Convergence of the Huber Regression M-Estimate in the Presence of Dense Outliers

Efthymios Tsakonas, Joakim Jaldén, Nicholas D. Sidiropoulos, and Björn Ottersten

***Abstract*—We consider the problem of estimating a deterministic unknown vector which depends linearly on n noisy measurements, additionally contaminated with (possibly unbounded) additive outliers. The measurement matrix of the model (i.e., the matrix involved in the linear transformation of the sought vector) is assumed known, and comprised of standard Gaussian i.i.d. entries. The outlier variables are assumed independent of the measurement matrix, deterministic or random with possibly unknown distribution. Under these assumptions we provide a simple proof that the minimizer of the Huber penalty function of the residuals converges to the true parameter vector with a $\sqrt{n}$-rate, even when outliers are dense, in the sense that there is a constant linear fraction of contaminated measurements which can be arbitrarily close to one. The constants influencing the rate of convergence are shown to explicitly depend on the outlier contamination level.**

***Index Terms*—Breakdown point (BP), dense outliers, Huber estimator, performance analysis.**

I. Introduction

It is often the case in robust statistical inference that we wish to estimate a signal from a set of measurement samples, where a fraction of them violates our standard modeling assumptions. The statistical signal processing literature often refers to these deviating samples as *outliers*, and dealing with them is critical for the successful application of any given learning method in practice.

Consider for example the linear regression model in which measurements are collected as $y_i = \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}_0 + e_i, \quad i = 1, \cdots, n$, for which the sequence $\{\mathbf{d}_i\}_{i=1}^n$ with $\mathbf{d}_i \in \mathbb{R}^p$ is known, $\boldsymbol{\omega}_0 \in \mathbb{R}^p$ is the parameter of interest, and $\{e_i\}_{i=1}^n$ are i.i.d. Gaussian noise variables with $e_i \sim \mathcal{N}(0, \sigma^2)$. In order to find a suitable estimate for $\boldsymbol{\omega}_0$, the standard least squares approach minimizes the sum of squares of the residuals,

$$\hat{\boldsymbol{\omega}}_{\mathrm{LS}} \triangleq \arg\min_{\boldsymbol{\omega}} \sum_{i=1}^{n} \left(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}\right)^2.$$

This coincides with the maximum likelihood (ML) estimator, when the noise is indeed i.i.d. Gaussian. However, even if a very limited number of observations does not follow the assumed Gaussian density, the least squares estimate can be very far away from the true value. The issue of robustness against outliers has been widely studied in the context of linear regression, and it has a long history in robust statistics [1].

Following ideas similar to [2], the author in [3] proposed to explicitly model outliers $\{o_i\}_{i=1}^n$ as sparse additive auxiliary variables in the linear regression model, and then regularize the ML estimator with their ℓ_1-norm as a viable method to detect them. The resulting ℓ_1-regularized ML variant (the terminology LASSO from [2] also applies)

$$\underset{\boldsymbol{\omega}, \{o_i\}_{i=1}^n}{\text{minimize}} \sum_{i=1}^{n} (y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega} - o_i)^2 + \lambda \sum_{i=1}^{n} |o_i|, \tag{1}$$

was proven in [3] to be equivalent to the famous Huber M–estimator $\hat{\boldsymbol{\omega}} \triangleq \arg\min_{\boldsymbol{\omega}} f_H(\boldsymbol{\omega})$, where

$$f_H(\boldsymbol{\omega}) \triangleq \sum_{i=1}^{n} \rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}), \tag{2}$$

and $\rho : \mathbb{R} \to \mathbb{R}$ is the (convex) Huber penalty function defined as

$$\rho(u) \triangleq \begin{cases} u^2, & \text{if } |u| \leq \lambda/2 \\ \lambda|u| - \lambda^2/4, & \text{otherwise.} \end{cases} \tag{3}$$

The parameter $\lambda > 0$ in (1) and (3) is a fixed regularization parameter which controls the outlier rejection in the method.

