine Learning. This line of thought has been continued in a number of articles, in particular [13] introduced some general concept of sub-Gaussian estimators that have been then used successfully in other applications, see [9, 14, 5, 31, 39, 35]. See also some comprehensive lecture notes on the subject in [30].

It is interesting to note that contrary to works from classical robust theory from the 70's, the influence functions of the M-estimators used by Catoni are not necessarily bounded. In this article, we initiate the analysis of the effect of unbounded influence function on the robustness of M-estimators, Huber [26] told us that the influence function must be bounded while Catoni uses unbounded influence function and he still shows robust properties for this type of estimator, the difference is in their vision of what is a robust estimator.

There will be three parts in our analysis of this problem, first we extend Catoni's non-asymptotic analysis of M-estimators as we analyze M-estimators with more general influence functions and in a multivariate setting using the properties of the influence function, making a link between the deviations of the influence function and the deviations of $T(X_1^n)$. Second we apply our theory to three specific M-estimators for which we show tight upper bound on the rate of convergence to the mean. Finally, we investigate an algorithm to compute $T(X_1^n)$ and we show that this algorithm converges in a reasonable number of steps.

More precisely, In Section 3, we show that concentration inequalities for M-estimators derive from concentration inequalities on the influence function by showing roughly that

$$\|T(X_1^n) - T(P)\| \simeq \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - T(P)}{\|X_i - T(P)\|} \psi(\|X_i - T(P)\|) \right\|. \tag{6}$$

From equation (4), the right hand side of equation (6) can be interpreted as the deviation of the influence function. The right hand side of equation (6) can be controlled under classical assumptions. For example in $\mathbb{R}$, if ψ is bounded by $\beta > 0$ (Huber estimator), we can use Hoeffding or Bernstein inequality to get a control on $\|T(X_1^n) - T(P)\|$. Using Hoeffding inequality, we obtain a concentration rate similar to the rate of the empirical mean on Gaussian data.

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \psi(X_i - T(P)) \right| \geq \frac{\beta}{\sqrt{n}} \lambda \right) \leq e^{-2\lambda^2}.$$

Remark that using Bernstein inequality, we don't need to have ψ bounded, hence this gives us a mean to show sub-Gaussian rates for M-estimators with unbounded ψ function.

In Section 4, we provide bounds on the bias $\|T(P) - \mathbb{E}[X]\|$ and on the variance terms in the concentration inequality from Section 3. Bounding the bias has often been a problem in robust statistics, if the distribution is skewed and the bias is not controlled we can only say that we estimate a quantity meant to quantify a central tendency of P but we do not estimate $\mathbb{E}[X]$ directly. However in statistical learning for example, estimating the mean is not just an arbitrary choice and we don't want to estimate a central tendency of the dataset, we want to estimate its mean. In this article, we give explicit bounds on the bias and we use those bounds (in Section 5) to give concentration results on $T(X_1^n)$ around $\mathbb{E}[X]$ in the context of heavy-tailed and adversarially corrupted datasets even beyond the L^2 case. Indeed, we show that the L^2 assumption is needed only to be able to handle the bias $\|T(P) - \mathbb{E}[X]\|$ but on the other hand if the distribution is symmetric, $T(P) = \mathbb{E}[X]$ and we can obtain sharp deviation bounds even in the case of L^1 distributions. Bounds on the bias are already present for the specific case Huber estimator in regression in [37]. Our bound on the bias has the same flavor as [37, Proposition 1] but applied to multivariate location parameter for more general M-estimators. Similarly to the discussion in [37], our analysis leads us to the study of a bias-variance trade-off for M-estimator depending on the value of β.

In this context, in Section 5, we show that $T(X_1^n)$ is suitable to estimate the mean in high dimension in a heavy-tailed and corrupted setting (even though our estimators are not minimax in corrupted setting). In the literature, there are estimators that have strong theoretical guarantees but that are intractable, for example one can see estimators based on the aggregation of one-dimensional estimators (same idea as projection pursuits), see [30, Theorem 44] and reference therein, see also [32] and there has also been estimators based on depth, for

example Tukey's median [7]. On the other hand, there are tractable algorithms but with non minimax optimal rates of convergence for example the coordinate-wise median or the geometrical median [33, 6], our work belong to this type of methods, our estimator is easily computable and even though the obtained error bounds are much better than for the coordinate-wise median, at least in corrupted setting it is not minimax. Recently there have been several propositions of algorithms whose goal was to be at the same time tractable and minimax, see [15, 40, 23, 12, 24, 8] however these algorithms are often hard to implement and in practice the complexity makes them intractable for high-dimensional problems.

In a corrupted setting where inliers have a q^{th} finite moment, for $q > 2$, we control the deviations of M-estimators. Let us suppose that the data are an adversarial ε_n corrupted dataset (see Assumption 1). Then, under some assumptions on ψ, for all $0 < t \lesssim n\frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}}$, with probability larger than $1 - 8\exp(-t)$,

$$\|\mathbb{E}[X] - T(X_1^n)\| \lesssim \frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{op} t}}{\sqrt{n}} \bigvee \mathbb{E}\left[\|X - \mathbb{E}[X]\|^q\right]^{1/q} \varepsilon_n^{1-1/q}.$$

See Proposition 2 below for the formal and more precise statement in the case of Huber's estimator, and see Section 5 for other examples in particular for M-estimators with unbounded score function. In the heavy-tailed setting where $\varepsilon_n = 0$, the bound is almost minimax optimal, the difference is that t is allowed to be up to $n\frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}}$ instead of the usual $t \lesssim n$. This estimator constitutes one of the few tractable and efficient estimators of the multivariate mean in heavy-tailed setting.

The error due to corruption is of order $O(\mathbb{E}\left[\|X - \mathbb{E}[X]\|^q\right]^{1/q} \varepsilon_n^{1-1/q})$, the term $\varepsilon_n^{1-1/q}$ is optimal (see [34, Lemma 5.4]) but the factor $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^q\right]^{1/q}$ is sub-optimal. In Section 6 we show that the dependency in the dimension due to the $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^q\right]^{1/q}$ is unavoidable (to understand the magnitude of this term, one may think about the case where all the marginal have same law, then we can show that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^q\right]^{1/q} \lesssim \sqrt{d}$). We can't avoid this dimension dependence for our estimator contrary to [12, 29] in which the authors show that the optimal error due to corruption is $O(\varepsilon\sqrt{\|\Sigma\|_{op}})$ in the case of Gaussian distributions. Hence our estimator is not optimal with respect to corruption.

Finally, in Section 7, we exhibit an algorithm to compute $T(X_1^n)$ and we show that this algorithm converges in a finite number of steps resulting in a complexity of order $O(nd\log(\mathrm{Tr}(\Sigma)))$ where n is the sample size and Σ is the covariance matrix of the inliers (our analysis is only valid for $d \ll e^n$). This algorithm is practically efficient, it is illustrated in Section 7.3 and we invite the interested reader to check out the github repository https://github.com/TimotheeMathieu/RobustMeanEstimator for the python code.

2. Setting and notations

2.1. Setting

In all the article, we consider $X_1, \dots, X_n$ in a Hilbert space $\mathcal{H}$ that have been corrupted by an adversary using the following process.

Assumption 1. *There exist $X'_1, \dots, X'_n \in \mathcal{H}$ i.i.d. following a law P that have been modified by an "adversary" to obtain $X_1, \dots, X_n$. The adversary can modify at most $|\mathcal{O}|$ points such that there exist $\mathcal{I}, \mathcal{O}$ partition of $\{1, \dots, n\}$, with $\mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$ and $|\mathcal{O}| < |\mathcal{I}|$.*

The points $(X_i)_{i\in\mathcal{O}}$ are arbitrary and are called outliers, the points $(X_i)_{i\in\mathcal{I}}$ are called inliers. Remark that the statistician do not know the sets $\mathcal{I}, \mathcal{O}$ and that although we have $X_i = X'_i$ for any $i \in \mathcal{I}$, the sample $(X_i)_{i\in\mathcal{I}}$ is in general not i.i.d. because the adversary can choose which data are corrupted using knowledge on the inliers. For instance it is possible that the adversary decided to corrupt the $|\mathcal{O}|$ points from $X'_1, \dots, X'_n$ that were the closest to the theoretical mean in which case the $(X_i)_{i\in\mathcal{I}}$ are not independent.

We consider the functional T defined by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi(\|X - T(P)\|)\right] = 0, \tag{7}$$

for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, the existence and unicity of $T(P)$ is discussed in Lemma 2. We are interested in the behavior of the associated M-estimator $T(X_1^n)$ defined by

$$\sum_{i=1}^{n} \frac{X_i - T(X_1^n)}{\|X_i - T(X_1^n)\|}\psi(\|X_i - T(X_1^n)\|) = 0, \tag{8}$$

where ψ is a function that satisfies the following properties.

Assumption 2. *The function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ verifies the following properties:*

(i) ψ *is continuous and differentiable almost everywhere*
(ii) $\psi(0) = 0$
(iii) ψ *is concave*
(iv) There exist $\beta, \gamma > 0$ such that

$$\forall x \geq 0, \qquad 1 \geq \psi'(x) \geq \gamma \mathbb{1}\{x \leq \beta\},$$

where $\mathbb{1}$ is the indicator function.

When we want to emphasize the dependency in β, we will use the notation ψ_1 to be such that for all $x \geq 0$, $\psi(x) = \beta\psi_1(x/\beta)$.

Because ψ is concave, non decreasing and not identically zero, there are always a couple of positive constants β, γ such that Assumptions 2 holds. For our results to hold we will ask that β and γ are not too small. A first result that can

be derived from Assumptions 2 and some additional assumptions is that our problem is well defined. This comes from the fact that the problem is a convex problem.

Let

$$Z_{P,\psi} : \theta \mapsto \mathbb{E}_{X\sim P}\left[\frac{X-\theta}{\|X-\theta\|}\psi\left(\|X-\theta\|\right)\right],$$

we have the following lemma whose proof is in Section C.1.

Lemma 1. *Let ψ satisfy Assumptions 2, let $u \in S$ and $\theta \in \mathcal{H}$,*

$$u^T Jac(Z_{P,\psi})(\theta)u \leq -\mathbb{E}\left[\psi'\left(\|X-\theta\|\right)\right]$$

where Jac denotes the Jacobi operator.

Using the previous lemma, we can prove using additional hypotheses that the problem is well defined.

Lemma 2. *Let ψ satisfy Assumptions 2, define $\rho : x \mapsto \int_0^x \psi(t)dt$ and let X satisfy $\mathbb{E}[\rho(\|X-\mathbb{E}[X]\|)] < \rho(\beta)$, then $T(P)$ defined by equation (7) exists and is unique.*

This lemma is proven in Section C.2. Remark that the condition $\mathbb{E}[\rho(\|X-\mathbb{E}[X])\|)] < \rho(\beta)$ is implied by $\mathrm{Tr}(\Sigma) = \mathbb{E}[\|X-\mathbb{E}[X]\|^2] < 2\rho(\beta)$ because $\rho(x) \leq x^2/2$ however an hypothesis on $\mathrm{Tr}(\Sigma)$ is a lot stronger because it supposes a finite second moment. In the whole article, we will suppose that $T(P)$ is unique, we do not necessarily suppose that the assumptions of Lemma 2 hold because they are not minimal assumptions for unicity and existence of $T(P)$. However, for simplicity, we will suppose that the following condition is verified.

Assumption 3. *X satisfies $\mathbb{E}[\rho(\|X-\mathbb{E}[X]\|)] < \min(\rho(\beta/3), \psi(\beta/2)^2/2)$.*

Assumption 3 means that β should be large enough to encompass most of the weight of X around the expectation. When ψ is bounded, this is only a first moment assumption. This assumption is due to limitations in our methods and we believe that this is not necessary. One can show that $\rho(x) \geq \psi(x)^2/2$ which may be used to simplify Assumption 3.

Assumptions 2 and Assumptions 3 will be supposed true. The behavior of ψ at 0 allows us to control the deviations of the estimator using the influence function, see Section 3 and it is also important to control the bias of the resulting estimator, see Section 4. On the other hand, the growth rate of ψ at $+\infty$ is central to derive concentration bounds of $T(X_1^n)$, as will become clear all along Section 3 and 5. Assumptions 2 do not always apply to M-estimators, for example the sample median is not an estimator derived from a function ψ satisfying these assumptions. On the other hand, we provide three examples of score functions satisfying Assumptions 2, with three different growth rates when x goes to infinity.

Huber's estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_H(x) = x\,\mathbb{1}\{x \leq \beta\} + \beta\,\mathbb{1}\{x > \beta\}. \tag{9}$$

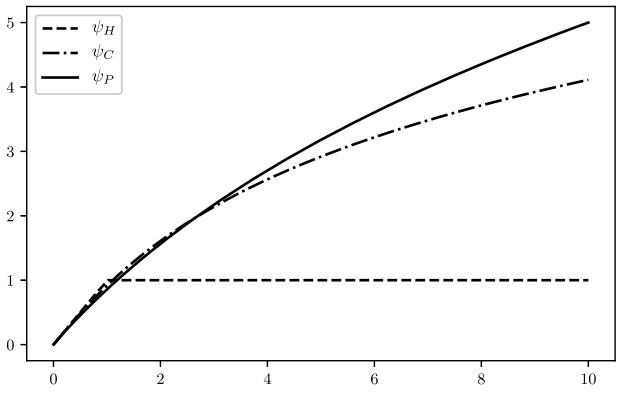


FIG 1. *Plot of ψ_H and ψ_C for $\beta = 1$. ψ_P is plotted for $\beta = 10$ and $p = 5$.*

In dimension 1, the M-estimator constructed from this score function is called Huber's estimator [26].

Catoni's estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{1}{2}\left(\frac{x}{\beta}\right)^2\right). \tag{10}$$

The associated M-estimator is one of the estimators considered by Catoni in [5]. We call the resulting M-estimator Catoni's estimator.

Polynomial estimator. Let $p \in \mathbb{N}^*$, $\beta > 0$. For all $x \geq 0$, let

$$\psi_P(x) = \frac{x}{1 + \left(\frac{x}{\beta}\right)^{1-1/p}}. \tag{11}$$

We call Polynomial estimator the M-estimator obtained using this score function.

The following result shows that the score functions from the previous three examples satisfy Assumptions 2.

Lemma 3. *For all $x \geq 0$, we have $\psi_H'(x) = \mathbb{1}\{x \leq \beta\}$, $\psi_C'(x) \geq \frac{4}{5}\mathbb{1}\{x \leq \beta\}$, $\psi_P'(x) \geq \frac{1}{4}\mathbb{1}\{x \leq \beta\}$.*

The proof of Lemma 3 is postponed to Section C.3.

2.2. Notations

Let $\mathcal{P}$ denote the set of probability distributions on $\mathcal{H}$, $S = \{x \in \mathcal{H} : \|x\| = 1\}$ where $\|\cdot\|$ is a norm associated with a scalar product $\langle\cdot,\cdot\rangle$ on $\mathcal{H}$. For any $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, let $\mathcal{P}_\psi = \{P \in \mathcal{P} : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$. We denote $a \lesssim b$ if there exists a numerical constant $C > 0$ such that $a \leq Cb$.

Let T_H, T_C and T_P denote the functionals associated to the score functions ψ_P,ψ_H and ψ_C respectively, using Equation (7). Define the following variance terms

$$V_\psi = \mathbb{E}[\psi(\|X - T_H(P)\|)^2], \quad \text{and}$$
$$v_\psi = \left\| \mathbb{E}\left[\frac{(X - T_H(P))(X - T_H(P))^T}{\|X - T_H(P)\|^2} \psi(\|X - T_H(P)\|)^2 \right] \right\|_{op}.$$

These variance terms are to be compared with $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ in the multivariate Gaussian setting, Hanson-Wright inequality (see Equation 18) tells us that $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ describe the spread of the empirical mean in high dimension. Here we are not in a Gaussian setting and for example in the case of Huber's estimator, V_{ψ_H} and v_{ψ_H} will describe the spread of the influence function of Huber's estimator. See Section 4.2 for a more formal study of the link between V_{ψ_H}, v_{ψ_H} and $Tr(\Sigma), \|\Sigma\|_{op}$.

3. Tail probabilities of M-estimator and influence function

The main result of the paper compares the tail probabilities of $\|T(X_1^n) - T(P)\|$ and those of its influence function.

Definition 1. *We call t_T and t_{IF} the tail probability functions defined for all $\lambda > 0$ and $\theta \in \mathcal{H}$ by*

$$t_T(\lambda; \theta, X_1^n) := \mathbb{P}\left(\|T(X_1^n) - \theta\| \geq \lambda\right) \tag{12}$$
$$t_{IF}(\lambda; \theta, X_1^n) := \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \frac{X_i - \theta}{\|X_i - \theta\|} \psi(\|X_i - \theta\|) \right\| \geq \lambda \right). \tag{13}$$

The main theorem of Section 3 is the following.

Theorem 1. *Suppose that Assumptions 1, 2 and 3 hold. If moreover $V_\theta = \mathbb{E}_P[\psi(\|X' - \theta\|)^2] \leq \psi(\beta/2)^2/2 < \infty$ and $|\mathcal{O}| \leq n\gamma/8$, then for all $\lambda \in (0, \beta/2)$ and for all $\theta \in \mathcal{H}$,*

$$t_T(\lambda; \theta, X_1^n) \leq t_{\mathrm{IF}}\left(\lambda\gamma/4; \theta, X_1^n\right) + e^{-n\gamma^2/32}. \tag{14}$$

The proof of this result is given in Section A.1. In the heavy-tailed setting, we will use Theorem 1 with $\theta = T(P)$ in order to have a small value of $t_{\mathrm{IF}}(\lambda; \theta, X_1^n)$. On the other hand, in a corrupted setting, θ will be set to $T(P)$ where P is the law of inliers. For now, there is no hypothesis on the outliers in $\mathcal{O}$, in what follows we will see that if ψ is unbounded, we will need some hypothesis on $X_i, i \in \mathcal{O}$ in order to have a control of the value of t_{IF}.

Remark that although M-estimators with bounded ψ are proven to have a breakdown point of $1/2$ (see [27]), our result is only valid for a proportion of outliers $|\mathcal{O}_n|/n \leq \gamma/8$. This is an artifact of the proof and with more stringent condition on β, we could allow for a higher breakdown point at the cost of a more complicated analysis. Remark however that when the corruption is close

to 1/2, the error due to corruption is the principal source of error and a looser deviation analysis could then be allowed to reach a higher breakdown point.

Example: Huber's estimation in dimension $d = 1$. In the case of Huber's estimator, ψ is bounded and from Lemma 3, $\gamma = 1$. Because ψ_H is bounded by β, we have directly from Bernstein inequality for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\text{sign}(X - T_H(P))\psi_H(|X - T_H(P)|)\right| > \sqrt{\frac{2V_{\psi_H}t}{n}} + \frac{\beta t}{n}\right) \leq 2e^{-t}.$$

Then, by Theorem 1, if $V_{\psi_H} \leq \beta^2/8$, for all $t > 0$ such that $4\sqrt{2V_{\psi_H}t/n} + 4\beta t/n \leq \beta/2$,

$$\mathbb{P}\left(|T_H(X_1^n) - T_H(P)| > 4\sqrt{\frac{2V_{\psi_H}t}{n}} + 4\frac{\beta t}{n}\right) \leq 2e^{-t} + e^{-n/32}. \tag{15}$$

Remark that choosing $\beta = \sqrt{V_{\psi_H}n}$ gives us a sub-Gaussian concentration around $T_H(P)$, this is similar to the concentrations inequalities introduced in [13] except that we concentrate around $T_H(P)$ instead of $\mathbb{E}[X]$. Remark also that the condition $V_{\psi_H} \leq \beta^2/8$ is rather weak because we already have $V_{\psi_H} \leq \beta^2$, the condition asks that there is enough weight in the interval $[-\beta, \beta]$.