Several works have considered small errors in addition to outliers in the measurements [4], [5], [7]. Much of the existing literature has focused on assuming nothing else about the outliers $\{o_i\}_{i=1}^n$ other than that they are *sparse* [4], [5], [6], [7]. Assuming sparsity of gross errors, Candes *et al.* in [5] proposed convex optimization algorithms which bound the reconstruction error by a constant times the ideal reconstruction error (under certain restricted isometry conditions), i.e., the error had there been no outliers in the measurements. It is worth noting that [5] requires outlier sparsity, but in turn accommodates outliers which can possibly be *malicious*, i.e., possibly taking values dependent on the particular realization of the sequence $\{\mathbf{d}_i\}_{i=1}^n$.

The focus of this letter is on exploring the performance of the simple Huber estimator when outliers are not malicious, i.e., when they can be assumed deterministic or random with possibly unknown distribution, however independent from the sequence $\{\mathbf{d}_i\}_{i=1}^n$ and the Gaussian noise variables $\{e_i\}_{i=1}^n$ affecting all the measurements. For the case where $\{\mathbf{d}_i\}_{i=1}^n$ is a sequence of standard Gaussian i.i.d. vectors, we provide a simple proof that the minimizer of the Huber penalty function of the residuals converges to the true parameter vector with a $\sqrt{n}$-rate, equal to the decay rate of the oracle Cramér-Rao Lower Bound

Manuscript received April 16, 2014; revised June 05, 2014; accepted June 06, 2014. Date of publication June 10, 2014; date of current version June 18, 2014. This work was supported by the European Research Council under the European Community's Seventh Framework Programme/ERC Grant 228044. The work of N. Sidiropoulos was also supported by the DTC/DTI, University of Minnesota. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jeronimo Arenas-Garcia.

E. Tsakonas and J. Jaldén are with the ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm, Sweden (e-mail: tsakonas@kth.se).

N. D. Sidiropoulos is with the University of Minnesota, Minneapolis, MN 55455 USA.

B. Ottersten is with the ACCESS Linnaeus Centre, Royal Institute of Technology (KTH), Stockholm, Sweden, and also with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2329811

1070-9908 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

(CRLB) [8], [9], even when *outliers are dense* (and possibly unbounded). The constants influencing the rate of convergence are shown to explicitly depend on the outlier contamination level.

When the Gaussian noise variables $\{e_i\}_{i=1}^n$ are absent, recovery in the presence of dense (and possibly malicious) outliers is possible under certain conditions, using the methods proposed in [10], assuming however that the *parameter vector of interest is sufficiently sparse*. Do note that the result in [10] relies on a specific construction model for the sequence $\{\mathbf{d}_i\}_{i=1}^n$, termed as the cross-and-bouquet model.

Notation: Bold lower (upper) case letters stand for vectors (matrices). The cardinality (number of elements) of a set $\mathcal{C}$ is denoted as $|\mathcal{C}|$. We denote the ℓ_0-(pseudo)norm (the operator that counts the number of non-zeros in a vector) as $\|\mathbf{x}\|_0$, the minimum eigenvalue and the trace of a positive definite matrix $\mathbf{X}$ as $\lambda_{\min}(\mathbf{X})$ and $\mathrm{Tr}(\mathbf{X})$, respectively. The complement of an event $\mathcal{G}$ is denoted as $\mathcal{G}^c$. The symbol $\mathbb{N}$ stands for the set of natural numbers. The gradient of a function $f(\cdot)$ is denoted as ∇f. Symbol $Q(\cdot)$ stands for the complementary CDF of the standard Gaussian density and $\Gamma(\cdot)$ stands for the Gamma function.

II. Main Contribution and Context

We consider the linear regression model

$$y_i = \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}_0 + o_i + e_i, \quad i = 1, \cdots, n \tag{4}$$

where $\{\mathbf{d}_i\}_{i=1}^n \in \mathbb{R}^p$ are assumed zero-mean i.i.d. standard Gaussian vectors whose realizations are known, and $\{e_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ Gaussian noise variables.

The outlier variables $\mathbf{o} \triangleq \{o_i\}_{i=1}^n$ are assumed independent of $\{\mathbf{d}_i, e_i\}_{i=1}^n$, either deterministic or random with possibly unknown distribution. A (possibly dense) *linear growth* outlier model is assumed, in which $\|\mathbf{o}\|_0 \leq \lfloor \kappa_o n \rfloor$, where $0 \leq \kappa_o < 1$ is a fixed constant. Bursty impulsive noise in electrical circuits is an example where this linear growth outlier model can be applied. There are, however, many other application examples for the model in (4), see, e.g., [5] for an example in Orthogonal Frequency-Division Multiplexing.

The objective is to derive an upper bound on the reconstruction performance (measured in terms of the Euclidean distance from the true parameter $\boldsymbol{\omega}_0$) of the Huber estimate

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} S_n(\boldsymbol{\omega}) \triangleq \frac{1}{n}\sum_{i=1}^n \rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}), \tag{5}$$

with $\rho(\cdot)$ defined in (3). Note that neither the fraction of outliers, nor their positions are assumed known to the estimator in (5). The main result of the letter is summarized in the next theorem, which we state now and prove in Section III. The theorem precisely establishes that $\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2 = O_p(1/\sqrt{n})$, where O_p is the big-O in probability notation, see [16].

Theorem 1: Consider the data model in (4) and let $\hat{\boldsymbol{\omega}}$ be defined as in (5). For any fixed $\lambda > 0$ and $0 < v < 1$, there exists a constant $n_0 \in \mathbb{N}$ such that

$$\Pr\left(\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2 \leq \frac{K_0}{(1-\kappa_o)\sqrt{n}}\right) \geq 1 - v, \forall n \geq n_0$$
$$\text{with } K_0 \triangleq \frac{\Gamma\left[(p+1)/2\right]}{\Gamma(p/2)} \frac{4\lambda\sqrt{2}}{v\left[1 - 2Q(\frac{\lambda}{4\sigma})\right]}. \tag{6}$$

An immediate implication of the above theorem is that the Huber estimator is a consistent estimator of $\boldsymbol{\omega}_0$ in (4), even when outliers are dense and possibly unbounded, or even possibly introducing a very large bias, as long as they do not depend on $\{\mathbf{d}_i, e_i\}_{i=1}^n$. Observe also that $\hat{\boldsymbol{\omega}}$ approaches $\boldsymbol{\omega}_0$ at a $\sqrt{n}$-rate, which is the same rate that would have been achieved by simple LS had there been no outliers in the observations. The outlier effect becomes obvious upon examining the constant factor before $1/\sqrt{n}$, where one sees that the bound is inversely proportional to the fraction of outlier-free measurements.

It is important to emphasize that the above result does not contradict the notion of the *breakdown point* (BP) which is central in robust estimation [11]. The BP is defined as the maximal fraction of outliers in the observations which can be handled by the estimator (i.e., the maximal fraction of errors above which the estimation error cannot be bounded) [1], [11]. It is well-known that no estimator can succeed if more than 50% of the observations are arbitrarily corrupted (hence the BP has maximum value $\frac{1}{2}$), but this implicitly assumes that the corruptions may possibly be malicious. Such malicious outliers are excluded in our context by virtue of the independence assumption on the variables $\{o_i\}_{i=1}^n$. It should also be noted that the existence of a breakdown point for a given estimator does not imply that this estimator is consistent below its breakdown point, only that the error magnitude can be bounded.