4. Bias and variance of M-estimators when considered as estimators of the mean

If P is symmetric then we can avoid the problem of the bias and $T(P) = \mathbb{E}[X]$, but unfortunately in the case of skewed distribution the bias $\|T(P) - \mathbb{E}[X]\|$ can be very large and the choice of β will determine how large the bias is. In this section, we show how the bias behaves as β grows, we also provide bounds on the variance terms defined in Section 2.2, these bounds will be useful to derive concentration inequalities on $T(X_1^n)$. We will use the notation $\psi(x) = \beta\psi_1(x/\beta)$ when we want to emphasize the dependency on β.

We introduce the following function

$$Z_\beta : \theta \mapsto \mathbb{E}\left[\beta\frac{(X-\theta)}{\|X-\theta\|}\psi_1\left(\frac{\|X-\theta\|}{\beta}\right)\right].$$

The following theorem links Z_β with the distance between $T(P)$ and $\mathbb{E}[X]$.

Theorem 2. *Let X be a random vector in $\mathcal{H}$, $X \sim P$ with finite expectation and suppose Assumptions 2 and 3. Then we have $\|\mathbb{E}[X] - T(P)\| \leq \frac{2}{\gamma}\|Z_\beta(\mathbb{E}[X])\|$.*

We postpone the proof to Section A.2. From Theorem 2, it is sufficient to upper bound $\|Z_\beta(\mathbb{E}[X])\|$ to get a bound on the bias.

4.1. Bias of M-estimators

We begin with the bias of Huber's estimator obtained from equation (3) with $\psi = \psi_H$. The following lemma gives a bound on the bias of Huber's estimator for a distribution with a finite number of finite moments.

Lemma 4. *Let X be a random variable with $\mathbb{E}[\|X\|^q] < \infty$ for $q \in \mathbb{N}^*$ and suppose that $\rho_1(1/3) \geq \mathbb{E}[\rho_1(\|X - \mathbb{E}[X]\|/\beta)]$. Then*

$$\|\mathbb{E}[X] - T_H(P)\| \leq \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}.$$

The proof is in Section C.4. Lemma 4 is not exactly tight as can be seen for instance if we do the computation with the Pareto distribution for $d = 1$ for which if the shape parameter is $\alpha = 2$, we have only one finite moment but can show that in this parametric case, Huber's estimator achieves rates $1/\beta$.

The choice of β is a very important problem when estimating $\mathbb{E}[X]$ using $T(X_1^n)$ and in particular we will need to choose β carefully as a function of n in order to have $T(X_1^n)$ that converges to $\mathbb{E}[X]$. The choice of β will entail a sort of bias-variance tradeoff. Remark that we do not need a finite second moment for our analysis to work, we only need $\mathbb{E}[\rho(\|X - T(P)\|/\beta)] < \infty$ which translates in a finite first moment in the case of $\psi = \psi_H$.

In addition to Lemma 4 we can also show an exponential bound on the bias when the random variable X is sub-exponential however because the primary use of Huber's estimator is with robust statistics, we only state the result for a finite number of finite moments as it is what will interest us. An interested reader can adapt the proof to lighter-tailed distributions.

For a ψ function that is not Huber's score function, the bias also depends strongly on the behavior of ψ near 0.

Lemma 5. *Suppose that ψ is k times differentiable with bounded k^{th} derivative and that Assumptions 2 and 3 hold, $\psi'(0) = 1$ and for $2 \leq j \leq k-1$, $\psi^{(j)}(0) = 0$. Let X be a random variable such that $\mathbb{E}[\|X\|^k] < \infty$, then,*

$$\|T(P) - \mathbb{E}[X]\| \leq \frac{\|\psi_1^{(k)}\|_\infty}{\gamma k! \beta^{k-1}} \mathbb{E}\left[\|X - \mathbb{E}[X]\|^k\right]. \tag{16}$$

Moreover, if X follows a Bernoulli distribution of parameter p, this bound is tight in its dependency in β. When $\beta \to \infty$, we have

$$Z_\beta(\mathbb{E}[X]) = \psi_1^{(k)}(0) \frac{p(1-p)^k - (1-p)p^k}{k!\beta^{k-1}} + o\left(\frac{1}{\beta^{k-1}}\right)$$

This Lemma is proven in Section C.5. For example, we can show that for Catoni's score function $\psi(x) = \log(1 + x + x^2/2)$ whose second derivative is $\psi''(x) = -(x + x^2/2)/(1 + x + x^2/2)^2$, we have that $\psi(x) = x - x^3/6 + o(x^3)$ and then the bias of Catoni's estimator is in general of order $1/\beta^2$. Lemma 5 shows that the bias depends on the smoothness of the function near 0 and also the number of finite moments.

4.2. Bound on the variance of M-estimators

First, we have to control the variability of $T(X_1^n)$ in order to control its deviations. The following lemma gives an upper bound on both V_ψ and v_ψ defined in Section 2.2.

Lemma 6. *Suppose that Assumptions 2 and 3 are satisfied, suppose that X has a finite second moment with covariance operator Σ we have that $V_\psi \leq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] = \mathrm{Tr}(\Sigma)$, and $v_\psi \leq \|\Sigma\|_{op} + \|\mathbb{E}[X] - T(P)\|^2$.*

Lemma 6 (proven in Section C.6) gives a control on V_ψ and v_ψ using the properties of X. Next we show that Lemma 6 is tight in the case of Huber's estimator as long as X is sufficiently concentrated using the following lemma whose proof is provided in Section C.7.

Lemma 7. *Suppose that Assumptions 2 and 3 are satisfied and that X is such that $\mathbb{E}[\|X\|^{2q}] < \infty$ for some $q > 1$, then*

$$V_{\psi_H} \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}}.$$

and

$$V_{\psi_H} \geq \|\Sigma\|_{op} - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}}$$

Lemmas 6 and 7 imply that if X has enough moments, say with 4 finite moments, and if β is sufficiently large, then the behavior of the variance term is the same as the variance term for the empirical mean. On the other hand, if X is not very concentrated, Lemma 6 can be a very rough bound and in the case of Huber estimator if X has only a finite first moment but no finite variance, then V_{ψ_H} and V_{ψ_H} are finite even though $Tr(\Sigma) = \|\Sigma\|_{op} = \infty$.

5. Application to the concentration of M-estimators around the mean in corrupted datasets

In this section, we investigate the concentration of the three M-estimators taken as example in this article in a corrupted, heavy-tailed setting. The goal will be to recover deviations similar to the one we would have in a Gaussian setting, but when the data are not Gaussian. The gold standard in this context is the deviation of the empirical mean in a Gaussian setting (see [4]). If $X_1, \dots, X_n$ are i.i.d from $\mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$, then for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \sigma\sqrt{\frac{t}{2n}}\right) \leq e^{-t}. \tag{17}$$

An equivalent of this in the multi-dimensional setting is Hanson-Wright inequality [21]: let $X \sim \mathcal{N}(\mu, \Sigma)$ for Σ a positive definite matrix, $\mu \in \mathbb{R}^d$. Then, for any $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\|^2 > \frac{2Tr(\Sigma)}{n} + \frac{9t\|\Sigma\|_{op}}{n}\right) \leq e^{-t}. \tag{18}$$

This form of Hanson-Wright inequality can be found for example in [30]. Our aim is to obtain deviations similar to the ones in equations (17) and (18) but in a non-Gaussian setting.

The results we show in this section are not optimal, they are nearly optimal in the heavy-tailed setting but the effect due to corruption is sub-optimal compared to [12]. One of our goals is to illustrate the use of the influence function and particularly Theorem 1 for an easy derivation of concentration inequalities for M-estimators. We also illustrate an interesting phenomenon derived from Theorem 1 by showing that the concentration of $T(X_1^n)$ around $T(P)$ can be much faster than the concentration of $T(X_1^n)$ around $\mathbb{E}[X]$, because the variance term is not $Tr(\Sigma)$ but V which can be a lot smaller than $Tr(\Sigma)$ (for instance if $Tr(\Sigma)$ is not finite). It also shows that $T(X_1^n)$ is never arbitrarily bad as long as $\mathbb{E}[\psi(\|X\|)] < \infty$ contrary to the empirical mean. We use the following corollary of [1, Theorem 4] recalled in Section D, this lemma is proven in Section C.8.

Lemma 8. *Let $Y_1, \dots, Y_n$ be i.i.d random variables taking values in $\mathcal{H}$, centered and with covariance operator Σ, and such that the Orlicz norm of Y is finite:*

$$\|Y\|_{\psi_1} = \inf\left\{\lambda > 0 : \mathbb{E}\left[\exp\left(\|Y\|/\lambda\right) - 1\right] \leq 1\right\} < \infty.$$

There exists an universal constant $C > 0$ such that, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Y_i\right\| \geq \frac{3}{2}\sqrt{\mathbb{E}\left[\|Y\|^2\right]}\, n + 2\sqrt{nt\|\Sigma\|_{op}} + Ct\left\|\max_{1\leq i\leq n}\|Y_i\|\right\|_{\psi_1}\right) \leq 4\exp\left(-t\right). \tag{19}$$

The last term in equation (19) can be handled using [38, Lemma 2.2.2] from which we get that there exists an absolute constant $K > 0$ such that

$$\left\|\max_{1\leq i\leq n}\|Y_i\|\right\|_{\psi_1} \leq K\log(n)\left\|Y_i\right\|_{\psi_1}. \tag{20}$$

However, note that Hanson-Wright's inequality for Gaussian random variables shows that this logarithm factor is not optimal. This extra logarithm factor can be removed if Y is bounded, which will be the case when we apply this result to Huber's estimator but not for Catoni's estimator.

In the rest of the section, we prove concentration inequalities for the estimators featured in Section 2 using Lemma 8 applied to $Y = \frac{X-T(P)}{\|X-T(P)\|}\psi(\|X - T(P)\|)$ and using the bounds on the bias from Section 4. For simplicity, we will not keep track of all the constants and we will give the names C_1, C_2, C_3 to numerical constants that do not depend on any of the parameters of the model (in particular they do not depend on P or β).

5.1. Huber's estimator

Let $\beta > 0$, and, for all $x \geq 0$, let $\psi_H(x) = x\,\mathbb{1}\{x \leq \beta\} + \beta\,\mathbb{1}\{x > \beta\}$. From Lemma 3, Assumptions 2 hold in this example with $\gamma = 1$. As ψ_H is bounded

by β, Hoeffding's lemma (see [4, Section 2.3]) shows that $\max_{1\leq i\leq n}\|\psi_H(\|X_i - T_H(P)\|)\|_{\psi_1} \leq \beta$. From Lemma 8, we have the following proposition, whose proof is given Section B.1.

Proposition 1. *Suppose that Assumption 1 and Assumption 3 are verified with $|\mathcal{O}| \leq n/32$. Then, there exist some numerical constants $C_1, C_2, C_3 > 0$ such that if $\beta^2 \geq V_{\psi_H}\max(8, C_1/n)$, then for any $t \leq C_2 n$, with probability larger than $1 - 4e^{-t} - e^{-n/32}$, we have*

$$\|T_H(X_1^n) - T_H(P)\| \leq 6\sqrt{\frac{V_{\psi_H}}{n}} + 8\sqrt{\frac{tv_{\psi_H}}{n}} + \frac{C_3 t\beta}{n} + 8\beta\varepsilon_n,$$

where $\varepsilon_n = |\mathcal{O}|/n$ is the proportion of outliers.

Remark that the condition on β can be simplified if needed, using Lemma 6, to $\beta^2 \geq 8Tr(\Sigma)$. The second step is to choose the value of β, the choice of β will be a tradeoff between the bias term from Lemma 4 and the concentration in Proposition 1

Proposition 2. *Suppose the same assumptions as in Proposition 1 and suppose additionally that $\mathbb{E}_P[\|X - \mathbb{E}_P[X]\|^q] < \infty$ for some $q \geq 2$. Then, there exist $C_1, C_2, C_3 > 0$ numerical constants such that for all $t > 0$ with*

$$t \leq C_1 n \min\left(\frac{\mathbb{E}_P[\|X - \mathbb{E}_P[X]\|^q]}{Tr(\Sigma)^{q/2}}, 1\right),$$

we have with probability larger than $1 - 4e^{-t} - e^{-n/32}$ that

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{Tr(\Sigma)}{n}} + 8\sqrt{\frac{t\|\Sigma\|_{op}}{n}} + C_2\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$$

where β is fixed to the value

$$\beta^q = \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{4\varepsilon_n + C_3 t/n}.$$

In particular, if $q > 2$ and $t \leq C_1 n \frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{\frac{1}{q-1}}}$ then we have the following bound on the deviations.

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{Tr(\Sigma)}{n}} + 9\sqrt{\frac{t\|\Sigma\|_{op}}{n}} + C_2\varepsilon_n^{1-1/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}. \tag{21}$$

When $\varepsilon_n = 0$ the previous proposition guarantees an optimal sub-Gaussian rate. Notice that ε_n is multiplied by a quantity that increases with the dimension in general, this bound is not minimax, see [12] which achieve a sharper bound in

the Gaussian setting. We see that the dependency on ε_n is $O(\varepsilon_n^{1-1/q})$, this type of bound is already present for example in [25] and the power $1-1/q$ is optimal (see [34, Lemma 5.4]) however, the $\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}$ factor in front of it is not optimal, we show in Section 6 that this factor is unavoidable for M-estimators. On the other hand, when $\varepsilon_n = 0$, we obtain sub-Gaussian rates of convergence as soon as $t \lesssim n$.

For Equation (21) to hold, we must have $t \leq C_1 n \frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{\frac{1}{q-1}}}$. Remark that this is dependent on the dimension, we could have stated a similar deviation bound under the alternative condition $t \leq O\left(n^{\frac{q-2}{2q-2}}\right)$, avoiding the dimension dependence at the price of a worse dependency on n. For simplicity we did not state it in the proposition.

Remark that in $\mathbb{R}^d$, with respect to the dimension, $\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}$ behaves similarly to $\sqrt{Tr(\Sigma)}$. Indeed, if $\Sigma = I_d$, we have $\sqrt{Tr(\Sigma)} = \sqrt{d}$ and on the other hand, by Jensen's inequality, if we denote by $X^{(i)}$ be the i^{th} coordinate of X,

$$\begin{aligned}\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q} &= \mathbb{E}\left[\left(\sum_{i=1}^d (X^{(i)}-\mathbb{E}[X^{(i)}])^2\right)^{q/2}\right]^{1/q} \\ &\leq d^{1/2-1/q}\left(\sum_{i=1}^d \mathbb{E}\left[|X^{(i)}-\mathbb{E}[X^{(i)}]|^q\right]\right)^{1/q}.\end{aligned}$$

The dependency in the dimension is similarly of order $\sqrt{d}$.

Finally, we present the symmetric case for which there is no need for Lemma 4 because we have right-away that $T_H(P) = \mathbb{E}_P[X]$ and this simplifies the computations. In particular we only need a finite first moment and we can directly pick the minimal value of β in Proposition 1 to get,

Proposition 3. *Suppose the same assumptions as in Proposition 1 and moreover, suppose that P is symmetric with $\mathbb{E}_P[\|X\|] < \infty$. Then, there exist some numerical constants $C_1, C_2, C_3 > 0$, such that for any $t \leq C_1 n$, with probability larger than $1-4e^{-t}-e^{-n/32}$, we have*

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{V_{\psi_H}}{n}} + 8\sqrt{\frac{t v_{\psi_H}}{n}} + C_2\mathbb{E}[\|X-\mathbb{E}[X]\|]\left(\frac{t}{n}+\varepsilon_n\right). \tag{22}$$

where $\varepsilon_n = |\mathcal{O}|/n$ is the proportion of outliers and $T_H, V_{\psi_H}, v_{\psi_H}$ are computed with $\beta = C_3\mathbb{E}[\|X-\mathbb{E}[X]\|]$.

We see with Proposition 3 that we can relax the Gaussian inliers assumption made in [12] to inliers that are symmetric or inliers with infinite number of finite moments and still we have a linear dependency in ε_n. We see also that this bound does not need a finite second moment, the law only need to be symmetric and have a finite first moment.

5.2. Catoni's estimator

In the case of Catoni's estimator and Polynomial estimator, we will only prove a proposition similar to Proposition 2 but we could also state the equivalents of Propositions 1 and 3 using the same reasoning.

Let $\beta > 0$ and, for all $x \geq 0$, let $\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{x^2}{2\beta^2}\right)$. From Lemma 3, ψ_C satisfies Assumptions 2 with $\gamma = 4/5$. Lemma 8 in addition to Theorem 1 can be used to obtain the following proposition.

Proposition 4. *Suppose Assumption 1 and Assumption 3 are verified with $\mathbb{E}_P[\|X\|^2] < \infty$, suppose that there exist some constants $C_O > 0$ and $\delta_O \in (0,1)$ such that with probability larger than $1 - \delta_O$, we have*

$$\frac{1}{|\mathcal{O}|}\sum_{i\in\mathcal{O}} \psi_C(\|X_i - T_C(P)\|) + \psi_C(\|X_i' - T_C(P)\|) \leq \beta C_O.$$

Fix the value of β to

$$\beta^3 = \frac{5\mathbb{E}[\|X - \mathbb{E}[X]\|^3]}{120C_O\varepsilon_n + Ct\log(n)/n},$$

and suppose $|\mathcal{O}| \leq n/(20C_O)$.

Then, there exist some numerical constants $C_1, C_2, C_3 > 0$ such that for any $t \leq C_2 \frac{n}{\log(n)^4}\frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^3]^{1/3}}$, we have with probability larger than $1 - 4e^{-t} - e^{-n/50} - \delta_O$,

$$\|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq \frac{15}{2}\sqrt{\frac{Tr(\Sigma)}{n}} + 11\sqrt{\frac{t\|\Sigma\|_{op}}{n}} + C_2\varepsilon_n^{2/3}\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}.$$

Proposition 4 gives results that are similar to Huber's estimator in Proposition 2 using stronger assumptions on the corruption. If $\mathbb{E}[\|X\|^3]$ is finite, we can use similar arguments to show a bound of order $O(\varepsilon^{2/3})$ the interested reader could adapt the proof of Proposition 2. Contrary to Proposition 2, our method does not allow us to go further than the 3^{rd} order because of the bias bound from Lemma 5.

The condition on outliers is not very unusual. Indeed, if we suppose that the outliers are i.i.d with law P_O and such that $\mathbb{E}_{P_O}[\psi_C(\|X - T_C(P)\|)] < \infty$, we can use Chebychev inequality to say that for any $C_O > 0$

$$\mathbb{P}\left(\forall i \in \mathcal{O}, \quad \psi_C(\|X_i - T_C(P)\|) \leq C_O\beta\right) \geq \left(1 - \frac{\mathbb{E}_{P_O}[\psi_C(\|X - T_C(P)\|)]}{C_O\beta}\right)^{|\mathcal{O}|}$$

and then, we can choose C_O such that the right-hand-side is strictly positive. Remark that we only suppose a finite moment for $\psi_C(\|X - T_C(P)\|)$, which is a logarithmic moment, this is a rather mild requirement on the outliers. On the other hand, if there is a fixed number of outliers, i.e. $\varepsilon_n = C/n$, we deduce from Proposition 4 that the speed of convergence does not deteriorate if the outliers are bounded almost surely by $\exp(\sqrt{nV_{\psi_C}})$ in which case $\delta_O = 0$. This is also a rather mild requirement on outliers.