Relation with the oracle CRLB: It is worth pointing out that the rate of convergence coincides with the rate dictated by the CRLB, derived explicitly in [8], in the case where the variables $\{o_i\}_{i=1}^n$ are unknown deterministic. In particular, when $\mathbf{o}$ has maximal support, we know from [8] that the CRLB for a fixed known set of $\{\mathbf{d}_i\}_{i=1}^n$ depends only on the outlier-free measurements (hence the term *oracle* CRLB) and is simply given by

$$\mathbb{E}\left\{\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2^2\right\} \geq \sigma^2 \mathrm{Tr}\left(\sum_{i=1}^n L_{ii}\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}\right)^{-1}, \tag{7}$$

where $L_{ii} = 0$ if $o_i \neq 0$ and $L_{ii} = 1$ otherwise. As long as there is an outlier-free linear fraction of measurements this lower bound on the root mean square error decays at a rate $\sqrt{n}$ as well. In the next section we prove Theorem 1.

III. Proof of Theorem 1

Consider an r-ball centered around $\boldsymbol{\omega}_0$ as $\mathcal{W}(\boldsymbol{\omega}_0) \triangleq \{\boldsymbol{\omega} \in \mathbb{R}^p | \|\boldsymbol{\omega} - \boldsymbol{\omega}_0\|_2 \leq r\}$ with $r > 0$ and define the index sets

$$\begin{aligned} \mathcal{A} &\triangleq \{i | \|\mathbf{d}_i\|_2 \leq \beta\sqrt{n}, |e_i| \leq \alpha, o_i = 0\} \\ \mathcal{B} &\triangleq \{i | |e_i| \leq \alpha, o_i = 0\}, \end{aligned} \tag{8}$$

with β and α some positive constants. Define the probability $q \triangleq \Pr(|e_i| \leq \alpha) = 1 - 2Q(\alpha/\sigma)$. Further, let $K_0' > 0$ be a positive constant that we will suitably choose later on in the proof, and set

$$r \triangleq \frac{4K_0'}{q(1-\kappa_o)\sqrt{n}}, \beta \triangleq \frac{\lambda q(1-\kappa_o)}{16K_0'}, \ \alpha \triangleq \frac{\lambda}{4},$$

where λ is the regularization parameter in the Huber penalty. Note that by picking r, β and α in this manner we have that $i \in \mathcal{A} \Rightarrow |y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}| = |\mathbf{d}_i^{\mathrm{T}}(\boldsymbol{\omega} - \boldsymbol{\omega}_0) + e_i| \leq \|\mathbf{d}_i\|_2 r + \alpha \leq \lambda/2$ for all $\boldsymbol{\omega} \in \mathcal{W}(\boldsymbol{\omega}_0)$. Consider the matrix $\Phi \triangleq \frac{1}{n}\sum_{i\in\mathcal{A}} \mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}$.

The proof consists of two parts: In the first part we show that, assuming that events

$$\begin{aligned}&(\mathrm{i})\mathcal{E}_a : |\mathcal{A}| > (1-\kappa_o)qn/2\\ &(\mathrm{ii})\mathcal{E}_b : \lambda_{\min}(\Phi) > (1-\kappa_o)q/2\\ &(\mathrm{iii})\mathcal{E}_c : \|\nabla S_n(\boldsymbol{\omega}_0)\|_2 \le K_0'/\sqrt{n}, \forall n \ge n_0\end{aligned}$$

happen jointly, the following bound [cf. Theorem 1]

$$\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2 \le \frac{K_0}{(1-\kappa_o)\sqrt{n}}, \forall n \ge n_0 \tag{9}$$

holds, with $K_0 \triangleq 2K_0'/q$. In the second part we complete the proof of the theorem by bounding the probability of the event $\mathcal{E}_a \cap \mathcal{E}_b \cap \mathcal{E}_c$ [i.e., the probability that (i)-(iii) happen jointly].