5.3. Polynomial estimator

Finally, we look at the polynomial estimator defined for $p, \beta > 0$ by

$$\psi_P(x) = \frac{x}{1 + (x/\beta)^{1-1/p}}.$$

Proposition 5. *Suppose Assumption 1 and Assumption 3 are verified with law P that verifies $\mathbb{E}_P[\|X\|^2] < \infty$ and its covariance matrix is denoted Σ. We suppose that there exist some constants $C_O \geq 1$ and $\delta_O \in (0,1)$ such that with probability $1 - \delta_O$, we have*

$$\frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \psi_P(\|X_i - T_P(P)\|) + \psi_P(\|X_i' - T_P(P)\|) \leq C_O \beta.$$

We fix the value of β to $\beta^2 = Tr(\Sigma)$. Then, there exist some numerical constants $C_1, C_2, C_3 > 0$ such that for any $t \leq C_1 n$, with probability larger than $1 - e^{-C_2 t} - e^{-n/512} - \delta_O$, we have

$$\|T_P(X_1^n) - \mathbb{E}[X]\| \leq 16\sqrt{65\frac{Tr(\Sigma)t}{n}} + 16\varepsilon_n C_O \sqrt{Tr(\Sigma)}.$$

where $\varepsilon_n = |\mathcal{O}|/n$ is the proportion of outliers, supposed positive, and the parameter p is fixed to $p = C_3 t$.

The proof of this proposition can be found Section B.5 In this proposition we see that we obtain weaker guarantees for the polynomial estimator because the t in the right-hand-side of the bound is multiplied by $Tr(\Sigma)$ as compared to the smaller $\|\Sigma\|_{op}$ factor that we had before, hence this bound is not optimal. Nonetheless, this result is valid with very high probability and this may be surprising to the reader but recall that p is considered a parameter that we tuned using the level t so that in fact we use a function ψ_P that gets very close to a bounded function when t gets large.

6. Lower bound in corruption bias for M-estimators

In this section we suppose ψ bounded and $\mathcal{H} = \mathbb{R}^d$. Define $P_\varepsilon = (1-\varepsilon)P + \varepsilon Q$ for some outlier probability Q and some $\varepsilon \in (0, 1/2)$. The goal is to estimate the error we would incur if we want to estimate the expectation of P using data from P_ε. Remark that this setting is more restrictive than the adversarial corruption setting we used until now, indeed we can see P_ε as a corrupted setting in which the adversary chose to modify a random number of randomly chosen outliers.

Theorem 3. *Suppose $P = \mathcal{N}(0, \sigma^2 I_d)$. There exist a distribution Q and $\varepsilon_{\max} > 0$ such that for all $\varepsilon \in (0, \varepsilon_{\max})$,*

$$\|T(P) - T(P_\varepsilon)\| = \|T(P_\varepsilon)\| \geq \sigma \frac{\varepsilon\sqrt{d-2}}{12}$$

where $P_\varepsilon = (1-\varepsilon)P + \varepsilon Q$.

Theorem 3 (proven in Section A.3) gives us a lower bound on the bias due to the corruption. Remark that a similar result already existed for the case of the geometric median, see [29, Proposition 2.1]. Our result extends the result of [29] to more general ψ functions using an alternative proof.

To make the link with the corrupted setting from Assumption 1, remark that if an adversary chose $\lfloor(1-\varepsilon)n\rfloor$ inliers i.i.d from P and $\lceil\varepsilon n\rceil$ outliers i.i.d from Q, then the resulting empirical distribution converges to P_ε and by continuity of T with respect to the weak topology (see Hampel's theorem in [27]), we have that $T(X_1^n) \to T(P_\varepsilon)$.

Then, remark that we have

$$\|T(X_1^n) - \mathbb{E}_P[X]\| \geq \|T(P_\varepsilon) - \mathbb{E}[X]\| - \|T(X_1^n) - T(P_\varepsilon)\|. \tag{23}$$

In the right-hand-side of (23), we have the term $\|T(P_\varepsilon) - T(P)\| = \|T(P_\varepsilon) - \mathbb{E}[X]\|$ that stays bounded by below for any value of β from Theorem 3 and the term $\|T(X_1^n) - T(P_\varepsilon)\|$ that goes to 0 as n goes to infinity. More precisely, we have the following corollary of Theorem 3 consequence of the consistency of the plug-in estimator $T(\widehat{P}_n) = T(X_1^n)$ for any distribution P (see [27]).

Corollary 1. *Suppose $P = \mathcal{N}(\mu, \sigma^2 I_d)$. There exist a distribution Q and $\varepsilon_{\max} > 0$ such that for all $\varepsilon \in (0, \varepsilon_{\max})$ and if an adversary chose $\lfloor(1-\varepsilon)n\rfloor$ inliers i.i.d from P and $\lceil\varepsilon n\rceil$ outliers i.i.d from Q to form a sample $X_1, \ldots, X_n$, we have*

$$\mathbb{P}\left(\|T(X_1^n) - \mathbb{E}_P[X]\| \geq \sigma\frac{\varepsilon\sqrt{d-2}}{24}\right) \xrightarrow[n\to\infty]{} 1.$$

It is possible to quantify the rates in Corollary 1 but this is not really necessary as this already proves that we can't hope to achieve a rate that does not depend on d in the corruption error, i.e. we can't attain the minimax rates of convergence which is in this case of order $O(\sigma\varepsilon)$.

7. M-estimators in practice

In this section, we give results for $\mathcal{H} = \mathbb{R}^d$ but they could be extended to more general Hilbert spaces provided that one use a sufficiently accurate initialization instead of the coordinate median used here.

7.1. Algorithm and convergence using iterative re-weighting

To compute $T(X_1^n)$, we use an iterative re-weighting algorithm. This algorithm is rather well known to compute M-estimators, see [27, Section 7] and it has already been extensively studied. The principle is to rewrite the definition of $T(X_1^n)$ from equation (3) as

$$T(X_1^n)\sum_{i=1}^n \frac{\psi\left(\|X_i - T(X_1^n)\|\right)}{\|X_i - T(X_1^n)\|} = \sum_{i=1}^n X_i \frac{\psi\left(\|X_i - T(X_1^n)\|\right)}{\|X_i - T(X_1^n)\|},$$

then, denote $w_i = \frac{\psi(\|X_i - T(X_1^n)\|)}{\|X_i - T(X_1^n)\|}$, we get an expression of $T(X_1^n)$ as a weighted sum:

$$T(X_1^n) = \sum_{i=1}^n X_i \frac{w_i}{\sum_{i=1}^n w_i}.$$

The weights w_i depend on $T(X_1^n)$ and the principle of the algorithm is as follows. Initialize θ_0 with the coordinate-wise median and iterate the following

$$w_i^{(m)} = \frac{\psi\left(\|X_i - \theta^{(m)}\|\right)}{\|X_i - \theta^{(m)}\|} \qquad \text{,and} \qquad \theta^{(m+1)} = \sum_{i=1}^n X_i \frac{w_i^{(m)}}{\sum_{i=1}^n w_i^{(m)}}.$$

We show that this algorithm allows us to find a minimizer of

$$J_n(\theta) = \frac{1}{n} \sum_{i=1}^n \rho\left(\|X_i - \theta\|\right).$$

Let $r_n, \delta > 0$ be such that

$$\mathbb{P}\left(\|T(X_1^n) - \mathbb{E}_P[X]\| \geq r_n\right) \leq \delta,$$

for instance, one can use the bound given in Section 5. We have the following theorem.

Theorem 4. *Let $X_1, \ldots, X_n$ be in the $\mathcal{I} \cup \mathcal{O}$ setting with $(X_i)_{i \in \mathcal{I}}$ i.i.d with law P whose variance is finite and the covariance matrix is denoted Σ. Suppose also that $|\mathcal{O}| \leq n/8$ and $\beta \geq 2\sqrt{2Tr(\Sigma)} + r_n + \psi^{-1}\left(\sqrt{2V_\psi}\right)$. Then, for all $\mathbb{N} \in \mathbb{N}$, with probability larger than $1 - (d+4)e^{-n/8} - \delta$, we have*

$$\|\theta^{(m)} - T(X_1^n)\| \leq \frac{1}{(1+\gamma/2)^m} \|\theta^{(0)} - T(X_1^n)\| \leq \frac{1}{(1+\gamma/2)^m} \left(2\sqrt{2Tr(\Sigma)} + r_n\right).$$

Said differently, the iterative reweighting algorithm is such that for any ε, we have $\|\theta^{(m)} - T(X_1^n)\| \leq \varepsilon$ after a number of iterations

$$m \geq \log\left(\frac{2\sqrt{2Tr(\Sigma)} + r_n}{\varepsilon}\right) / \log(1 + \gamma/2).$$

The proof of Theorem 4 can be found in Section A.4. To prove this theorem, we use techniques similar to those used to prove the convergence of Weiszfeld's Method (see [3]).

We obtain an exponential rate of convergence. Remark that because the objective function is convex, even if the initialization was not as good as the coordinate-wise median or if β was not large enough, we would converge nonetheless but with a linear rate of convergence (similar to convergence rates in [3]) until we are close enough to $T(X_1^n)$ for the rate to be exponential.

7.2. Discussion on the choice of β

The choice of β is a frequent problem when using Huber estimator. One solution is to use Lepski's method but this is computationally expensive and not always efficient, another (often used) approach is to use a heuristic for β based on the median absolute deviation by saying that β must be of order σ the standard deviation of the inliers however in Section 5 we see that this choice is very conservative and β would often be too small if it is estimated using the median absolute deviation.

In view of Section 5 where we see that depending on the number of finite moments, we may want to choose β between $\sqrt{Tr(\Sigma)}$ and $\sqrt{Tr(\Sigma)n}$, we propose to choose β in the interval $[0, \text{MAD}\sqrt{n}]$ where $\text{MAD} = \text{Med}\left(\|X_i - \text{GMed}(X_1^n)\|\right)$ (GMED being the geometric median).

We propose the following Heuristic to choose β:

$$\widehat{\beta} = \arg\min_{\beta \in [0, \text{MAD}\sqrt{n}]} \frac{\widehat{V_{\psi_\beta}}}{n} + C_\psi \frac{\text{MAD}^4}{\beta^2} + (0.05\beta)^2$$

where $\widehat{V_{\psi_\beta}} = \frac{1}{n}\sum_{i=1}^n \psi_\beta(\|X_i - T_{\psi_\beta}(X_1^n)\|)^2$ and C_ψ is a constant that depend on ψ as described by the bounds on the bias (Section 4, $C_{\psi_H} = 1$, $C_{\psi_C} = 5/32$ and $C_{\psi_P} = 1/16$). This is a bias-variance trade-off, the first term converges to the asymptotic variance of $T_{\psi_\beta}(X_1^n)$, the second term is a bound on the squared bias and the third term is a bound on the corruption bias if we suppose that $\varepsilon_n \leq 0.05$ (In the Robust literature, it is often said that there is less than 5 or 10 percent of outliers).

Remark that the objective function may not be convex and hence there can be local minima, we restrict the search space $[0, \text{MAD}\sqrt{n}]$ in order to be able to choose β efficiently using a grid-search.

7.3. Illustrations

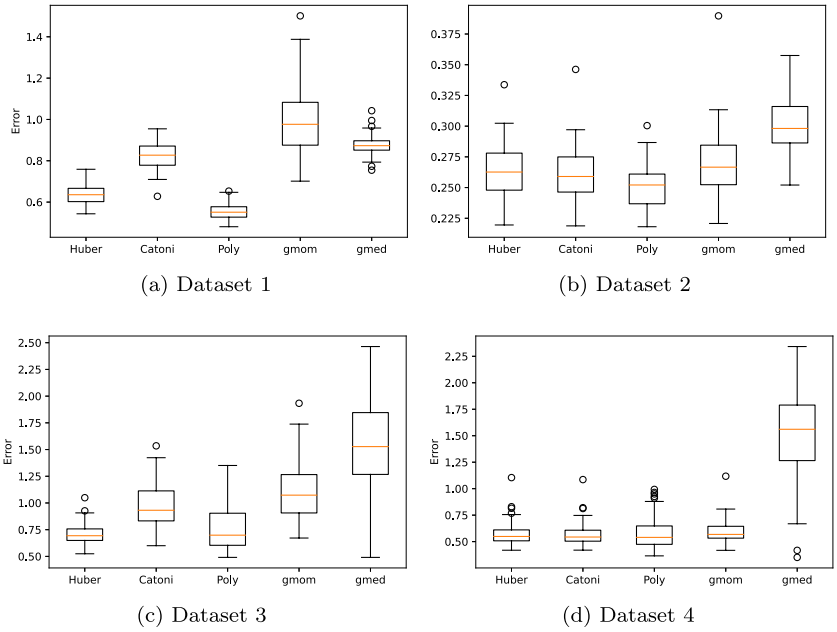
To illustrate the behavior of M-estimators in heavy-tailed setting we consider a multivariate Pareto law for which the coordinates are drawn, independently of each other, from a Pareto distribution and a multivariate student distribution. All of the dataset present a finite variance but infinite third moment.

Dataset 1: Coordinates drawn from a Pareto with shape parameter $\alpha = 2.1$ and scale parameter 1, with two corrupted samples situated in $300\mathbf{1}_d$ where $\mathbf{1}_d$ is the vector with all coordinates equal to 1.

Dataset 2: Coordinates drawn from a Pareto with shape parameter $\alpha = 3$ and scale parameter 1. Uncorrupted dataset.

Dataset 3: Data i.i.d from a mixture $0.4\mathcal{T}(0, 2.1)+0.6\mathcal{T}(2\mathbf{1}_d, 2.1)$, where $\mathcal{T}(\mu, \nu)$ is the multivariate student distribution with mean μ and degree of freedom ν. In this dataset, there is also two corrupted samples situated in $300\mathbf{1}_d$.

Dataset 4: Data i.i.d from the mixture $0.4\mathcal{T}(0, 3) + 0.6\mathcal{T}(2\mathbf{1}_d, 3)$. Uncorrupted dataset.



(a) Dataset 1 (b) Dataset 2

(c) Dataset 3 (d) Dataset 4

FIG 2. *Boxplots of the error* $\|\widehat{\mu} - \mathbb{E}[X]\|$ *of various estimators* $\widehat{\mu}$ *using* 100 *independent copies of the three dataset considered.*

In these three datasets, we consider $n = 1000$ samples and the dimension is $d = 100$.

We consider 5 estimators : Huber and Catoni estimator with β chosen using Section 7.2, the polynomial estimator with $p = 5$ and β chosen using Section 7.2, the geometric median (gmed) and the geometric median of means described in [33] with $k = 9$ blocks. The result is represented in Figure 2

In Figure 2, M-estimators are outperforming the geometric median by quite a lot because the geometric median is very biased when estimating the mean of an asymmetric distribution. The geometric median of means is closer in performance to M-estimators but it is not as good, maybe because there is no adaptivity the number of blocks have been fixed to 9.

Remark that the multivariate robust mean estimators described in [15, 12, 24, 8] are not presented here because they are either too computationally intensive or too hard to implement for the purpose of comparison. Remark also that the choice of β from Section 7.2 is heuristic. In a learning setting, one may prefer to use cross-validation to tune β using directly the learning criterion.

Appendix A: Proof of Theorems

A.1. Proof of Theorem 1

Hypotheses in this proof. *(i)* $V_\theta = \mathbb{E}_P[\psi(\|X' - \theta\|)^2] \leq \psi(\beta/2)^2/2 < \infty$ *(ii)* $|\mathcal{O}| \leq n\gamma/8$

For all $n \in \mathbb{N}^*$, $\lambda \in \mathbb{R}, \theta \in \mathcal{H}$ and $\theta \neq T(X_1^n)$, let

$$f_n(\lambda;\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\langle X_i - \theta - \lambda u_n(\theta), u_n(\theta)\rangle}{\|X_i - \theta - \lambda u_n(\theta)\|}\psi(\|X_i - \theta - \lambda u_n(\theta)\|),$$

where $u_n(\theta) = \frac{\theta - T(X_1^n)}{\|\theta - T(X_1^n)\|}$.

Step 1. For any $\lambda > 0$, $\mathbb{P}(\|\theta - T(X_1^n)\| \geq \lambda) \leq \mathbb{P}(f_n(\lambda;\theta) \geq 0)$.

Proof. For all $\theta \in \mathcal{H}$, let $J_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho(\|X_i - \theta\|)$ from Lemma 1 applied to the empirical density $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, we have for any $u \in S$ and $\theta \in \mathcal{H}$,

$$u^T Hess(J_n)(\theta)u \geq \frac{1}{n}\sum_{i=1}^{n} \psi'\left(\|X_i - \theta\|\right).$$

In particular, $f_n(\lambda;\theta) = -\langle \nabla J_n(\theta - \lambda u_n(\theta)), u_n(\theta)\rangle$ and if we take the derivative of f_n with respect to λ, we have the following equation

$$\begin{aligned} f_n'(\lambda;\theta) &= -u_n(\theta)^T Hess(J_n)(\theta - \lambda u_n(\theta))u_n(\theta) \\ &\leq -\frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - \theta - \lambda u_n(\theta)\|). \end{aligned} \tag{24}$$

where $f_n'(\lambda;\theta)$ is the derivative of $f_n(\lambda;\theta)$ with respect to λ. Then, because ψ' is non-negative (Assumption 2-(iv)), the function $f_n(\,\cdot\,;\theta)$ is non-increasing. Hence, for all $n \in \mathbb{N}^*$, $\lambda \in \mathbb{R}$

$$\|\theta - T(X_1^n)\| \geq \lambda \Rightarrow f_n(\|\theta - T(X_1^n)\|;\theta) = 0 \leq f_n(\lambda;\theta).$$

And then,

$$\mathbb{P}(\|\theta - T(X_1^n)\| \geq \lambda) \leq \mathbb{P}(f_n(\lambda;\theta) \geq 0). \tag{25}$$

□

Step 2. For all $\lambda > 0$,

$$f_n(\lambda;\theta) \leq f_n(0) - \lambda \inf_{t\in[0,\lambda]} |f_n'(t;\theta)|.$$

Proof. We apply Taylor's inequality to the function $\lambda \mapsto f_n(\lambda;\theta)$. As $\lambda \mapsto f_n(\lambda;\theta)$ is non-increasing (because its derivative is non-positive, see Equation (24)), we get the result. □

Step 3. Let $m = \gamma \mathbb{P}\left(\|X_i' - \theta\| \leq \beta/2\right)$. With probability than $1 - e^{-nm^2/8}$, we have

$$\inf_{t\in[0,\lambda]} |f_n'(\lambda;\theta)| \geq \frac{3m}{4} - \frac{|\mathcal{O}|}{n}.$$

Proof. From equation (24),

$$\begin{aligned}
|f_n'(t;\theta))| \geq & \frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - \theta - tu_n(\theta)\|) \\
\geq & \frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i' - \theta - tu_n(\theta)\|) \\
& + \frac{1}{n}\sum_{i\in\mathcal{O}} \left(\psi'(\|X_i - \theta - tu_n(\theta)\|) - \psi'(\|X_i' - \theta - tu_n(\theta)\|)\right).
\end{aligned} \tag{26}$$

Hence, because $\psi' \in [0,1]$, we have

$$\begin{aligned}
\inf_{t\in[0,\lambda]} |f_n'(t;\theta))| &\geq \inf_{t\in[0,\lambda]} \frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i' - \theta - tu_n(\theta)\|) - \frac{|\mathcal{O}|}{n} \\
&\geq \inf_{t\in[0,\lambda]} \frac{\gamma}{n}\sum_{i=1}^{n} \mathbb{1}\{\|X_i' - \theta - tu_n(\theta)\| \leq \beta\} - \frac{|\mathcal{O}|}{n} \\
&\geq \frac{\gamma}{n}\sum_{i=1}^{n} \mathbb{1}\{\|X_i' - \theta\| \leq \beta/2\} - \frac{|\mathcal{O}|}{n}
\end{aligned}$$

The last line is a consequence of $\lambda \in (0, \beta/2)$. Then, this is a sum of bounded random variables and by Hoeffding's inequality, with probability larger than $1 - e^{-2n\varepsilon^2}$, we have

$$\sup_{t\in[0,\lambda]} -|f_n'(\lambda;\theta)| \leq -\gamma \mathbb{E}\left[\mathbb{1}\{\|X_i' - \theta\| \leq \beta/2\}\right] + \varepsilon + \frac{|\mathcal{O}|}{n}.$$

take $\varepsilon = \gamma \mathbb{P}\left(\|X_i' - \theta\| \leq \beta/2\right)/4$ to conclude □

Step 4. For $\lambda \in (0, \beta/2)$, $\mathbb{P}(\|\theta - T(X_1^n)\| \geq \lambda) \leq t_{\text{IF}}\left(\lambda\gamma/4; \theta, X_1^n\right) + e^{-n\gamma^2/32}$.