For the first part, we begin by proving that [assuming (i)-(iii)]

$$S_n(\boldsymbol{\omega}) \ge S_n(\boldsymbol{\omega}_0) + \nabla S_n(\boldsymbol{\omega}_0)^{\mathrm{T}}(\boldsymbol{\omega} - \boldsymbol{\omega}_0) + \frac{m}{2}\|\boldsymbol{\omega} - \boldsymbol{\omega}_0\|_2^2 \tag{10}$$

is satisfied for all $\boldsymbol{\omega} \in \mathcal{W}(\boldsymbol{\omega}_0)$, if one chooses $m = 2\lambda_{\min}(\Phi)$. Let us partition the sum of $S_n(\boldsymbol{\omega})$ into

$$S_n(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i\in\mathcal{A}} \rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}) + \frac{1}{n}\sum_{i\notin\mathcal{A}} \rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}), \tag{11}$$

and define $S_n^{\mathcal{A}}(\boldsymbol{\omega}) \triangleq \frac{1}{n}\sum_{i\in\mathcal{A}}\rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega})$ and $S_n^{\mathcal{A}^c}(\boldsymbol{\omega}) \triangleq \frac{1}{n}\sum_{i\notin\mathcal{A}}\rho(y_i - \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega})$. Since $S_n^{\mathcal{A}^c}(\boldsymbol{\omega})$ is a convex function, the inequality

$$S_n^{\mathcal{A}^c}(\boldsymbol{\omega}) \ge S_n^{\mathcal{A}^c}(\boldsymbol{\omega}_0) + \nabla S_n^{\mathcal{A}^c}(\boldsymbol{\omega}_0)^{\mathrm{T}}(\boldsymbol{\omega} - \boldsymbol{\omega}_0)$$

holds for all $\boldsymbol{\omega} \in \mathcal{W}(\boldsymbol{\omega}_0)$. Thus, it suffices to prove that

$$S_n^{\mathcal{A}}(\boldsymbol{\omega}) \ge S_n^{\mathcal{A}}(\boldsymbol{\omega}_0) + \nabla S_n^{\mathcal{A}}(\boldsymbol{\omega}_0)^{\mathrm{T}}(\boldsymbol{\omega} - \boldsymbol{\omega}_0) + \frac{m}{2}\|\boldsymbol{\omega} - \boldsymbol{\omega}_0\|_2^2.$$

Since λ in the definition of $\mathcal{A}$ is such that samples with index $i \in \mathcal{A}$ fall in the quadratic region of the Huber function, we have that

$$S_n^{\mathcal{A}}(\boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathrm{T}}\Phi\boldsymbol{\omega} - \frac{2}{n}\sum_{i\in\mathcal{A}} y_i\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{n}\sum_{i\in\mathcal{A}} y_i^2. \tag{12}$$

Recalling assumption (ii), observe from (12) that $S_n^{\mathcal{A}}(\boldsymbol{\omega})$ is a positive definite quadratic function and hence, strongly convex [12]. Therefore, (10) holds if one chooses $m = m_o \triangleq 2\lambda_{\min}(\Phi)$.

Now, let $\hat{\boldsymbol{\omega}}$ be the minimizer of $S_n(\boldsymbol{\omega})$ in the neighborhood $\mathcal{W}(\boldsymbol{\omega}_0)$. By definition, $S_n(\hat{\boldsymbol{\omega}}) \le S_n(\boldsymbol{\omega}_0)$ and therefore (10) implies that

$$\nabla S_n(\boldsymbol{\omega}_0)^{\mathrm{T}}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0) + \frac{m_o}{2}\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2^2 \le 0$$

which in turn implies that $\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2 \le 2\|\nabla S_n(\boldsymbol{\omega}_0)\|_2/m_o$. From (ii) and (iii) we therefore have that

$$\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\|_2 \le \frac{K_0}{(1-\kappa_o)\sqrt{n}} < r, \forall n \ge n_0, \tag{13}$$

with $K_0 = 2K_0'/q$. Since $\hat{\boldsymbol{\omega}}$ lies in the interior of $\mathcal{W}(\boldsymbol{\omega}_0)$ and $S_n(\boldsymbol{\omega})$ is convex, it follows that $\hat{\boldsymbol{\omega}}$ is also the global minimizer of $S_n(\boldsymbol{\omega})$, and the first step of the proof is complete. We next turn into the second part, where we bound the probability of the event $\mathcal{E}_a \cap \mathcal{E}_b \cap \mathcal{E}_c$.

Observe that (i) happens with high probability as n increases. In fact, a simple counting argument can show that $\lim_{n\to\infty}|\mathcal{A}|/n = (1-\kappa_o)q$ with probability 1, from which it follows that $\Pr(\mathcal{E}_a) \to 1$ as $n \to \infty$.

To show that (ii) also happens with high probability as n increases, recall the definition of set $\mathcal{B}$ in (8) and notice that

$$\lambda_{\min}(\Phi) = \frac{|\mathcal{B}|}{n}\lambda_{\min}\left[\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}\mathbb{I}\left(\|\mathbf{d}_i\|_2 \le \beta\sqrt{n}\right)\right],$$

where $\mathbb{I}(\cdot)$ is the Boolean indicator function. We have the following equalities: $\lim_{n\to\infty}\lambda_{\min}(\Phi) =$

$$\overset{(a)}{=} \lim_{n\to\infty}\frac{|\mathcal{B}|}{n}\lambda_{\min}\left(\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}\right) \overset{(b)}{=} (1-\kappa_o)q. \tag{14}$$

The equality (a) is due to the tail probability of the Gaussian density, since $\Pr(\|\mathbf{d}_i\|_2 \le \beta\sqrt{n}, i = 1,\cdots,n) \to 1$ as $n \to \infty$. To prove the equality (b), note that the vectors $\{\mathbf{d}_i\}_{i=1}^n$ are standard Gaussian i.i.d independent of $\{e_i\}_{i=1}^n$, therefore

$$\lim_{n\to\infty}\lambda_{\min}\left(\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}\right) = 1$$

with probability 1. Moreover, note that $\lim_{n\to\infty}|\mathcal{B}|/n = (1-\kappa_o)q$ with probability 1. The limit of the product is equal to the product of the individual limits whenever these limits exist [13, Th. 3.4]. Hence, the equality in (b) holds, which implies that $\Pr(\mathcal{E}_b) \to 1$ as $n \to \infty$.

Finally, we prove that (iii) happens with constant probability as n increases, which can however be made arbitrarily close to one by selecting K_0'. From Markov's inequality we get that

$$\Pr\left(\|\nabla S_n(\boldsymbol{\omega}_0)\|_2 > \frac{K_0'}{\sqrt{n}}\right) \le \frac{\mathbb{E}\{\|\nabla S_n(\boldsymbol{\omega}_0)\|_2\}}{K_0'/\sqrt{n}}, \tag{15}$$

and we may further upper bound the expected value of the norm of the gradient in (15) as follows. First, observe that

$$\begin{aligned}\mathbb{E}\{\|\nabla S_n(\boldsymbol{\omega}_0)\|_2\} &= \mathbb{E}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \rho'(o_i + e_i)\mathbf{d}_i\right\|_2\right\} &(16\mathrm{a})\\ &= \mathbb{E}\left\{\mathbb{E}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \rho'(o_i + e_i)\mathbf{d}_i\right\|_2 \middle| o_i, e_i\right\}\right\}, &(16\mathrm{b})\end{aligned}$$

where the outer expectation in (16b) is only with respect to the variables $\{o_i, e_i\}_{i=1}^n$. The variables $\{o_i, e_i\}_{i=1}^n$ and $\{\mathbf{d}_i\}_{i=1}^n$ have been assumed independent, therefore (16b) yields

$$\mathbb{E}\{\|\nabla S_n(\boldsymbol{\omega}_0)\|_2\} = \mathbb{E}\left\{\frac{1}{n}\sqrt{\sum_{i=1}^n \rho'^2(o_i + e_i)}\,\mathbb{E}\left\{\|\mathbf{d}\|_2\right\}\right\}, \tag{17}$$