Proof. For any $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}(\|\theta - T(X_1^n)\| \geq \lambda) &\overset{(a)}{\leq} \mathbb{P}(f_n(\lambda;\theta) \geq 0) \\
&\overset{(b)}{\leq} 1 - \mathbb{P}\left(f_n(0;\theta) - \lambda \inf_{t\in[0,\lambda]} |f_n'(t;\theta)| \leq 0\right) \\
&\overset{(c)}{\leq} 1 - \mathbb{P}\left(f_n(0;\theta) \leq \lambda\left(\frac{3}{4}m - \frac{|\mathcal{O}|}{n}\right)\right) + e^{-nm^2/8} \\
&= t_{\text{IF}}\left(\lambda\left(\frac{3}{4}m - \frac{|\mathcal{O}|}{n}\right); \theta, X_1^n\right) + e^{-nm^2/8},
\end{aligned} \tag{27}$$

where (a) follows from Step 1, (b) follows from Step 2 and (c) follows from Step 3. Finally, we bound m because ψ is increasing on $[0, \beta]$, for any $\lambda < \beta/2$,

$$\mathbb{P}\left(\|X_i' - \theta\| \leq \beta/2\right) = \mathbb{P}\left(\psi(\|X_i' - \theta\|) \leq \psi(\beta/2)\right).$$

As $V_\theta = \mathbb{E}[\psi(\|X - \theta\|)^2] < \infty$, by Markov's inequality, it follows that

$$m = \gamma \mathbb{P}\left(\|X_i' - \theta\| \leq \beta/2\right) \geq \gamma\left(1 - \frac{V_\theta}{\psi(\beta/2)^2}\right). \tag{28}$$

From Hypothesis (i), $V_\theta \leq \psi(\beta/2)^2/2$, we get $m \geq \gamma/2$. Plugging this bound in equation (27), we get

$$\mathbb{P}(\|\theta - T(X_1^n)\| \geq \lambda) \leq t_{\mathrm{IF}}\left(\lambda\left(3\gamma/8 - \frac{|\mathcal{O}|}{n}\right); \theta, X_1^n\right) + e^{-n\gamma^2/32}. \tag{29}$$

we conclude using that $|\mathcal{O}| \geq n\gamma/8$ (from Hypothesis (ii)). □

A.2. Proof of Theorem 2

Step 1. We have

$$\begin{aligned}&\|Z_\beta(\mathbb{E}[X]) - Z_\beta(T(P))\| \\ &\quad \geq \gamma\|\mathbb{E}[X] - T(P)\|\mathbb{P}\big(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|\big).\end{aligned}$$

Proof. The function $\theta \mapsto Z_\beta(\theta)$ is differentiable and by the mean value theorem, we have

$$\begin{aligned}&\|Z_\beta(\mathbb{E}[X]) - Z_\beta(T(P))\| \\ &\quad \geq \|\mathbb{E}[X] - T(P)\| \inf_{t\in[0,1]} \|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op}\end{aligned} \tag{30}$$

Where Jac denotes the Jacobi operator. From Lemma 1 and Assumption 2-(iv), we get

$$\|Jac(Z_\beta)(\theta)\|_{op} \geq \mathbb{E}\left[\psi_1'\left(\left\|\frac{X - \theta}{\beta}\right\|\right)\right] \geq \gamma\mathbb{P}\left(\|X - \theta\| \leq \beta\right).$$

Hence, for all $t \in [0, 1]$,

$$\begin{aligned}&\|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op} \\ &\quad \geq \gamma\mathbb{P}\left(\|X - t\mathbb{E}[X] - (1-t)T(P)\| \leq \beta\right) \\ &\quad \geq \gamma\mathbb{P}\left(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|\right)\end{aligned}$$

inject this in Equation (30) to get the result. □

Step 2. $\mathbb{P}\left(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|\right) \geq \frac{1}{2}$.

Proof. Use the following lemma proven in Section C.9.

Lemma 9. *If* $\rho(1/3) \geq \mathbb{E}[\rho(\|X - \mathbb{E}[X]\|/\beta)]$, *then* $\|\mathbb{E}[X] - T(P)\| \leq \frac{\beta}{3}$.

We get,

$$\begin{aligned}
&\mathbb{P}\left(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|\right) \\
&\quad \geq \mathbb{P}\left(\|X - T(P)\| \leq 2\beta/3\right) \\
&\quad = \mathbb{P}\left(\rho\left(\frac{\|X - T(P)\|}{\beta}\right) \leq \rho(2/3)\right) \\
&\quad \geq \mathbb{P}\left(\rho\left(\frac{\|X - T(P)\|}{\beta}\right) \leq 2\rho(1/3)\right)
\end{aligned}$$

because ρ is increasing and super-additive on $\mathbb{R}_+$ (ρ is increasing because $\psi(0) = 0$ and ψ is non-decreasing because $\psi' \geq 0$, hence $\psi = \rho' \geq 0$). Hence, by Markov's inequality and using the hypothesis,

$$\begin{aligned}
&\mathbb{P}\left(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|\right) \\
&\qquad \geq 1 - \frac{1}{2\rho(1/3)}\mathbb{E}\left[\rho\left(\frac{\|X - T(P)\|}{\beta}\right)\right] \geq \frac{1}{2}
\end{aligned}$$

□

Step 3. $\|\mathbb{E}[X] - T(P)\| \leq \frac{2}{\gamma}\left\|Z_\beta(\mathbb{E}[X])\right\|$.

Proof. Use Step 1 and Step 2 □

A.3. Proof of Theorem 3

P is symmetric around 0, hence $T(P) = 0$. By definition of the influence function we have

$$\mathrm{IF}(x) = \lim_{\varepsilon \to 0} \frac{T(P_\varepsilon) - T(P)}{\varepsilon}$$

where $P_\varepsilon = (1-\varepsilon)P + \varepsilon\delta_x$ and we have, from [20] that $\mathrm{IF}(x) = M^{-1}\frac{x}{\|x\|}\psi(\|x\|)$, where

$$M = -\mathbb{E}\left[\left(\frac{I_d}{\|X\|} - \frac{XX^T}{\|X\|^3}\right)\psi(\|X\|) + \frac{XX^T}{\|X\|^2}\psi'(\|X\|)\right].$$

Hence, taking $\|x\|$ such that $\psi(\|x\|) \geq \|\psi\|_\infty/2$ (this fixes the norm of x) and $\|M^{-1}x\| = \|M^{-1}\|_{op}\|x\|$ (this fixes the direction of x), we have

$$\|\mathrm{IF}(x)\| \geq \frac{\|M^{-1}\|_{op}}{2}\|\psi\|_\infty \geq \frac{1}{2\|M\|_{op}}\|\psi\|_\infty \tag{31}$$

where $\|M\|_{op}$ is the operator norm of M. Let us control this operator norm. We have for all $u \in S$ where S is the sphere in $\mathbb{R}^d$,

$$u^T M u = -\mathbb{E}\left[\left(\frac{1}{\|X\|} - \frac{\langle X, u\rangle^2}{\|X\|^3}\right)\psi(\|X\|) + \frac{\langle X, u\rangle^2}{\|X\|^2}\psi'(\|X\|)\right]$$

Hence,

$$|u^T M u| \leq 3\mathbb{E}\left[\frac{\psi(\|X\|)}{\|X\|}\right] \leq 3\|\psi\|_\infty \mathbb{E}\left[\frac{1}{\|X\|}\right] \leq 3\frac{\|\psi\|_\infty}{\sigma}\mathbb{E}\left[\frac{\sigma^2}{\|X\|^2}\right]^{1/2}$$

Then, use that $X \sim \mathcal{N}(0, \sigma^2 I_d)$ to have that $\sigma^2/\|X\|^2$ have a law inverse-χ^2 of parameter d. Hence,

$$|u^T M u| \leq 3\frac{\|\psi\|_\infty}{\sigma\sqrt{d-2}}.$$

Inject this in equation (31) to conclude that $\|\mathrm{IF}(x)\| \geq \sqrt{d-2}/6$. Now, use that $\varepsilon \mapsto T(P_\varepsilon)$ is continuous (it is in fact Lipshitz continuous because the Influence function is bounded) to conclude.

A.4. Proof of Theorem 4

This proof mimic the proof of convergence for EM algorithm or for the algorithm used to compute the geometric median.

Let

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho_1\left(\frac{\|X_i - \theta\|}{\beta}\right),$$

we are searching for the argmin of J_n.

First, we show that the initialization is not too far away from the optimum.

Lemma 10. *If Assumption 1 is verified, Σ is the covariance matrix of P and $|\mathcal{O}| \leq n/8$. Then, with probability larger than $1 - de^{-n/8} - \delta$,*

$$\|\theta_0 - T(X_1^n)\| \leq 2\sqrt{2Tr(\Sigma)} + r_n$$

The proof is provided in Section C.10. Then, we note that J_n is strongly convex with high probability, using Lemma 1, for all $u \in S$ and all $\theta \in \mathbb{R}^d$,

$$u^T Hess(J_n)(\theta) u \geq \frac{1}{\beta^2 n}\sum_{i=1}^n \psi_1'\left(\frac{\|X_i - \theta\|}{\beta}\right). \tag{32}$$

This allows us to show that the sequence of iterates is equivalent to minimizing a convex majorant of J.

Lemma 11. *Assume that ρ is convex, that $x \mapsto \psi(x)/x$ is bounded and non-increasing, define*

$$\boldsymbol{U}_\kappa(\theta) = \sum_{i=1}^n \left(\frac{w_i(\kappa)}{2}\left(\frac{\|X_i - \theta\|}{\beta}\right)^2 + \rho_1(r_i(\kappa)) - \frac{1}{2}r_i(\kappa)\psi(r_i(\kappa))\right).$$

where $r_i(\kappa) = \|X_i - \kappa\|/\beta$ and $w_i(\kappa) = \psi_1(r_i(\kappa))/r_i(\kappa)$. We have that

- *$\boldsymbol{U}_{\theta^{(m)}}$ is minimized at $\theta^{(m+1)}$*

- *for all* $\kappa \in \mathbb{R}^d$, $\boldsymbol{U}_\kappa \geq J_n$
- *The approximation error* $h = J_n - U_\kappa$ *is differentiable and* ∇h *is* L*-Lipshitz continuous with* $L = \frac{1}{n\beta^2}\sum_{i=1}^n \frac{\psi(r_i)}{r_i}$
- $h(\kappa) = 0$ *and* $\nabla h(\kappa) = 0$.

The proof is provided in Section C.11.

From Lemma 11, we get that the direction from $\theta^{(m)}$ to $\theta^{(m+1)}$ is a proper descent direction through the following lemma proved in Section C.12.

Lemma 12. *For all* $\theta \in \mathbb{R}^d$ *and for all* $m \in \mathbb{N}$, *we have*

$$J(\theta^{(m+1)}) \leq J(\theta) + \frac{1}{2\beta^2 n}\sum_{j=1}^n w_j(\theta^{(m)})\left(\|\theta^{(m)} - \theta\| - \|\theta^{(m+1)} - \theta\|\right)$$

Use Lemma 12 with $\theta = T(X_1^n)$.

$$J(\theta^{(m+1)}) \leq J(T(X_1^n)) + \frac{1}{2\beta^2 n}\sum_{j=1}^n w_j(\theta^{(m)})\left(\|\theta^{(m)} - T(X_1^n)\| - \|\theta^{(m+1)} - T(X_1^n)\|\right)$$

Hence, using the fact that $T(X_1^n)$ is a minimizer of J, we get the result that

$$\|\theta^{(m+1)} - T(X_1^n)\| \leq \|\theta^{(m)} - T(X_1^n)\| \tag{33}$$

This allows us to restrict ourselves to a bounded domain. From Lemma 10, with probability larger than $1 - de^{-n/8} - \delta$, we have that θ_0 is in

$$\Theta = \{\theta \in \mathbb{R}^d : \quad \|\theta - T(X_1^n)\| \leq 2\sqrt{2Tr(\Sigma)} + r_n\}.$$

and then, equation (33) assures us that we stay in Θ for the other iterations: $\forall m \in \mathbb{N}, \theta^{(m)} \in \Theta$. The following Lemma is proven in Section C.13.

Lemma 13. *Let* $\theta \in \Theta$, *if* $X_1, \ldots, X_n$ *are corrupted by an adversary with* $X_1', \ldots, X_n'$ *i.i.d with law* P *whose variance is finite and the covariance matrix is denoted* Σ. *Moreover, we suppose* $|\mathcal{O}| \leq n/8$, *then if* $\beta \geq 2\sqrt{2Tr(\Sigma)} + r_n + \psi_1^{-1}\left(\sqrt{2V_\psi}\right)$, *we have with probability greater than* $1 - e^{-n/32}$,

$$\frac{1}{\beta^2 n}\sum_{i=1}^n \psi_1'\left(\frac{\|X_i - \theta\|}{\beta}\right) \geq \frac{\gamma}{4\beta^2}$$

This allows us to quantify the speed at which the sequence $\|\theta^{(m)} - T(X_1^n)\|$ decreases. Indeed, from Lemma 12 for $\theta = T(X_1^n)$, we have

$$J(\theta^{(m+1)}) \leq J(T(X_1^n)) + \frac{1}{2\beta^2 n}\sum_{j=1}^n w_j(\theta^{(m)})\left(\|\theta^{(m)} - T(X_1^n)\| - \|\theta^{(m+1)} - T(X_1^n)\|\right)$$

and by Lemma 13 and Lemma 10 and the convexity equation (32), we have by Taylor's theorem that with probability larger than $1-(d+4)e^{-n/8}-\delta$ (because $e^{-n/32} \leq 4e^{-n/8}$),

$$J(\theta^{(m+1)}) - J(T(X_1^n)) \geq \frac{\gamma}{4\beta^2}\|\theta^{(m+1)} - T(X_1^n)\|.$$

Hence,

$$\frac{\gamma}{4\beta^2}\|\theta^{(m+1)} - T(X_1^n)\| \leq \frac{1}{2\beta^2 n}\sum_{j=1}^{n} w_j(\theta^{(m)}) \left(\|\theta^{(m)} - T(X_1^n)\| - \|\theta^{(m+1)} - T(X_1^n)\|\right).$$

Solve this equation for $\|\theta^{(m+1)} - T(X_1^n)\|$ to obtain,

$$\|\theta^{(m+1)} - T(X_1^n)\| \leq \frac{\frac{1}{2n\beta^2}\sum_{j=1}^{n} w_j(\theta^{(m)})}{\frac{1}{2n\beta^2}\sum_{j=1}^{n} w_j(\theta^{(m)}) + \frac{\gamma}{4\beta^2}}\|\theta^{(m)} - T(X_1^n)\| \tag{34}$$

then use that $\sum_{j=1}^{n} w_j(\theta^{(m)}) \leq n$ and the fact that the right-hand-side of Equation (34) is increasing in $\sum_{j=1}^{n} w_j(\theta^{(m)})$ to get for all $m \in \mathbb{N}$,

$$\|\theta^{(m+1)} - T(X_1^n)\| \leq \frac{1/2}{1/2 + \frac{\gamma}{4}}\|\theta^{(m)} - T(X_1^n)\| = \frac{1}{1+\gamma/2}\|\theta^{(m)} - T(X_1^n)\| \tag{35}$$

Then, this implies directly that with probability larger than $1-(d+4)e^{-n/8}-\delta$, we have

$$\|\theta^{(m)} - T(X_1^n)\| \leq \frac{1}{(1+\gamma/2)^m}\|\theta^{(0)} - T(X_1^n)\| \leq \frac{1}{(1+\gamma/2)^m}\left(2\sqrt{2Tr(\Sigma)} + r_n\right)$$

Appendix B: Proof of propositions

B.1. Proof of Proposition 1

Hypotheses in this proof. *(i)* $|\mathcal{O}| \leq n/32$
(ii) $\beta^2 \geq V_{\psi_H} \max(8, C_1/n)$

From Theorem 1 with $\gamma = 1$, we have that if $V_{\psi_H} = \mathbb{E}_P[\psi_H(\|X - T_H(P)\|)^2] \leq \psi(\beta/2)^2/2 = \beta^2/8$, then for all $\lambda \in (0, \beta/2)$ and for all $\theta \in H$,

$$t_T(\lambda; T_H(P), X_1^n) \leq t_{\mathrm{IF}}(\lambda/4; T_H(P), X_1^n) + e^{-n/32}. \tag{36}$$

let us find an upper bound of $t_{\mathrm{IF}}(\lambda/4; T_H(P), X_1^n)$.

Step 1. $t_{\mathrm{IF}}(\lambda_t/4; T_H(P), X_1^n) \leq 4e^{-t}$.

Proof. We have,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-T(P)}{\|X_i-T_H(P)\|}\psi_H\left(\|X_i-T_H(P)\|\right)\right\|\geq\lambda/4\right)$$
$$\leq\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i'-T(P)}{\|X_i'-T_H(P)\|}\psi_H\left(\|X_i'-T_H(P)\|\right)\right\|\geq\lambda/4-2\frac{|\mathcal{O}|}{n}\beta\right) \tag{37}$$

because ψ_H is bounded by β. Then, we use Lemma 8 to bound the sum of i.i.d. bounded random variables. Having ψ_H bounded, we see that by Hoeffding's lemma (see [4, Section 2.3]), $\max_{1\leq i\leq n}\|\psi_H(\|X_i-T_H(P)\|)\|_{\psi_1}\leq\beta$. Then, by Lemma 8 for all $t>0$, with probability larger than $1-4e^{-t}$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i'-T(P)}{\|X_i'-T_H(P)\|}\psi_H\left(\|X_i'-T_H(P)\|\right)\right\|\leq\frac{3}{2}\sqrt{\frac{V_{\psi_H}}{n}}+2\sqrt{\frac{tv_{\psi_H}}{n}}+\frac{Ct\beta}{n},$$

inject this in Equation (37) with $\lambda_t=4\left(\frac{3}{2}\sqrt{\frac{V_{\psi_H}}{n}}+2\sqrt{\frac{tv_{\psi_H}}{n}}+\frac{Ct\beta}{n}\right)+8\beta\frac{|\mathcal{O}|}{n}$ to get $t_{\mathrm{IF}}\left(\lambda_t/4;T_H(P),X_1^n\right)\leq 4e^{-t}$. □

Step 2. There exists a constant $C_2>0$ such that condition $\lambda_t\leq\beta/2$ is verified for any $t\leq C_2 n$.

Proof. Use Hypothesis (i) to get that $\lambda_t\leq\beta/2$ is implied by

$$\frac{\beta}{4}\geq 6\sqrt{\frac{V_{\psi_H}}{n}}+8\sqrt{\frac{tv_{\psi_H}}{n}}+4\frac{Ct\beta}{n}.$$

Then, this is implied by the following system of equation with C_1,C_2, $C_3>0$ three numerical constants,

$$\begin{cases}\beta\geq C_1\sqrt{\frac{V_{\psi_H}}{n}}\\ \beta\geq\sqrt{C_2\frac{tv_{\psi_H}}{n}}\\ \beta\geq\frac{t\beta}{C_3 n}\end{cases}\quad\Leftarrow\quad\begin{cases}\beta\geq C_1\sqrt{\frac{V_{\psi_H}}{n}}\\ t\leq C_2 n\beta^2/v_{\psi_H}\\ t\leq C_3 n\end{cases}\quad\Leftarrow\quad\begin{cases}\beta\geq C_1\sqrt{\frac{V_{\psi_H}}{n}}\\ t\leq C_2 n\\ t\leq C_3 n\end{cases}$$

The last implication is from Hypothesis (ii): $\beta^2\geq 8V_{\psi_H}\geq 8v_{\psi_H}$. Inject this in Equation (36) to get the desired result. □

B.2. Proof of Proposition 2

Hypotheses in this proof. *(i)* $\mathbb{E}_P[\|X-\mathbb{E}_P[X]\|^q]<\infty$ *for some* $q\geq 2$.
(ii) $t\leq C_1 n\min\left(\frac{\mathbb{E}_P[\|X-\mathbb{E}_P[X]\|^q]}{Tr(\Sigma)^{q/2}},1\right)$,

Step 1. $\beta = \left(\frac{2\mathbb{E}_P[\|X-\mathbb{E}_P[X]\|^q]}{8\varepsilon_n + C_3 t/n}\right)^{1/q}$ verifies the conditions of Proposition 1.

Proof. We will check the stronger assumption that $\beta^2 \geq Tr(\Sigma)\max(8, C_1/n)$.
Having $\varepsilon_n \geq 0$, we can bound the chosen β by $\beta^q \geq n\frac{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]}{C_3 t}$, then there exists $C > 0$ such that the hypothesis on β is verified if

$$t \leq C\frac{n\mathbb{E}[\|X-\mathbb{E}[X]\|^q]}{Tr(\Sigma)^{q/2}}.$$

This is condition on t is verified in Hypothesis (ii). □

Step 2. For any $t \leq C_2 n$, we have

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{V_{\psi_H}}{n}} + 8\sqrt{\frac{tv_{\psi_H}}{n}} + 4\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q} \mathbb{E}[\|X-\mathbb{E}_P[X]\|^q]^{1/q}.$$

Proof. From the previous step, we can apply Proposition 1, and by Lemma 4 and Hypothesis (i), if $\beta^2 \geq V_{\psi_H}\max(8, C_1/n)$ then with probability larger than $1 - 4e^{-t} - e^{-n/32}$,

$$\begin{aligned}
&\|T_H(X_1^n) - \mathbb{E}_P[X]\| \\
&\leq 6\sqrt{2\frac{V_{\psi_H}}{n}} + 8\sqrt{2\frac{tv_{\psi_H}}{n}} + \frac{C_3 t\beta}{n} + 8\beta\varepsilon_n + \|\mathbb{E}_P[X] - T_H(P)\| \\
&\leq 6\sqrt{2\frac{V_{\psi_H}}{n}} + 8\sqrt{2\frac{tv_{\psi_H}}{n}} + \frac{C_3 t\beta}{n} + 8\beta\varepsilon_n + \frac{2\mathbb{E}[\|X-\mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}. \quad (38)
\end{aligned}$$

To simplify, we don't take into account the effect of β on V_{ψ_H} and v_{ψ_H} when choosing β then, we choose β such that

$$\beta^q = \frac{2\mathbb{E}_P[\|X-\mathbb{E}_P[X]\|^q]}{8\varepsilon_n + C_3 t/n}$$

inject this in Equation (38) to get

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{V_{\psi_H}}{n}} + 8\sqrt{\frac{tv_{\psi_H}}{n}} + 2^{1/q}\left(1 + \frac{1}{q-1}\right)\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q} \mathbb{E}_P[\|X-\mathbb{E}_P[X]\|^q]^{1/q}.$$

The result follow because $2^{1/q}(1 + \frac{1}{q-1}) \leq 4$. □

Step 3. For any $t \leq C_2 n$,

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{2\frac{Tr(\Sigma)}{n}} + 8\sqrt{2\frac{t\|\Sigma\|_{op}}{n}} + C_4\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q} \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}.$$

Proof. From Lemma 6, $V_{\psi_H} \leq Tr(\Sigma)$ and

$$\begin{aligned} v_{\psi_H} \leq & \|\Sigma\|_{op} + \|\mathbb{E}[X] - T_H(P)\|^2 \\ \leq & \|\Sigma\|_{op} + \frac{4\mathbb{E}[\|X - \mathbb{E}[X]\|^{2q}]}{(q-1)^2\beta^{2q-2}} \leq \|\Sigma\|_{op} \\ & + \frac{2^{2/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{2/q}}{(q-1)^2}\left(8\varepsilon_n + C_3\frac{t}{n}\right)^{2-2/q} \end{aligned}$$

by sub-linearity of the square root we obtain

$$\sqrt{v_{\psi_H}} \leq \sqrt{\|\Sigma\|_{op}} + \frac{2^{1/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}}{(q-1)}\left(4\varepsilon_n + C_3\frac{t}{n}\right)^{1-1/q}$$

inject this in Step 2, to get

$$\begin{aligned} & \|T_H(X_1^n) - \mathbb{E}_P[X]\| \\ \leq & 6\sqrt{2\frac{Tr(\Sigma)}{n}} + 8\sqrt{2\frac{t\|\Sigma\|_{op}}{n}} \\ & + \left(4 + 2\sqrt{t/n}\right)\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q} \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \\ \leq & 6\sqrt{2\frac{Tr(\Sigma)}{n}} + 8\sqrt{2\frac{t\|\Sigma\|_{op}}{n}} + C_4\left(\frac{C_3 t}{n} + 8\varepsilon_n\right)^{1-1/q} \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}. \end{aligned}$$

□

B.3. Proof of Proposition 3

From Proposition 1, if $\beta^2 \geq V_{\psi_H}\max(8, C_1/n)$ and having $\mathbb{E}_P[X] = T_H(P)$, we get with probability larger than $1 - 4e^- t - e^{-n/32}$,

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{2\frac{V_{\psi_H}}{n}} + 8\sqrt{2\frac{t v_{\psi_H}}{n}} + \frac{C_3 t\beta}{n} + 8\beta\varepsilon_n \tag{39}$$

Take $\beta = C_2\mathbb{E}[\|X - \mathbb{E}[X]\|]$ with $C_2^2 \geq \max(8, C_1)$, this value of β verifies that $\beta^2 \geq V_{\psi_H}\max(8, C_1/n)$ hence it verifies the condition of Proposition 1. Inject the value of β in Equation (39) to get,

$$\|T_H(X_1^n) - \mathbb{E}_P[X]\| \leq 6\sqrt{\frac{V_{\psi_H}}{n}} + 8\sqrt{\frac{t v_{\psi_H}}{n}} + C_2\mathbb{E}[\|X - T(P)\|]\left(\frac{C_3 t}{n} + 8\varepsilon_n\right). \tag{40}$$

B.4. Proof of Proposition 4

Hypotheses in this proof. *(i) With probability larger than* $1-\delta_O$*, we have* $\frac{1}{|\mathcal{O}|}\sum_{i\in\mathcal{O}}\psi_C(\|X_i-T_C(P)\|)+\psi_C(\|X_i'-T_C(P)\|)\leq\beta C_O$.
(ii) $\beta^3=\frac{5\mathbb{E}[\|X-\mathbb{E}[X]\|^3]}{120C_O\varepsilon_n+24Ct\log(n)/n}$,
(iii) $|\mathcal{O}|\leq n/(20C_O)$

Step 1. β verify that $\beta^2\geq 10Tr(\Sigma)$.

Proof. Having $\varepsilon_n\geq 0$ and $s\geq 1$, we can bound the chosen β by $\beta^3\geq\frac{5\mathbb{E}[\|X-\mathbb{E}[X]\|^3]}{48Ct\log(n)/n}$. Then, because there exists $C_3>0$ such that $t\leq C_3 n/\log(n)$, we have $\beta^2\geq 10\mathbb{E}[\|X-\mathbb{E}[X]\|^3]^{2/3}$ (if needed, C_3 can be decreased to obtain the right constant), hence $\beta^2\geq 10Tr(\Sigma)$. □

Step 2.

$$\begin{aligned}&t_{\mathrm{IF}}(\lambda;T_C(P),X_1^n)\\&\leq\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\frac{X_i'-T_C(P)}{\|X_i'-T_C(P)\|}\psi_C\left(\|X_i'-T_C(P)\|\right)\right\|\geq\lambda/5-\varepsilon_nC_O\beta\right)+\delta_O\end{aligned}$$

Proof. From Theorem 1 with $\gamma=4/5$, we have that if $V_{\psi_C}=\mathbb{E}_P[\psi_C(\|X-T_C(P)\|)^2]\leq\psi_C(\beta/2)^2/2=\beta^2\log(13/8)^2/2$, then for all $\lambda\in(0,\beta/2)$,

$$t_T(\lambda;T_C(P),X_1^n)\leq t_{\mathrm{IF}}\left(\lambda/5;T_C(P),X_1^n\right)+e^{-n/50}. \tag{41}$$

The condition $V_{\psi_C}\leq\beta^2\log(13/8)^2/2$ is verified because $\beta^2\geq 10Tr(\Sigma)$ and $V_{\psi_C}\leq Tr(\Sigma)$ by Lemma 6.
Now, we take care of the outliers in the right-hand side of Equation (41),

$$\begin{aligned}&t_{\mathrm{IF}}\left(\lambda/5;T_C(P)\right)\\&\leq\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\frac{X_i'-T_C(P)}{\|X_i'-T_C(P)\|}\psi_C\left(\|X_i'-T_C(P)\|\right)\right\|\geq\lambda/5\right.\\&\quad\left.-\frac{1}{n}\sum_{i\in\mathcal{O}}\psi_C(\|X_i-T_C(P)\|)+\psi_C(\|X_i'-T_C(P)\|)\right)\\&\leq\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\frac{X_i'-T_C(P)}{\|X_i'-T_C(P)\|}\psi_C\left(\|X_i'-T_C(P)\|\right)\right\|\geq\lambda/5-\varepsilon_nC_O\beta\right)+\delta_O\end{aligned} \tag{42}$$

□

Step 3. $\mathbb{E}[\|X-T_C(P)\|^2]\leq 2Tr(\Sigma)$.

Proof. We apply Lemma 5 with $\gamma = 4/5$, $\|\psi'\|_\infty \leq 1$ and two finite moments to get, $\|T_C(P) - \mathbb{E}_P[X]\| \leq \frac{5Tr(\Sigma)}{4\beta}$. We get $\|T_C(P) - \mathbb{E}_P[X]\| \leq 5Tr(\Sigma)/(4\sqrt{10}) \leq Tr(\Sigma)$. Hence, by bias-variance decomposition, $\mathbb{E}[\|X - T_C(P)\|^2] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + \|T_C(P) - \mathbb{E}_P[X]\|^2 \leq 2Tr(\Sigma)$. □

Step 4. We have

$$t_{\mathrm{IF}}\left(\lambda_t/5; T_C(P)\right) \leq 4e^{-t} + \delta_0$$

where

$$\lambda_t = 5\left(\frac{3}{2}\sqrt{\frac{V_{\psi_C}}{n}} + 2\sqrt{\frac{tv_{\psi_C}}{n}} + \frac{Ct\beta}{n}\log(n)\right) + 5\varepsilon_n C_O \beta.$$

Proof. In order to use Lemma 8, we use the following Lemma.

Lemma 14. *If X satisfies $\mathbb{E}_P[\|X\|^2] < \infty$, then, for all $q \in \mathbb{N}^*$,*

$$\mathbb{E}_P[\psi_C(\|X - T_C(P)\|)^q] \leq q!(s\beta)^q,$$

where

$$s = \max\left(e, \log\left(1 + \frac{\mathbb{E}_P[\|X - T_C(P)\|])}{\beta} + \frac{\mathbb{E}[\|X - T_C(P)\|^2]}{2\beta^2}\right)\right).$$

The proof of Lemma 14 is postponed to Section C.14. By Step 3 and the fact that $\beta^2 \geq Tr(\Sigma)$, we have $s \leq \max(e, \log(1 + 1 + 1/2)) \leq e$. Then, by Lemma 14, for any $q \in \mathbb{N}^*$,

$$\mathbb{E}_P[\psi_C(\|X - T_C(P)\|)^q] \leq q!(e\beta)^q,$$

Then, using the power series expansion of the exponential function, we get that, for all $t > \beta e$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\psi_C(\|X - T_C(P)\|)}{t}\right)\right] &= \sum_{q=0}^{\infty} \frac{\mathbb{E}[\psi_C(\|X - T(P)\|^q]}{t^q q!} \\ &\leq \sum_{q=0}^{\infty} \frac{\beta^q e^q}{t^q} = \frac{1}{1 - \beta e/t}.\end{aligned}$$

Choosing $t = 2\beta e$ shows that $\|\psi_C(\|X - T_C(P)\|)\|_{\psi_1} \leq 2e\beta$. Then, using Lemma 8, we get for all $t > 0$, with probability larger than $1 - 4e^{-t}$,

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n} \frac{X_i' - T_C(P)}{\|X_i' - T_C(P)\|}\psi_H\left(\|X_i' - T_C(P)\|\right)\right\| \\ &\quad\leq \frac{3}{2}\sqrt{\frac{V_{\psi_C}}{n}} + 2\sqrt{\frac{tv_{\psi_C}}{n}} + \frac{Ct\beta}{n}\log(n),\end{aligned}$$

Hence, from Step 2, we have

$$t_{\mathrm{IF}}\left(\lambda_t/5; T_C(P)\right) \leq 4e^{-t} + \delta_0$$

where

$$\lambda_t = 5\left(\frac{3}{2}\sqrt{\frac{V_{\psi_C}}{n}} + 2\sqrt{\frac{tv_{\psi_C}}{n}} + \frac{Ct\beta}{n}\log(n)\right) + 5\varepsilon_n C_O \beta.$$

□

Step 5. The condition $\lambda_t \leq \beta/2$ is implied by $t \leq C_3 n/\log(n)$ for some absolute constant C_3.

Proof. We use that $\varepsilon_n C_O \leq 1/20$ to get that $\lambda_t \leq \beta/2$ is implied by

$$\frac{\beta}{4} \geq \frac{15}{2}\sqrt{\frac{V_{\psi_C}}{n}} + 10\sqrt{\frac{tv_{\psi_C}}{n}} + 5\frac{Cts\beta}{n}\log(n).$$

Then, this is implied by the following system of equation with C_1, C_2, $C_3 > 0$ three numerical constants,

$$\begin{cases} \beta \geq C_1\sqrt{\frac{V_{\psi_C}}{n}} \\ \beta \geq \sqrt{C_2\frac{tv_{\psi_C}}{n}} \\ \beta \geq \frac{t\beta}{C_3 n}\log(n) \end{cases} \Leftarrow \begin{cases} \beta \geq C_1\sqrt{\frac{V_{\psi_C}}{n}} \\ t \leq C_2 n\beta^2/v_{\psi_C} \\ t \leq C_3\frac{n}{s\log(n)} \end{cases} \Leftarrow \begin{cases} \beta \geq C_1\sqrt{\frac{V_{\psi_C}}{n}} \\ t \leq C_2 n \\ t \leq C_3\frac{n}{\log(n)} \end{cases}$$

The last implication is because $\beta^2 \geq 8V_{\psi_C} \geq 8v_{\psi_C}$. The first inequality is necessarily verified because $\beta^2 \geq 10Tr(\Sigma)$, hence the only remaining condition is $t \leq C_3 n/\log(n)$ for some absolute constant C_3. □

Step 6. There exists a numerical constant $C_2 > 0$ such that with probability larger than $1 - 4e^{-t} - \delta_0$,

$$\begin{aligned} \|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq{}& \frac{15}{2}\sqrt{\frac{Tr(\Sigma)}{n}} + 10\sqrt{\frac{t\|\Sigma\|_{op}}{n}} \\ &+ C_2\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + Ct\frac{\log(n)}{n}\right)^{2/3}. \end{aligned} \tag{43}$$

Proof. From Lemma 5, with 3 finite moments, we have

$$\|T_C(P) - \mathbb{E}_P[X]\| \leq \frac{5\mathbb{E}[\|X - \mathbb{E}[X]\|^3]}{24\beta^2}. \tag{44}$$

Then, from Step 4 with probability larger than $1 - 4e^{-t} - e^{-n/50}$,

$$\|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq \frac{15}{2}\sqrt{\frac{V_{\psi_C}}{n}}$$

$$+ 10\sqrt{\frac{tv_{\psi_C}}{n}} + \frac{Ct\beta}{n}\log(n) + 5\beta C_O\varepsilon_n + \frac{5\mathbb{E}[\|X - \mathbb{E}[X]\|^3]}{24\beta^2}.$$

Then, there exists a constant $C_4 > 0$ such that

$$\|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq \frac{15}{2}\sqrt{\frac{V_{\psi_C}}{n}} + 10\sqrt{\frac{tv_{\psi_C}}{n}} + C_4\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + t\frac{\log(n)}{n}\right)^{2/3}, \quad (45)$$

To simplify, we also use the bounds on the variance from 6, $V_{\psi_C} \leq Tr(\Sigma)$ and also

$$\begin{aligned} v_{\psi_C} &\leq \|\Sigma\|_{op} + \|\mathbb{E}[X] - T_C(P)\|^2 \\ &\leq \|\Sigma\|_{op} + \frac{25\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^2}{576\beta^2}. \end{aligned}$$

By sub-linearity of the square root we obtain

$$\begin{aligned} \sqrt{v_{\psi_C}} &\leq \sqrt{\|\Sigma\|_{op}} + \frac{5\mathbb{E}[\|X - \mathbb{E}[X]\|^3]}{24\beta^2} \\ &= \sqrt{\|\Sigma\|_{op}} + C_4\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + ts\frac{\log(n)}{n}\right)^{2/3}, \end{aligned}$$

inject this in Equation (45), to get

$$\|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq \frac{15}{2}\sqrt{\frac{Tr(\Sigma)}{n}} + 10\sqrt{\frac{t\|\Sigma\|_{op}}{n}} + C_4\left(1 + \sqrt{\frac{2t}{n}}\right)\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + Ct\frac{\log(n)}{n}\right)^{2/3}.$$

Having $t \lesssim n$, we get that there exists a numerical constant $C_2 > 0$ such that

$$\|T_C(X_1^n) - \mathbb{E}_P[X]\| \leq \frac{15}{2}\sqrt{\frac{Tr(\Sigma)}{n}} + 10\sqrt{\frac{t\|\Sigma\|_{op}}{n}} + C_2\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + Ct\frac{\log(n)}{n}\right)^{2/3}. \quad (46)$$

□

Step 7. Using the sub-linearity of $x \mapsto x^{2/3}$ and using that

$$t \leq C_2\frac{n}{\log(n)^4}\frac{\sqrt{\|\Sigma\|_{op}}}{\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/3}}$$

for C_2 small enough in equation (46), we have

$$C_2\mathbb{E}[\|X-\mathbb{E}[X]\|^3]^{1/3}\left(C_O\varepsilon_n + Ct\frac{\log(n)}{n}\right)^{2/3} \leq C_3\mathbb{E}[\|X-\mathbb{E}[X]\|^3]^{1/3}\varepsilon_n^{2/3} + \sqrt{\frac{t\|\Sigma\|_{op}}{n}}$$

which finishes the proof.