where the outer expectation is with respect to $\{o_i, e_i\}_{i=1}^n$ and the inner is with respect to the variable $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The variable $\|\mathbf{d}\|_2$ is Chi-distributed [14], therefore $\mathbb{E}\{\|\mathbf{d}\|_2\} = \sqrt{2}\Gamma[(p+1)/2]/\Gamma(p/2)$[14]. Moreover, the function $\rho(\cdot)$ is Lipschitz continuous with $|\rho'(o_i + e_i)| \le \lambda\,\forall i$, therefore (15)-(17) yields the bound

$$\Pr\left(\|\nabla S_n(\boldsymbol{\omega}_0)\|_2 > K_0'/\sqrt{n}\right) \le \frac{\Gamma[(p+1)/2]}{\Gamma(p/2)}\frac{\lambda\sqrt{2}}{K_0'}. \tag{18}$$

For any fixed $v > 0$, selecting K_0' such that the right hand side in (18) becomes equal to $v/2$, yields the constant K_0 in (6).

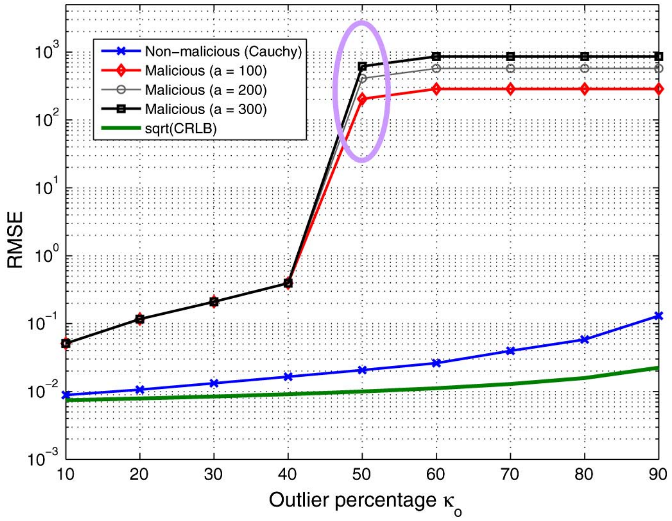

Fig. 1. RMSE performance of Huber estimator in the presence of malicious, non-malicious (Cauchy) outliers, as well as the oracle CRLB as a function of κ_o. Observe the RMSE behavior before and after the breakdown point of the estimator (see the eclipse).

Now, to conclude the proof, let $\mathcal{E}_a^c$, $\mathcal{E}_b^c$ and $\mathcal{E}_c^c$ denote the complementary events to (i), (ii) and (iii), respectively. From the union bound we obtain that

$$\Pr(\mathcal{E}_a \cap \mathcal{E}_b \cap \mathcal{E}_c) \geq 1 - \Pr(\mathcal{E}_a^c) - \Pr(\mathcal{E}_b^c) - \Pr(\mathcal{E}_c^c). \quad (19)$$

The right hand side of (19) can be made greater than or equal to $1 - v$, since $\Pr(\mathcal{E}_a^c) \to 0$ and $\Pr(\mathcal{E}_b^c) \to 0$ as $n \to \infty$ and $\Pr(\mathcal{E}_c^c) < v/2$. The proof of Theorem 1 is now complete.