B.5. Proof of Proposition 5

Hypotheses in this proof. *(i) With probability larger than* $1-\delta_O$*, we have* $\frac{1}{|\mathcal{O}|}\sum_{i\in\mathcal{O}}\psi_P(\|X_i - T_P(P)\|) + \psi_P(\|X_i' - T_P(P)\|) \leq \beta C_O$.
(ii) $\beta^2 = Tr(\Sigma)$

Step 1. For $t \leq C_1 n$, with probability larger than $1 - e^{-C_1 t} - e^{-n/512} - \delta_O$,

$$\|T_P(X_1^n) - T_P(P)\| \leq 16\sqrt{\frac{\mathbb{E}_P[\|X-T_P(P)\|^2]t}{n}} + 16\varepsilon_n C_O\beta.$$

Proof. From Theorem 1 with $\gamma = 1/4$, we have that if $V_{\psi_P} = \mathbb{E}_P[\psi_P(\|X - T_P(P)\|)^2] \leq \psi_P(\beta/2)^2/2 = \beta^2/8$, then for all $\lambda \in (0, \beta/2)$,

$$t_T(\lambda; T_P(P), X_1^n) \leq t_{\mathrm{IF}}(\lambda/16; T_P(P), X_1^n) + e^{-n/512}. \tag{47}$$

The following lemma applies.

Lemma 15. *Let* $n \in \mathbb{N}^*$*, suppose* $X_1, \ldots, X_n$ *are i.i.d. Let* $q \in \mathbb{N}^*$ *and suppose* $\mathbb{E}_P[\|X\|^q] < \infty$*. There exists an absolute constant* $K > 0$ *such that*

$$t_{\mathrm{IF}}(\lambda) \leq \frac{\mathbb{E}_P[\|X - T_P(P)\|^q]}{\beta^q}\left(\frac{Kpq\beta}{\sqrt{n}\lambda}\right)^{qp}.$$

The proof is postponed to Section C.15.
Take $q = 2$ and $\lambda = \sqrt{\mathbb{E}_P[\|X - T_P(P)\|^2]t/n}$, we get for all $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \frac{X_i' - T_P(P)}{\|X_i' - T_P(P)\|}\psi_P(\|X_i' - T_P(P)\|)\right\| \geq \sqrt{\frac{\mathbb{E}_P[\|X-T_P(P)\|^2]t}{n}}\right) \leq \left(\frac{\beta^2}{\mathbb{E}_P[\|X-T_P(P)\|^2]}\right)^{p-1}\left(\frac{4K^2p}{t}\right)^{2p}.$$

Take having $Tr(\Sigma) \leq \mathbb{E}_P[\|X - T(P)\|^2]$, we have $\beta^2 \leq \mathbb{E}_P[\|X - T(P)\|^2]$ hence

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \frac{X_i' - T_P(P)}{\|X_i' - T_P(P)\|}\psi_P(\|X_i' - T_P(P)\|)\right\| \geq \sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}}\right)$$

$$\leq \left(\frac{4K^2 p}{t}\right)^{2p}.$$

Take $p = \frac{t}{4K^2}e^{-1}$, and there is a constant $C_1 > 0$ such that with probability larger than $1 - e^{-C_1 t}$, we have

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{X_i' - T_P(P)}{\|X_i' - T_P(P)\|}\psi_P(\|X_i' - T_P(P)\|) \right\| \leq \sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}}. \tag{48}$$

Now, we take care of the outliers in a similar way as in Proposition 1 and Proposition 4. We have from Hypothesis (i),

$$t_{\mathrm{IF}}\left(\lambda/16; T_P(P), X_1^n\right) \leq t_{\mathrm{IF}}\left(\frac{\lambda}{16} - \varepsilon_n C_O \beta; T_P(P), (X')_1^n\right) + \delta_O \tag{49}$$

Hence, from Equation (48) and Equation (49) with

$$\lambda = \lambda_t = 16\sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}} + 16\varepsilon_n C_O \beta,$$

we have with probability larger than $1 - e^{-C_1 t} - e^{-n/512} - \delta_O$,

$$\begin{aligned}\|T_P(X_1^n) - T_P(P)\| &\leq 16\sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}} + 16\varepsilon_n C_O \beta \\ &\leq 16\sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}} + 16\varepsilon_n C_O \beta.\end{aligned}$$

□

Step 2. There exist a constant $C_1' > 0$ such that for any $t \leq C_1' n$, we have $\lambda_t \leq \beta/2$.

Proof. The condition $\lambda_t \leq \beta/2$ is implied by

$$16\sqrt{\frac{\mathbb{E}_P[\|X - T_P(P)\|^2]t}{n}} \leq \beta/4$$

using the fact that $\varepsilon_n \leq 1/(64 C_O)$. This simplifies with $t \lesssim n\beta^2/\mathbb{E}_P[\|X - T_P(P)\|^2]$ and then, by Hypothesis (ii) $t \lesssim n$. □

Step 3. With probability larger than $1 - e^{-C_2 t} - e^{-n/512} - \delta_O$, we have

$$\|T_P(X_1^n) - \mathbb{E}[X]\| \leq 16\sqrt{65\frac{Tr(\Sigma)t}{n}} + 16\varepsilon_n C_O \sqrt{Tr(\Sigma)}.$$

Proof. For any $t \leq C_1 n\beta^2/\mathbb{E}[\|X - T_P(P)\|^2]$, with probability larger than $1 - e^{-C_2 t} - e^{-n/512} - \delta_O$,

$$\|T_P(X_1^n) - T_P(P)\| \leq 16\sqrt{\frac{\mathbb{E}[\|X - T_P(P)\|^2]t}{n}} + 16\varepsilon_n C_O \beta. \tag{50}$$

From Lemma 5, we also have that $\|T_P(P)-\mathbb{E}[X]\| \leq 8Tr(\Sigma)/\beta$. Hence,

$$\begin{aligned}\mathbb{E}[\|X - T_P(P)\|^2] &\leq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + \|\mathbb{E}[X] - T_P(P)\|^2 \\ &\leq Tr(\Sigma)\left(1 + \frac{64Tr(\Sigma)}{\beta^2}\right)\end{aligned}$$

from Hypothesis (ii) we have $\beta^2 = Tr(\Sigma)$, hence

$$\mathbb{E}[\|X - T_P(P)\|^2] \leq Tr(\Sigma)(1+64) = 65Tr(\Sigma),$$

and $\|T_P(P) - \mathbb{E}[X]\| \leq 8\sqrt{Tr(\Sigma)}$. Then, inject this in Equation (50),

$$\|T_P(X_1^n) - \mathbb{E}[X]\| \leq 16\sqrt{65\frac{Tr(\Sigma)t}{n}} + 16\varepsilon_n C_O\sqrt{Tr(\Sigma)}.$$

□

Appendix C: Proof of lemmas

C.1. Proof of Lemma 1

By direct differentiation of $Z_{P,\psi}$,

$$\begin{aligned}Jac(Z_{P,\psi})(\theta) =& \mathbb{E}\left[\frac{(X-\theta)\otimes(X-\theta)}{\|X-\theta\|^3}\psi\left(\|X-\theta\|\right)\right] - \mathbb{E}\left[\frac{I_{\mathcal{H}}}{\|X-\theta\|}\psi\left(\|X-\theta\|\right)\right] \\ &- \mathbb{E}\left[\frac{(X-\theta)\otimes(X-\theta)}{\|X-\theta\|^2}\psi'\left(\|X-\theta\|\right)\right]\end{aligned}$$

Hence, for any $u \in S$,

$$\begin{aligned}u^T Jac(Z_{P,\psi})(\theta)u = &-\mathbb{E}\left[\frac{1}{\|X-\theta\|}\psi\left(\|X-\theta\|\right)\right] + \mathbb{E}\left[\frac{\langle X-\theta, u\rangle^2}{\|X-\theta\|^3}\psi\left(\|X-\theta\|\right)\right] \\ &- \mathbb{E}\left[\frac{\langle X-\theta, u\rangle^2}{\|X-\theta\|^2}\psi'\left(\|X-\theta\|\right)\right] \quad (51)\end{aligned}$$

Then, remark that because ψ is concave, by Taylor's theorem we have $\forall y \geq 0$, $\psi(y) \geq \psi(0) + y\psi'(y)$. And then, because and $\psi(0) = 0$ we have for any $y \geq 0$, $\psi(y) \geq y\psi'(y)$.

Moreover, by Cauchy-Schwarz inequality we also have $1 - \left\langle \frac{X-\theta}{\|X-\theta\|}, u\right\rangle^2 \geq 0$. Hence, with these two inequalities, we get

$$\mathbb{E}\left[\frac{1 - \langle\frac{X-\theta}{\|X-\theta\|}, u\rangle^2}{\|X-\theta\|}\psi\left(\|X-\theta\|\right)\right] \geq \mathbb{E}\left[\left(1 - \langle\frac{X-\theta}{\|X-\theta\|}, u\rangle^2\right)\psi'\left(\|X-\theta\|\right)\right].$$

Inject this in (51) to get the result: for any $u \in S$,

$$u^T Jac(Z_{P,\psi})(\theta)u \leq -\mathbb{E}\left[\psi'\left(\|X-\theta\|\right)\right].$$

C.2. Proof of Lemma 2

Hypotheses in this proof. *(i)* $\mathbb{E}[\rho(\|X - \mathbb{E}[X]\|)] < \rho(\beta)$

Step 1. $T(P)$ is a critical point of $J(\theta) = \mathbb{E}[\rho(\|X - \theta\|)]$.

Proof. First, notice that from Assumptions 2-(i) and 2-(iv) ρ is two times derivable and increasing on $\mathbb{R}_+$. Hence, J is differentiable and its gradient is

$$\nabla J(\theta) = -\mathbb{E}\left[\frac{X-\theta}{\|X-\theta\|}\psi(\|X-\theta\|)\right].$$

Then, by definition of $T(P)$ (Equation (7)), $T(P)$ verifies $\nabla J(T(P)) = 0$, i.e. $T(P)$ critical point of J. □

Step 2. J is convex

Proof. Let $\mathrm{Hess}(J)$ denote the Hessian of J. From Lemma 1, for any $\theta \in \mathcal{H}$ and $u \in S$,

$$u^T \mathrm{Hess}(J)(\theta)u = -u^T Jac(Z_\psi)(\theta)u \geq \mathbb{E}\left[\psi'(\|X-\theta\|)\right] \geq 0$$

Hence, J is convex because its Hessian is positive. □

Step 3. $T(P)$ exists

Proof. Because ρ is increasing, we have $\rho(x) \xrightarrow[x\to\infty]{} \infty$ and hence $J(\theta) \xrightarrow[\|\theta\|\to\infty]{} \infty$. Hence, J is coercive and as J is also convex (Step 2), its minimum $T(P)$ exists. □

Step 4. J is strictly convex at $T(P)$.

Proof. From Assumption 2-(iv), and because ρ is increasing,

$$\mathbb{E}[\psi'(\|X-T(P)\|)] \geq \gamma \mathbb{P}(\|X-T(P)\| \leq \beta) = \gamma \mathbb{P}(\rho(\|X-T(P)\|) \leq \rho(\beta)).$$

Hence, by Markov's inequality,

$$\mathbb{E}[\psi'(\|X-T(P)\|)] \geq \gamma\left(1 - \frac{\mathbb{E}[\rho(\|X-T(P)\|)]}{\rho(\beta)}\right). \tag{52}$$

By Step 3, we have that $T(P)$ minimizer of J, hence

$$\frac{\mathbb{E}[\rho(\|X-T(P)\|)]}{\rho(\beta)} = \frac{J(T(P))}{\rho(\beta)} \leq \frac{J(\mathbb{E}[X])}{\rho(\beta)}.$$

Inject this in Equation (52) to get

$$\mathbb{E}[\psi'(\|X-T(P)\|)] \geq \gamma\left(1 - \frac{\mathbb{E}[\rho(\|X-\mathbb{E}[X])\|)]}{\rho(\beta)}\right).$$

Then, using Hypothesis (i), we get $\mathbb{E}[\psi'(\|X-T(P)\|)] > 0$, which implies from Lemma 1 that for all $u \in \mathcal{H}$, $u \neq 0$, $u^T \mathrm{Hess}(J)(T(P))u > 0$. Hence J is strictly convex at $T(P)$, the minimizer of J. □

Step 5. $T(P)$ as defined by Equation (7) exists and is unique.

Proof. As $T(P)$ is minimizer of J, it is also root of Equation (7) and the existence and unicity we proven in Step 2 and Step 4. □

C.3. Proof of Lemma 3

Huber's score function: The equality for the Huber's score function is immediate by derivation of ψ_H.
Catoni's score function: ψ_C is differentiable, and we have for all $x \geq 0$,

$$\psi_C'(x) = \frac{1 + \frac{x}{\beta}}{1 + \frac{x}{\beta} + \frac{x^2}{2\beta^2}}.$$

This function is decreasing on $\mathbb{R}_+$, positive and even, hence

$$\psi_C'(x) \geq \psi_C'(\beta)\mathbb{1}\{x \leq \beta\} = \frac{4}{5}\mathbb{1}\{x \leq \beta\}.$$

Polynomial score function: ψ_P is differentiable, and we have for all $x \geq 0$

$$\psi_P(x)' = \frac{1 + \frac{1}{p}\left(\frac{x}{\beta}\right)^{1-1/p}}{\left(1 + \left(\frac{x}{\beta}\right)^{1-1/p}\right)^2}.$$

As in the case of Catoni's score function, this function is decreasing over $\mathbb{R}_+$, positive and even. Then, we get

$$\psi_P'(x) \geq \psi_P'(\beta)\mathbb{1}\{x \leq \beta\} = \frac{1}{4}\left(1 + \frac{1}{p}\right)\mathbb{1}\{x \leq \beta\} \geq \frac{1}{4}\mathbb{1}\{x \leq \beta\}.$$

C.4. Proof of Lemma 4

From Theorem 2 we only need to control $Z_\beta(\mathbb{E}[X])$. We have,

$$\begin{aligned}
&Z_\beta(\mathbb{E}[X]) \\
&\quad = \mathbb{E}\left[\beta\frac{(X - \mathbb{E}[X])}{\|X - \mathbb{E}[X]\|}\psi_1\left(\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right)\right] \\
&\quad = \mathbb{E}\left[\beta\frac{(X - \mathbb{E}[X])}{\|X - \mathbb{E}[X]\|}\psi_1\left(\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right)\right] - \mathbb{E}\left[\beta\frac{(X - \mathbb{E}[X])}{\|X - \mathbb{E}[X]\|}\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right]
\end{aligned}$$

Hence, by triangular inequality,

$$\|Z_\beta(\mathbb{E}[X])\| \leq \beta\mathbb{E}\left[\left|\psi_1\left(\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right) - \left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right|\right]$$

We denote $Y = \|X - \mathbb{E}[X]\|/\beta$, we have

$$\begin{aligned}\|Z_\beta(\mathbb{E}[X])\| \leq \beta\mathbb{E}\left[|\psi_1(Y) - Y|\right] &= \beta\int |\psi_1(y) - y| dF_Y(y) \\ &= \beta\int_0^\infty (y - \psi_1(y)) dF_Y(y)\end{aligned}$$

because ψ_1 is 1-Lipshitz and $\psi_1(0) = 0$. Then, by integration by part,

$$\|Z_\beta(\mathbb{E}[X])\| \leq \beta\int_0^\infty (1 - \psi_1'(y))(1 - F_Y(y)) dy$$

Until now, the proof was valid for any ψ_1, for the specific case of Huber score function and using Theorem 2, we get that

$$\|\mathbb{E}[X] - T_H(P)\| \leq 2\beta\int_1^\infty \mathbb{P}\left(\|X - \mathbb{E}[X]\| \geq \beta y\right) dy$$

Then, use Markov's inequality,

$$\|\mathbb{E}[X] - T_H(P)\| \leq 2\beta\int_1^\infty \frac{\|X - \mathbb{E}[X]\|^q}{\beta^q y^q} dy = \frac{2\|X - \mathbb{E}[X]\|^q}{(q-1)\beta^{q-1}}.$$

C.5. Proof of Lemma 5

We have,

$$Z_\beta(\mathbb{E}[X]) = \mathbb{E}\left[\frac{X - \mathbb{E}[X]}{\|X - \mathbb{E}[X]\|}\beta\psi_1\left(\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right)\right]$$

Then, by Taylor expansion

$$\begin{aligned}\|Z_\beta(\mathbb{E}[X])\| \leq &\left\|\mathbb{E}\left[\frac{X - \mathbb{E}[X]}{\|X - \mathbb{E}[X]\|}\beta\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|\right]\right\| + \beta\mathbb{E}\left[\frac{\|\psi_1^{(k)}\|_\infty}{k!}\left\|\frac{X - \mathbb{E}[X]}{\beta}\right\|^k\right] \\ = &\mathbb{E}\left[\frac{\|\psi_1^{(k)}\|_\infty \mathbb{E}[\|X - \mathbb{E}[X]\|^k]}{k!\beta^{k-1}}\right]\end{aligned}$$

which proves the first part of the lemma using Theorem 2. In the case of Bernoulli distribution, the result follows from a Taylor expansion:

$$\begin{aligned}Z_\beta(\mathbb{E}[X]) = &\mathbb{E}\left[\text{sign}(X - \mathbb{E}[X])\beta\psi_1\left(\left|\frac{X - \mathbb{E}[X]}{\beta}\right|\right)\right] \\ = &p\left(\beta\psi_1\left(\frac{1-p}{\beta}\right)\right) + (1-p)\left(\beta\psi_1\left(\frac{-p}{\beta}\right)\right) \\ = &p\beta\left(\frac{1-p}{\beta} + \frac{1}{k!}\psi_1^{(k)}(0)\frac{(1-p)^k}{\beta^k} + o\left(\frac{1}{\beta^k}\right)\right) \\ &- (1-p)\beta\left(\frac{p}{\beta} + \frac{1}{k!}\psi_1^{(k)}(0)\frac{p^k}{\beta^k} + o\left(\frac{1}{\beta^k}\right)\right).\end{aligned}$$

C.6. Proof of Lemma 6

Step 1. For any ψ that satisfies Assumptions 2, $V_\psi \leq \mathrm{Tr}(\Sigma)$.

Proof. First, remark that we have for all $x \in \mathbb{R}_+$, $\psi_1^2(x) \leq 2\rho_1(x)$. Indeed, let $h(x) = \psi_1^2(x) - 2\rho_1(x)$, its derivative is $h'(x) = 2\psi_1(x)(\psi_1'(x) - 1)$ and because $\psi_1' \leq 1$ and $\psi_1(0) = 0$, we get that h is decreasing, the fact that $h(0) = 0$ implies that for all $x \in \mathbb{R}_+$, $\psi_1^2(x) \leq 2\rho_1(x)$. Then,

$$V_\psi \leq 2\beta^2 \mathbb{E}\left[\rho_1\left(\frac{\|X - T(P)\|}{\beta}\right)\right]. \tag{53}$$

Define $J(\theta) = \mathbb{E}\left[\rho_1\left(\frac{\|X-\theta\|}{\beta}\right)\right]$ by definition 1, $T(P)$ is the minimum of J and by equation (53),

$$V_\psi \leq 2\beta^2 \mathbb{E}\left[\rho_1\left(\frac{\|X - \mathbb{E}[X]\|}{\beta}\right)\right]$$

Then finally, using that by integration of $\psi_1' \leq 1$ we have $\rho_1(x) \leq x^2/2$, hence the result. □

Step 2. For any ψ that satisfies Assumptions 2, $x_\psi \leq \|\Sigma\|_{op} + \|\mathbb{E}[X] - T(P)\|^2$.

Proof. Note that because $\psi_1(x) \leq 2\rho_1(x) \leq x^2$,

$$\begin{aligned}
v_\psi &= \beta^2 \sup_{u \in S} \mathbb{E}\left[\frac{\langle u, X - T(P)\rangle^2}{\|X - T(P)\|^2} \psi_1\left(\frac{\|X - T(P)\|}{\beta}\right)^2\right] \\
&\leq \beta^2 \sup_{u \in S} \mathbb{E}\left[\frac{\langle u, X - T(P)\rangle^2}{\|X - T(P)\|^2} \left(\frac{\|X - T(P)\|}{\beta}\right)^2\right] \\
&= \sup_{u \in S} \mathbb{E}\left[\langle u, X - T(P)\rangle^2\right] \\
&= \sup_{u \in S} \mathbb{E}\left[\left(\langle u, X - \mathbb{E}[X]\rangle + \langle u, \mathbb{E}[X] - T(P)\rangle\right)^2\right] \\
&= \sup_{u \in S} \mathbb{E}\left[\langle u, X - \mathbb{E}[X]\rangle^2 + \langle u, \mathbb{E}[X] - T(P)\rangle^2\right] \\
&= \|\Sigma\|_{op} + \|\mathbb{E}[X] - T(P)\|^2
\end{aligned}$$

□

C.7. Proof of Lemma 7

Step 1. if X has $q > 2$ finite moments, then

$$\begin{aligned}
V_{\psi_H} &\geq \mathbb{E}[\|X - T_H(P)\|^2] \\
&- \mathbb{E}\left[\left(\|X - T_H(P)\|^2 + \beta^2\right)^q\right]^{1/q} \mathbb{P}\left(\|X - T_H(P)\| > \beta\right)^{1-1/q}.
\end{aligned} \tag{54}$$

Proof. We have

$$\begin{aligned} V_{\psi_H} &= \mathbb{E}\left[\psi_H\left(\|X - T_H(P)\|\right)^2\right] \\ &= \mathbb{E}[\beta^2 \wedge (\|X - T_H(P)\|)^2] \\ &= \mathbb{E}[\|X - T_H(P)\|^2] - \mathbb{E}[\left(\|X - T_H(P)\|^2 + \beta^2\right) \\ &\qquad \times \mathbb{1}\{\|X - T_H(P)\| > \beta\}] \end{aligned}$$

Then, by Hölder inequality, we get the result announced. □

Step 2. We have $V_{\psi_H} \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$.

Proof. We use the following lemma

Lemma 16. *Let Y be a positive real random variable, $\mathbb{E}[Y^q] < \infty$. We have for all $\lambda > 0$,*

$$\mathbb{P}\left(Y \geq \lambda\right) \leq 2^{q-1} \frac{\mathbb{E}\left[Y^q\right]}{\lambda^q + \mathbb{E}[Y^q]}$$

See Section C.16 for the proof. Then, for $Y = \|X - T_H(P)\|$ and from Step 1, we get,

$$\begin{aligned} V_{\psi_H} &\geq \mathbb{E}[\|X - T_H(P)\|^2] \\ &- \mathbb{E}\left[\left(\|X - T_H(P)\|^2 + \beta^2\right)^q\right]^{1/q} \left(2^{2q-1} \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]}{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}}\right)^{1-1/q}. \end{aligned}$$

Use the fact that $(a+b)^q \leq 2^{q-1}(a^q + b^q)$,

$$\begin{aligned} V_{\psi_H} \geq& \mathbb{E}[\|X - T_H(P)\|^2] \\ &- 2^{(q-1)/q + (2q-1)(1-1/q)} \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}} \\ \geq& \mathbb{E}[\|X - T_H(P)\|^2] - 2^{2q} \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}} \end{aligned}$$

And finally, because $\mathbb{E}[X]$ is the minimizer of the quadratic loss,

$$V_{\psi_H} \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}}.$$

□

Step 3. Similarly, if X has $q > 2$ finite moments, we have$v_{\psi_H} \geq \|\Sigma\|_{op} - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$.

Proof. Then, we operate the same manner for the bound on v_{ψ_H}. We have,

$$\begin{aligned} v_{\psi_H} =& \sup_{u\in S} \mathbb{E}\left[\frac{\langle u, X - T_H(P)\rangle^2}{\|X - T_H(P)\|^2}\left(\beta^2 \wedge (\|X - T_H(P)\|)^2\right)\right] \\ =& \sup_{u\in S}\Bigg(\mathbb{E}\left[\langle u, X - T_H(P)\rangle^2\right] \\ & - \mathbb{E}\Bigg[\frac{\langle u, X - T_H(P)\rangle^2}{\|X - T_H(P)\|^2}\left(\beta^2 - (\|X - T_H(P)\|)^2\right) \\ & \times \mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\Bigg]\Bigg) \end{aligned}$$

Then, use Cauchy-Schwarz inequality,

$$\begin{aligned} v_{\psi_H} \geq& \sup_{u\in S} \mathbb{E}\left[\langle u, X - T_H(P)\rangle^2\right] \\ & - \mathbb{E}\left[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\right] \\ \geq& \sup_{u\in S}\Big(\mathbb{E}\left[\langle u, X - \mathbb{E}[X]\rangle^2\right] + \langle u, \mathbb{E}[X] - T_H(P)\rangle^2 \\ & - -\mathbb{E}\left[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\right]\Big) \\ \geq& \sup_{u\in S} \mathbb{E}\left[\langle u, X - \mathbb{E}[X]\rangle^2\right] \\ & - \mathbb{E}\left[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\right] \end{aligned}$$

Then, use the same reasoning as for the bound on V_{ψ_H} to conclude that

$$v_{\psi_H} \geq \|\Sigma\|_{op} - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right] + \beta^{2q}\right)^{1-2/q}}$$

□

C.8. Proof of Lemma 8

From [1, Theorem 4] and because $\|Y\| = \sup_{\|u\|=1}\langle Y, u\rangle$, there exists an absolute constant C_1 such that, for all $t \geq 0$,

$$\begin{aligned} \mathbb{P}\left(\left\|\sum_{i=1}^n Y_i\right\| \geq \frac{3}{2}\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|\right] + t\right) & \\ \leq \exp\left(-\frac{t^2}{4n\sigma^2}\right) + 3\exp\left(-\frac{t}{C_1\|\max_{1\leq i\leq n}\|Y_i\|\|_{\psi_1}}\right). & \end{aligned}$$

where $\sigma^2 = n\sup_{u\in S}\mathbb{E}[\langle X, u\rangle^2]$. Remark that σ^2 can be rewritten

$$\sigma^2 = n\sup_{u\in S}\langle u\mathbb{E}[X\otimes X], u\rangle = n\|\Sigma\|_{op}.$$

Then, by Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} Y_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{n} Y_i\right\|^2\right]^{1/2} = \sqrt{n}\mathbb{E}\left[\|Y\|^2\right]^{1/2}.$$

C.9. Proof of Lemma 9

ρ is convex because $\rho'' = \psi' \geq 0$ and it is increasing because $\psi = \rho' \geq 0$ ($\psi(0) = 0$ and ψ increasing). Then, from triangular inequality and Jensen's inequality, we have

$$\rho\left(\frac{\|\mathbb{E}[X] - T(P)\|}{\beta}\right) \leq \rho\left(\frac{\mathbb{E}[\|X - T(P)\|]}{\beta}\right) \leq \mathbb{E}\left[\rho\left(\frac{\|X - T(P)\|}{\beta}\right)\right].$$

By definition of $T(P)$, it is a minimizer of $\theta \mapsto \mathbb{E}\left[\rho\left(\frac{\|X-\theta\|}{\beta}\right)\right]$, hence,

$$\rho\left(\frac{\|\mathbb{E}[X] - T(P)\|}{\beta}\right) \leq \mathbb{E}\left[\rho\left(\frac{\|X - \mathbb{E}[X]\|}{\beta}\right)\right]$$

then, use the hypothesis to upper bound the right-hand side by $\rho(1/3)$, we get

$$\rho\left(\frac{\|\mathbb{E}[X] - T(P)\|}{\beta}\right) \leq \rho(1/3).$$

Finally, because ρ is non-decreasing on $\mathbb{R}_+$ (its derivative is non-negative), we get the result.

C.10. Proof of Lemma 10

First, let us begin with $d = 1$. We have that for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left(Med(X_1^n) - \mathbb{E}_P[X] > t\right) &\leq \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i - \mathbb{E}_P[X] > t\} \geq \frac{n}{2}\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i' - \mathbb{E}_P[X] > t\} \geq \frac{n}{2} - |\mathcal{O}|\right) \qquad (55)\end{aligned}$$

By Hoeffding's inequality, we have

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i - \mathbb{E}_P[X] > t\} \geq \frac{n}{2} - |\mathcal{O}|\right) \\ &\qquad \leq \exp\left(-2n\left(\frac{1}{2} - \frac{|\mathcal{O}|}{n} - \mathbb{P}\left(X - \mathbb{E}_P[X] > t\right)\right)^2\right)\end{aligned}$$

and by Chebychev inequality, for the choice $t = 2\sqrt{2}\sigma$ we have $\mathbb{P}\left(X - \mathbb{E}_P[X] > t\right) \leq 1/8$. Then, from Equation (55),

$$\mathbb{P}\left(Med(X_1^n) - \mathbb{E}_P[X] > 2\sqrt{2}\sigma\right) \leq \exp\left(-2n\left(\frac{1}{2} - \frac{|\mathcal{O}|}{n} - \frac{1}{8}\right)^2\right).$$

and then, because $|\mathcal{O}| \leq n/8$,

$$\mathbb{P}\left(Med(X_1^n) - \mathbb{E}_P[X] > 2\sqrt{2}\sigma\right) \leq \exp\left(-n/8\right).$$

Now for dimension d, we use the dimension 1 result on each coordinate, and by union bound we have that with probability larger than $1 - de^{-n/8}$, for all $1 \leq j \leq d$,

$$Med(\langle X_1^n, e_j\rangle) - \mathbb{E}_P[\langle X, e_j\rangle] \leq 2\sqrt{2}\sigma_j$$

and then, by taking the sum of the squares, we get $\|\theta_0 - \mathbb{E}_P[X]\|^2 \leq 8\sum_{j=1}^d \sigma_j^2 = 8Tr(\Sigma)$. We conclude using that $\|T(X_1^n) - \mathbb{E}_P[X]\| \leq r_n$ with probability larger than $1 - \delta$.

C.11. Proof of Lemma 11

The proof is derived from the proof of iterative reweighting algorithm for regression found in [27, Section 7.8].

First point.

We have

$$\mathbf{U}_{\theta^{(m)}}(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\frac{w_i(\theta^{(m)})}{2}\left(\frac{\|X_i - \theta\|}{\beta}\right)^2 + \rho_1(r_i(\theta^{(m)})) - \frac{1}{2}r_i(\theta^{(m)})\psi_1(r_i(\theta^{(m)}))\right).$$

$\mathbf{U}_{\theta^{(m)}}$ is a convex function, let us take its gradient to find its minimum,

$$\nabla\mathbf{U}_{\theta^{(m)}}(\theta) = \frac{1}{n}\sum_{i=1}^n w_i(\theta^{(m)})\frac{\theta - X_i}{\beta^2}$$

Hence, the minimum is found for $\theta = \sum_{i=1}^n \frac{w_i(\theta^{(m)})}{\sum_{j=1}^n w_j(\theta^{(m)})}X_i = \theta^{(m+1)}$.

Second point.

For all $i \in \{1, \ldots, n\}$,

$$g_i(x) = \frac{w_i(\kappa)}{2}x^2 + \rho_1(r_i(\kappa)) - \frac{1}{2}r_i(\kappa)\psi_1(\kappa) - \rho_1(x)$$

We have that g_i is differentiable and $g_i'(x) = w_i(\kappa)x - \psi_1(x) = \frac{\psi_1(r_i(\kappa))}{r_i(\kappa)}x - \psi_1(x)$. Then, having that $x \mapsto \psi_1(x)/x$ is non-increasing on $\mathbb{R}_+$, we have that for

$x \in [0, r_i(\kappa)]$, $g_i'(x) \geq 0$ and for $x \geq r_i(\kappa)$, $g_i'(x) \leq 0$. Hence, g_i is minimal in $x = r_i(\kappa)$ and $g_i(x) \geq g_i(r_i(\kappa)) = 0$.

This prove that g_i is a majorant of ρ_1 and by taking the sum, this implies that $\mathbf{U}_\kappa$ is a majorant of J_n.

Third point.

Define

$$h_{i,\kappa}(\theta) = \frac{w_i(\kappa)}{2}\left(\frac{\|X_i - \theta\|}{\beta}\right)^2 + \rho_1(r_i(\kappa)) - \frac{1}{2}r_i(\kappa)\psi_1(r_i(\kappa)) - \rho_1\left(\frac{\|X_i - \theta\|}{\beta}\right),$$

we have by definition of $w_i(\kappa)$,

$$h_{i,\kappa}(\kappa) = \frac{w_i(\kappa)}{2}r_i(\kappa)^2 + \rho_1(r_i(\kappa)) - \frac{1}{2}r_i(\kappa)\psi_1(r_i(\kappa)) - \rho_1(r_i(\kappa)) = 0 \tag{56}$$

and moreover, h_i is a differentiable function whose gradient is

$$\nabla h_{i,\kappa}(\theta) = w_i(\kappa)\frac{(\theta - X_i)}{\beta^2} - \frac{\theta - X_i}{\beta\|\theta - X_i\|}\psi_1\left(\frac{\|X_i - \theta\|}{\beta}\right),$$

and we can verify that

$$\nabla h_{i,\kappa}(\kappa) = w_i(\kappa)\frac{(\kappa - X_i)}{\beta^2} - \frac{\kappa - X_i}{\beta^2 r_i}\psi\left(r_i\right) = 0. \tag{57}$$

Let us show that ∇h_κ is Lipshitz. We have

$$Hess(h_{i,\kappa}) = w_i(\kappa)\frac{I_d}{\beta^2} - Hess(\theta \mapsto \rho_1(\|X_i - \theta\|))$$

Then, use that $\theta \mapsto \rho_1(\|X_i - \theta\|)$ is convex, hence its Hessian is positive and we have for all $u \in S$,

$$u^T Hess(h_{i,\kappa})u \leq \frac{w_i}{\beta^2}$$

This conclude that $\nabla h_{i,\kappa}$ is w_i/β^2-Lipshitz continuous and hence by summing over i, ∇h_κ is Lipshitz continuous with Lipshitz constant $L = \frac{1}{n\beta^2}\sum_{i=1}^n \frac{\psi_1(r_i)}{r_i} \leq 1/\beta^2$.

Fourth point already verified using Equations (56) and (57).