IV. Numerical Experiments

We construct a simple simulation setup to illustrate how malicious and non-malicious outliers affect the RMSE performance of the Huber estimate. Consider the model (4) with $n = 10^4$, $p = 5$ and the Huber estimator in (5) with $\lambda = 1/2$. The true parameter vector $\boldsymbol{\omega}_0 \in \mathbb{R}^p$ was generated standard Gaussian with particular realization $\boldsymbol{\omega}_0 = [0.322; -0.569; -0.364; -2.735; 0.416]^{\mathrm{T}}$, and kept fixed throughout the experiment. We assess the RMSE performance of (5) using Monte Carlo (MC) simulations. In every trial, the $\{\mathbf{d}_i\}$'s are generated from the standard Gaussian distribution and $\{e_i\}$'s are independent $\mathcal{N}(0, 0.1)$. Non-malicious outliers are assumed standard independent Cauchy random variables (zero location, unit scale) [14], and they are added to the last $\lfloor \kappa_o n \rfloor$ measurements. The non-malicious outlier samples are further *biased*, by adding the constant value $\mu = 10^4$. On the other hand, malicious outliers are generated by adding to the last $\lfloor \kappa_o n \rfloor$ samples the values $o_i = a\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\omega}_0$ for $i = \lfloor \kappa_o n \rfloor, \cdots, n$, with a chosen either as 100, 200 or 300. Notice how outliers in the second case are specifically designed to adversarially affect the estimation of $\boldsymbol{\omega}_0$.

Fig. 1 depicts the RMSE of (5) in every case (malicious/non-malicious), as well as the root-CRLB, as a function of the outlier percentage κ_o. All curves are produced by averaging over 100 independent MC runs, and the interior-point method developed in [15] was used to solve the optimization problem in (5). Observe that the RMSE performance of (5) in the non-malicious outlier-case is comparable to that in the malicious case(s) when outliers are below 50%, but the latter RMSE degrades significantly after 50% (which is precisely the breakdown point of the estimator [1], [11]). On the other hand, notice that the Huber estimate is similar to the theoretical CRLB in (7) when outliers are not malicious, even for large outlier percentages.

V. Discussion and Concluding Remarks

This letter examined the performance of the Huber estimator in a scenario where the collected linear measurements are corrupted by an arbitrary linear fraction of gross errors, and when in addition, all measurements are contaminated by standard errors. When the measurement matrix of the model comprises i.i.d. Gaussian entries and gross errors are not malicious (i.e., when gross errors are independent of the measurement model matrix), it is shown that the Huber estimate converges to the sought parameter vector with the same rate as if there had been no gross corruptions. The result holds even if these corruptions are dense.

One can observe from the proof that the core properties that make this convergence behavior possible are the convexity and the Lipschitz continuity of the Huber penalty function. On the flip side, we believe that the Gaussianity assumption on the measurement matrix of the model is not critical to the analysis, and can possibly be replaced by a milder assumption. We leave this latter conjecture as a topic for future work.

References

[1] P. Huber, *Robust statistics*. New York, NY, USA: Wiley, 1981.

[2] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. B*, vol. 58, pp. 267–288, 1994.

[3] J. Fuchs, "An inverse problem approach to robust regression," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA, Mar. 15–19, 1999.

[4] G. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[5] E. Candes and P. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, 2008.

[6] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, pp. 4203–4215, 2005.

[7] Y. Jin and B. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Mar. 2010.

[8] E. Tsakonas, J. Jaldén, N. Sidiropoulos, and B. Ottersten, "Connections between sparse estimation and robust statistical learning," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.

[9] Z. Ben-Haim and Y. Eldar, "The Cramér-Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58, pp. 3384–3389, 2010.

[10] J. Wright and Y. Ma, "Dense error correction via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, pp. 3540–3560, 2009.

[11] A. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," in *IEEE Signal Process. Mag.*, Jul. 2012, pp. 61–80.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[13] W. Rudin, "Principles of mathematical analysis," in *International Series in Pure and Applied Mathematics*. New York, NY, USA: McGraw-Hill, 1976.

[14] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford, U.K.: Oxford Univ. Press, 2001.

[15] J. Jaldén, "Bi-criterion ℓ_1/ℓ_2-norm optimization," M.S. thesis, Dept. of Signals, Sensors and Systems, Signal Processing, Royal Inst. Technol. (KTH), Stockholm, Sweden, Sep. 2002.

[16] Y. Dodge, *The Oxford Dictionary of Statistical Terms*. Oxford, U.K.: Oxford Univ. Press, 2003.