C.12. Proof of Lemma 12

The function $\mathbf{U}_\kappa$ can be rewritten as follows:

$$\begin{aligned}\mathbf{U}_\kappa(\theta) =& \frac{1}{n}\sum_{i=1}^n \left(\frac{w_i(\kappa)}{2}\left(\frac{\|X_i - \theta\|}{\beta}\right)^2 + \rho_1(r_i(\kappa)) - \frac{1}{2}r_i(\kappa)\psi_1(r_i(\kappa))\right) \\ =& \frac{1}{n}\sum_{i=1}^n \frac{w_i(\kappa)}{2}\frac{\|X_i - \kappa\|^2 + \|\kappa - \theta\| + 2\langle X_i - \kappa, \kappa - \theta\rangle}{\beta^2}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n}\left(\rho_1(r_i(\kappa))-\frac{1}{2}r_i(\kappa)\psi_1(r_i(\kappa))\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{r_i(\kappa)\psi_1(r_i(\kappa))}{2}+\frac{w_i(\kappa)}{2}\frac{\|\kappa-\theta\|+2\langle X_i-\kappa,\kappa-\theta\rangle}{\beta^2}\\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\rho_1(r_i(\kappa))-\frac{1}{2}r_i(\kappa)\psi_1(r_i(\kappa))\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\rho_1(r_i(\kappa))+\frac{\langle X_i-\kappa,\kappa-\theta\rangle}{\beta\|X_i-\kappa\|}\psi_1(r_i(\kappa))+\frac{w_i(\kappa)\|\kappa-\theta\|}{2\beta^2}\right)\\
=&J(\kappa)+2\langle\nabla J(\kappa),\theta-\kappa\rangle+\frac{1}{2n\beta^2}\sum_{i=1}^{n}w_i(\kappa)\|\theta-\kappa\|^2
\end{aligned}
$$

Then, let $f(\theta)=\sum_{i=1}^{n}\frac{w_i(\theta)}{\sum_{j=1}^{n}w_j(\theta)}X_i$ where $w_i(\theta)=\psi_1(r_i(\theta))$ and $r_i(\theta)=\|X_i-\theta\|/\beta$. Inject $\theta=f(\kappa)$ to get,

$$
\mathbf{U}_\kappa(f(\kappa))=J(\kappa)+2\langle\nabla J(\kappa),f(\kappa)-\kappa\rangle+\frac{1}{2n\beta^2}\sum_{i=1}^{n}w_i(\kappa)\|f(\kappa)-\kappa\|^2
$$

Then, use that $\mathbf{U}_\kappa$ is a majorant of J (Lemma 11) and apply this to $\kappa=\theta^{(m)}$ to obtain

$$
\begin{aligned}
J(\theta^{(m+1)})\leq&\, J(\theta^{(m)})\\
&+\langle\nabla J(\theta^{(m)}),\theta^{(m+1)}-\theta^{(m)}\rangle+\frac{\sum_{j=1}^{n}w_j(\theta^{(m)})}{2n\beta^2}\|\theta^{(m+1)}-\theta^{(m)}\|^2
\end{aligned}
$$

Then, because J_n is convex, we have that for all $\theta\in\mathbb{R}^d$, $J(\theta^{(m)}\leq J(\theta)+\langle\nabla J(\theta^{(m)}),\theta^{(m)}-\theta\rangle$, hence

$$
\begin{aligned}
&J(\theta^{(m+1)})\\
\leq&J(\theta)+\langle\nabla J(\theta^{(m)}),\theta^{(m+1)}-\theta\rangle+\frac{\sum_{j=1}^{n}w_j(\theta^{(m)})}{2n\beta^2}\|\theta^{(m+1)}-\theta^{(m)}\|^2\\
=&J(\theta)+\langle\nabla U_{\theta^{(m)}}(\theta^{(m)}),\theta^{(m+1)}-\theta\rangle+\frac{\sum_{j=1}^{n}w_j(\theta^{(m)})}{2n\beta^2}\|\theta^{(m+1)}-\theta^{(m)}\|^2\\
=&J(\theta)+\left\langle\frac{1}{n}\sum_{i=1}^{n}w_i(\theta^{(m)})(X_i-\theta^{(m)}),\theta^{(m+1)}-\theta\right\rangle\\
&+\frac{\sum_{j=1}^{n}w_j(\theta^{(m)})}{2n\beta^2}\|\theta^{(m+1)}-\theta^{(m)}\|^2\\
=&J(\theta)+\sum_{j=1}^{n}w_j(\theta^{(m)})\left\langle\theta^{(m+1)}-\theta^{(m)},\theta^{(m+1)}-\theta\right\rangle
\end{aligned}
$$

$$+\frac{\sum_{j=1}^n w_j(\theta^{(m)})}{2n\beta^2}\|\theta^{(m+1)}-\theta^{(m)}\|^2$$

Then, use the following identity: for all $u,v,w\in\mathbb{R}^d$,

$$2\langle w-v,v-u\rangle=\|w-v\|^2-\|w-u\|^2+\|u-v\|^2$$

to conclude that

$$J(\theta^{(m+1)})\leq J(\theta)+\frac{1}{2n\beta^2}\sum_{j=1}^n w_j(\theta^{(m)})\left(\|\theta^{(m)}-\theta\|-\|\theta^{(m+1)}-\theta\|\right)$$

C.13. Proof of Lemma 13

By Assumption 2-(iv), we have

$$\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{\|X_i-\theta\|}{\beta}\right)\geq\frac{\gamma}{n}\sum_{i=1}^n\mathbb{1}\{\|X_i-\theta\|\leq\beta\}$$

Then, because $\theta\in\Theta$ and $\beta\geq 2\sqrt{2Tr(\Sigma)}+2r_n+\psi_1^{-1}\left(\sqrt{2V_{\psi_1}}\right)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{\|X_i-\theta\|}{\beta}\right)&\geq\frac{\gamma}{n}\sum_{i=1}^n\mathbb{1}\{\|X_i-T(X_1^n)\|\leq\beta-2\sqrt{2Tr(\Sigma)}-r_n\}\\&\geq\frac{\gamma}{n}\sum_{i=1}^n\mathbb{1}\{\|X_i-T(X_1^n)\|\leq\psi_1^{-1}\left(\sqrt{2V_{\psi_1}}\right)+r_n\}\end{aligned}$$

Now, having that with probability larger than $1-\delta$, $\|T(X_1^n)-T(P)\|\leq r_n$, we have with probability larger than $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{\|X_i-\theta\|}{\beta}\right)\geq\frac{\gamma}{n}\sum_{i=1}^n\mathbb{1}\{\|X_i-T(P)\|\leq\psi_1^{-1}\left(\sqrt{2V_{\psi_1}}\right)\}.$$

Then, as we don't have informations on the outliers, we take them out:

$$\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{\|X_i-\theta\|}{\beta}\right)\geq\frac{\gamma}{n}\sum_{i=1}^n\mathbb{1}\left\{\|X_i'-T(P)\|\leq\psi_1^{-1}\left(\sqrt{2V_{\psi_1}}\right)\right\}-\gamma\frac{|\mathcal{O}|}{n}$$

Then, by Hoeffding's inequality, we have that for $t>0$, with probability larger than $1-\delta-\exp(-2nt^2)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{\|X_i-\theta\|}{\beta}\right)&\geq\gamma\mathbb{P}\left(\|X-T(P)\|\leq\psi_1^{-1}\left(\sqrt{2V_{\psi_1}}\right)\right)-\gamma t-\gamma\frac{|\mathcal{O}|}{n}\\&\geq\gamma\mathbb{P}\left(\psi_1(\|X-T(P)\|)^2\leq 2V_{\psi_1}\right)-\gamma t-\gamma\frac{|\mathcal{O}|}{n}\end{aligned}$$

By Markov inequality and having $|\mathcal{O}| \leq n/8$, we have that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_1'\left(\frac{\|X_i - \theta\|}{\beta}\right) \geq \gamma\left(\frac{1}{2} - t - \frac{1}{8}\right).$$

Hence, with probability larger than $1 - \delta - e^{-n/32}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \psi_1'\left(\frac{\|X_i - \theta\|}{\beta}\right) \geq \frac{\gamma}{4}.$$

C.14. Proof of Lemma 14

For $q \in \mathbb{N}^*$, let

$$g_q : x \mapsto \begin{cases} q^q x/(e^q - 1) & if\ x \in [0, e^q - 1] \\ \log(1+x)^q & if\ x > e^q - 1 \end{cases}$$

Step 1. g_q is a concave function over $\mathbb{R}_+$.

Proof. g_q is continuous at $e^q - 1$, the left and right limits are equal to q^q. g_q is derivable on $[0, e^q - 1)$ and $(e^q - 1, \infty)$. This derivative is non-increasing on both intervals. At $e^q - 1$, the left derivative is $q^q(e^q - 1)^{-1}$ while the derivative on the right is $q^q e^{-q}$. Thus, the left derivative at $e^q - 1$ is larger than the right derivative. Hence the derivative is non-increasing on $\mathbb{R}_+$, g_q is concave on $\mathbb{R}_+$. □

Step 2. $\mathbb{E}[\log(1+Z)^q] \leq \mathbb{E}[g_q(Z)] \leq g_q(\mathbb{E}[Z])$.

Proof. By concavity, $\log(1+x)^q$ is smaller than its tangent in $e^{q-1} - 1$. This tangent is given by the function

$$x \mapsto (q-1)^q + \frac{q^q}{e^{q-1} - 1}\left(x - (e^{q-1} - 1)\right).$$

This last function is clearly smaller than the function $x \mapsto q^q x/(e^q - 1)$. Hence, $x \mapsto \log(1+x)^q$ is smaller than g_q, we found a concave upper bound of $x \mapsto \log(1+x)^q$.
Since g_q is concave (Step 1), by Jensen's inequality, for any positive random variable Z such that $\mathbb{E}[Z] < \infty$, we have $\mathbb{E}[\log(1+Z)^q] \leq \mathbb{E}[g_q(Z)] \leq g_q(\mathbb{E}[Z])$. □

Step 3. For any $q \in \mathbb{N}^*$, $\mathbb{E}[\psi_C(\|X - T_C(P)\|)^q] \leq q!(\beta s)^q$.

Proof. Then, for all x, we have $g_q(x) \leq \max(q^q, \log(1+x)^q)$, hence, by Step 2

$$\mathbb{E}[\log(1+Z)^q] \leq \max(q^q, \log(1 + \mathbb{E}[Z])^q).$$

Finally, use that $q^q \leq q! e^q$ to get

$$\mathbb{E}[\log(1+Z)^q] \leq q! \max(e, \log(1 + \mathbb{E}[Z]))^q. \tag{58}$$

Denote $s = \max\left(e, \log\left(1 + \frac{\mathbb{E}[\|X - T_C(P)\|])}{\beta} + \frac{\mathbb{E}[\|X - T_C(P)\|^2]}{2\beta^2}\right)\right)$, and apply equation (58) to $Z = X/\beta + X^2/(2\beta^2)$ to get $\mathbb{E}[\psi_C(\|X - T_C(P)\|)^q] \leq q!(\beta s)^q$. □

C.15. Proof of Lemma 15

By Markov's inequality, we have for any $\lambda > 0$,

$$t_{\mathrm{IF}}(\lambda; T_P(P), X_1^n) \leq \frac{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \frac{X_i - T_P(P)}{\|X_i - T_P(P)\|}\psi_P(\|X_i - T_P(P)\|)\right\|^{qp}\right]}{\lambda^{qp}}. \tag{59}$$

Let $Y_i = \frac{1}{n}\sum_{i=1}^n \frac{X_i - T_P(P)}{\|X_i - T_P(P)\|}\psi_P(\|X_i - T_P(P)\|)$ for $1 \leq i \leq d$. from [10, Theorem 1.2.5], there exists an absolute constant $K > 0$ such that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^{pq}\right]^{1/(pq)} \leq Kpq\left(\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right]^{1/2} + \mathbb{E}\left[\max_{1\leq i\leq n}\|Y_i\|^{pq}\right]^{1/(pq)}\right). \tag{60}$$

Step 1. We have $\mathbb{E}\left[\|\sum_{i=1}^n Y_i\|^2\right] \leq 4n\mathbb{E}[\|Y\|^{pq}]^{2/(pq)}$.

Proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d Rademacher random variable independents from $Y_1, \ldots, Y_n$. By the symmetrization lemma (see [10, Lemma 1.2.6]),

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right] \leq 4\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i Y_i\right\|^2\right] = 4\mathbb{E}\left[\sum_{i=1}^n \|Y_i\|^2\right] = 4n\mathbb{E}[\|Y\|^2].$$

Thus, by Jensen's inequality, $\mathbb{E}\left[\|\sum_{i=1}^n Y_i\|^2\right] \leq 4n\mathbb{E}[\|Y\|^{pq}]^{2/(pq)}$. □

Step 2. $\mathbb{E}\left[\max_{1\leq i\leq n}\|Y_i\|^{pq}\right] \leq n^{pq/2}\mathbb{E}\left[\|Y\|^{pq}\right]$.

Proof. As the max of n non-negative real numbers is smaller than their sum, we have

$$\mathbb{E}\left[\max_{1\leq i\leq n}\|Y_i\|^{pq}\right] \leq \mathbb{E}\left[\sum_{1\leq i\leq n}\|Y_i\|^{pq}\right] \leq n\mathbb{E}\left[\|Y\|^{pq}\right] \leq n^{pq/2}\mathbb{E}\left[\|Y\|^{pq}\right].$$

□

Step 3. For any $\lambda > 0$, $t_{\mathrm{IF}}(\lambda; T_P(P), X_1^n) \leq \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q}\left(\frac{Kpq\beta}{\sqrt{n}\lambda}\right)^{qp}$.

Proof. From Step 1 and Step 2 and equation (60), we get

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^{pq}\right]^{1/(pq)} \leq 3Kpq\sqrt{n}\mathbb{E}\left[\|Y\|^{pq}\right]^{1/(pq)}. \tag{61}$$

From equations (59) and (61) and if we re-inject the definition of Y_i's, we get

$$t_{\mathrm{IF}}(\lambda; T_P(P), X_1^n) \leq \mathbb{E}[\psi_P(\|X - T_P(P)\|)^{pq}] \left(\frac{Kpq}{\sqrt{n}\lambda}\right)^{qp}.$$

Then, use that $\psi_p(\|x\|) \leq \|x\|^{1/p}\beta^{1-1/p}$ to get

$$t_{\mathrm{IF}}(\lambda; T_P(P), X_1^n) \leq \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q} \left(\frac{Kpq\beta}{\sqrt{n}\lambda}\right)^{qp}.$$

□

C.16. Proof of Lemma 16

We have for all $u, \lambda > 0$,

$$\mathbb{P}(Y \geq \lambda) = \mathbb{P}((Y+u)^q \geq (\lambda+u)^q) \leq \frac{\mathbb{E}[(Y+u)^q]}{(\lambda+u)^q} \leq \frac{\mathbb{E}[(Y+u)^q]}{(\lambda^{q/2}+u^{q/2})^2}.$$

Then, use that by convexity of the q^{th}-power function, $(a+b)^q \leq 2^{q-1}(a^q+b^q)$ and also $(a+b)^q \geq a^q + b^q$,

$$\mathbb{P}(Y \geq \lambda) \leq 2^{q-1}\frac{\mathbb{E}[Y^q + u^q]}{(\lambda^{q/2}+u^{q/2})^2}.$$

Take $u = \mathbb{E}[Y^q]^{2/q}/\lambda$ to get,

$$\mathbb{P}(Y \geq \lambda) \leq 2^{q-1}\frac{\mathbb{E}[Y^q] + \frac{\mathbb{E}[Y^q]^2}{\lambda^q}}{\lambda^q(1+\frac{\mathbb{E}[Y^q]}{\lambda^q})^2} = 2^{q-1}\frac{\mathbb{E}[Y^q]}{\lambda^q(1+\frac{\mathbb{E}[Y^q]}{\lambda^q})} = 2^{q-1}\frac{\mathbb{E}[Y^q]}{\lambda^q + \mathbb{E}[Y^q]}.$$

Appendix D: Technical tools

We remind the reader of Bernstein inequality, a classical concentration inequality, this form of Bernstein inequality is borrowed from [4, Theorem 2.10].

Theorem 5. *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v$ and*

$$\sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq vc^{q-2} \quad \textit{for all integers } q \geq 3,$$

where $x_+ = \max(0, x)$. Then for all $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq \sqrt{2vt} + ct\right) \leq e^{-t}.$$

The following theorem is borrowed from [1, Theorem 4], it is a concentration inequality for suprema of sums of independent random variables.

Theorem 6. *Let $X_1, \dots, X_n$ be independent random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. Assume that for every $f \in \mathcal{F}$ and every i, $\mathbb{E}[f(X_i)] = 0$ and for any $\alpha \in (0, 1]$ and all i, $\| \sup_f |f(X_i)| \|_{\psi_\alpha} < \infty$. Let*

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right|.$$

Define moreover

$$\sigma^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}[f(X_i)^2].$$

Then, for all $0 < \eta < 1$ and $\delta > 0$, there exists a constant $C = C(\alpha, \eta, \delta) > 0$ such that for all $t \geq 0$,

$$\mathbb{P}(Z \geq (1+\eta)\mathbb{E}[Z] + t) \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\| \max_i \sup_{f\in\mathcal{F}} |f(X_i)| \|_{\psi_\alpha}}\right)^{\alpha}\right), \quad (62)$$

and

$$\mathbb{P}(Z \leq (1-\eta)\mathbb{E}[Z] - t) \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\| \max_i \sup_{f\in\mathcal{F}} |f(X_i)| \|_{\psi_\alpha}}\right)^{\alpha}\right). \quad (63)$$

Remark that because of the need for a countable set $\mathcal{F}$, this Theorem can only be applied to separable spaces.

Acknowledgments

The author would like to thank Elvezio Ronchetti, for the discussions that gave him the ideas behind this article, and his two Ph.D.-advisors Matthieu Lerasle and Guillaume Lecué for their advice and reviews.

References

[1] Adamczak, R. et al. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability* **13** 1000–1034. MR2424985

[2] Bahadur, R. R. (1966). A Note on Quantiles in Large Samples. *Ann. Math. Statist.* **37** 577–580.

[3] Beck, A. and Sabach, S. (2015). Weiszfeld's method: Old and new results. *Journal of Optimization Theory and Applications* **164** 1–40. MR3296283
[4] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford. MR3185193
[5] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.* **48** 1148–1185. MR3052407
[6] Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. (https://arxiv.org/abs/1712.02747).
[7] Chen, M., Gao, C., Ren, Z. et al. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics* **46** 1932–1960. MR3845006
[8] Cherapanamjeri, Y., Flammarion, N. and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory* 786–806. PMLR.
[9] Chinot, G., Lecué, G. and Lerasle, M. (2019). Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and related fields* 1–44. MR4087486
[10] De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media, New York. MR1666908
[11] Debruyne, M., Hubert, M. and Suykens, J. A. (2008). Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research* **9** 2377–2400. MR2452631
[12] Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time.
[13] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics* **44** 2695–2725. MR3576558
[14] Diakonikolas, I. and Kane, D. M. (2019). Recent Advances in Algorithmic High-Dimensional Robust Statistics.
[15] Diakonikolas, I., Kane, D. M. and Pensia, A. (2020). Outlier Robust Mean Estimation with Subgaussian Rates via Stability.
[16] Elisseeff, A. and Pontil, M. (2002). Leave-one-out error and stability of learning algorithms with applications Stability of Randomized Learning Algorithms Source. *International Journal of Systems Science - IJSySc* **6**. MR2249815
[17] Fernholz, L. T. (1983). *von Mises calculus for statistical functionals*. *Lecture Notes in Statistics* **19**. Springer-Verlag, New York. MR713611
[18] Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *Ann. Math. Statist.* **42** 1887–1896. MR0301858
[19] Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association* **69** 383–393. MR0362657

[20] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, 1st edition ed. *Wiley Series in Probability and Statistics*. Wiley missing. MR0829458

[21] Hanson, D. L. and Wright, F. T. (1971). A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables. *Ann. Math. Statist.* **42** 1079–1083. MR0279864

[22] He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M -estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971

[23] Hopkins, S., Li, J. and Zhang, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems* **33** 11902–11912.

[24] Hopkins, S. B. (2020). Mean Estimation with Sub-Gaussian Rates in Polynomial Time. *Annals of Statistics* **48** 1193–1213. (https://arxiv.org/abs/1809.07425). MR4102693

[25] Hopkins, S. B. and Li, J. (2019). How Hard Is Robust Mean Estimation? *CoRR* **abs/1903.07870**.

[26] Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[27] Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics; 2nd ed. Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795

[28] Koh, P. W. and Liang, P. (2017). Understanding Black-box Predictions via Influence Functions.

[29] Lai, K. A., Rao, A. B. and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* 665–674. IEEE. MR3631029

[30] Lerasle, M. (2019). Lecture Notes: Selected topics on robust statistical learning theory.

[31] Lugosi, G. and Mendelson, S. (2019). Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey. *Foundations of Computational Mathematics* **19** 1145–1190. MR4017683

[32] Lugosi, G., Mendelson, S. et al. (2019). Sub-Gaussian estimators of the mean of a random vector. *The annals of statistics* **47** 783–794. MR3909950

[33] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

[34] Minsker, S. (2018). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

[35] Minsker, S. and Mathieu, T. (2019). Excess risk bounds in robust empirical risk minimization. MR4353129

[36] Ronchetti, E. (1997). Robust inference by influence functions. *Journal of Statistical Planning and Inference* **57** 59–72. Robust Statistics and Data Analysis, Part I. MR1440230

[37] Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association* **115** 254–265. MR4078461

[38] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*

and empirical processes. Springer Series in Statistics. Springer-Verlag, New York With applications to statistics. MR1385671

[39] Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46** 1904–1931. MR3845005

[40] Zhu, B., Jiao, J. and Steinhardt, J. (2020). Robust estimation via generalized quasi-gradients. *arXiv preprint arXiv:2005.14073*.