Adaptive Hard Thresholding for Near-optimal Consistent Robust Regression

Arun Sai Suggala[*] ASUGGALA@CS.CMU.EDU
Carnegie Mellon University

Kush Bhatia KUSHBHATIA@BERKELEY.EDU
University of California, Berkeley

Pradeep Ravikumar PRADEEPR@CS.CMU.EDU
Carnegie Mellon University

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research, India

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the problem of robust linear regression with response variable corruptions. We consider the oblivious adversary model, where the adversary corrupts a fraction of the responses in complete ignorance of the data. We provide a nearly linear time estimator which consistently estimates the true regression vector, even with $1 - o(1)$ fraction of corruptions. Existing results in this setting either don't guarantee consistent estimates or can only handle a small fraction of corruptions. We also extend our estimator to robust sparse linear regression and show that similar guarantees hold in this setting. Finally, we apply our estimator to the problem of linear regression with heavy-tailed noise and show that our estimator consistently estimates the regression vector even when the noise has unbounded variance (e.g., Cauchy distribution), for which most existing results don't even apply. Our estimator is based on a novel variant of outlier removal via hard thresholding in which the threshold is chosen adaptively and crucially relies on randomness to escape bad fixed points of the non-convex hard thresholding operation.

Keywords: Robust regression, heavy tails, hard thresholding, outlier removal.

1. Introduction

We study robust least squares regression, where the goal is to robustly estimate a linear predictor from data which is potentially corrupted by an adversary. We focus on the setting where response variables are corrupted via an oblivious adversary. Such a setting has numerous applications such as click-fraud in a typical ads system, ratings-fraud in recommendation systems, as well as the less obvious application of regression with heavy tailed noise.

For the problem of oblivious adversarial corruptions, our goal is to design an estimator that satisfies three key criteria: (a) (**statistical efficiency**) estimates the optimal solution *consistently* with nearly optimal statistical rates, (b) (**robustness efficiency**) allows a high amount of corruption, *i.e.,* fraction of corruptions is $1 - o(1)$, (c) (**computational efficiency**) has the same or nearly the same computational complexity as the standard ordinary least squares (OLS) estimator. Most existing

[*] Part of the work done while interning at Microsoft Research, India.

. Extended abstract. Full version appears as [https://arxiv.org/abs/1903.08192, v2]

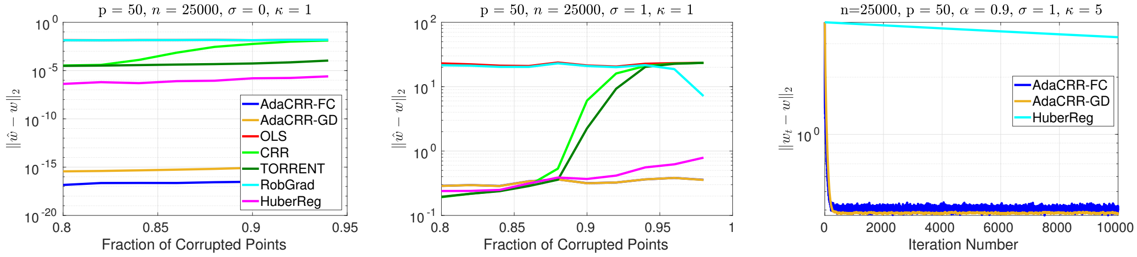

Figure 1: The first two plots show the parameter error (y-axis) of various estimators as we vary fraction of corruptions α in the robust regression setting (x-axis); noise variance is 0 for the first plot and 1 for the second. Plots indicate that AdaCRR is able to tolerate significantly higher fraction of outliers than most existing methods. The last plot shows parameter error over number of iterations for robust regression, indicating AdaCRR can be upto 100x faster as compared to Huber regression.

techniques do not even provide consistent estimates in this adversary model (Bhatia et al., 2015; Nasrabadi et al., 2011; Nguyen and Tran, 2013; Prasad et al., 2018; Diakonikolas et al., 2018; Wright and Ma, 2010). Bhatia et al. (2017) provides statistically consistent and computationally efficient estimator, but requires the fraction of corruptions to be less than a small constant ($\leqslant 1/100$). Tsakonas et al. (2014) study Huber-loss based regression to provide nearly optimal statistical rate with nearly optimal fraction of corruptions. But their sample complexity is sub-optimal, and more critically, the algorithm has super-linear computational complexity (in terms of number of points) and is significantly slower than the standard least squares estimator.

So the following is still an open question: *"Can we design a linear time consistent estimator for robust regression that allows almost all responses to be corrupted by an oblivious adversary?"*

We answer this question in affirmative, *i.e.,* we design a novel outlier removal technique that can ensure consistent estimation at nearly optimal statistical rates, assuming Gaussian data and sub-Gaussian noise. Our results hold as long as the number of points n is larger than the input dimensionality p by logarithmic factors, i.e., $n \geqslant p \log^2 p$, and allows $n - \frac{n}{\log \log n}$ responses to be corrupted; the number of corrupted responses can be increased to $n - \frac{n}{\log n}$ with a slightly worse generalization error rate.

Our algorithm, which we refer to as AdaCRR [1], uses a similar technique as Bhatia et al. (2015, 2017), where we threshold out points that we estimate as outliers in each iteration. However, we show that fixed thresholding operators as in Bhatia et al. (2015, 2017) can get stuck at poor fixed-points in presence of a large number of outliers. Instead, we rely on an adaptive thresholding operator that uses noise in each iteration to avoid such sub-optimal fixed-points. Similar to Bhatia et al. (2015, 2017), AdaCRR-FC solves a standard OLS problem in each iteration, so the overall complexity is $O(T \cdot T_{OLS})$ where T is the number of iterations and T_{OLS} is the time-complexity of an OLS solver. We show that $T = O(\log 1/\epsilon)$ iterations are enough to obtain ϵ-optimal solution, i.e., the algorithm is almost as efficient as the standard OLS solvers. Our simulations also demonstrate our claim, *i.e.,* we observe that AdaCRR-FC is significantly more efficient than Huber-loss based approaches (Tsakonas et al., 2014) while still ensuring consistency in presence of a large number of corruptions unlike existing thresholding techniques (Bhatia et al., 2015, 2017) (see Figure 1).

[1] To be more precise, AdaCRR is a framework and we study two algorithms instantiated from this framework, namely AdaCRR-FC, AdaCRR-GD which differ in how they update $\mathbf{w}$.

The above result requires $n \geqslant p \log^2 p$ which is prohibitively large for high-dimensional problems. Instead, we study the problem with sparsity structure on the regression vector (Wainwright, 2009). That is, we study the problem of sparse linear regression with oblivious response corruptions. We provide *first* (to the best of our knowledge) consistent estimator for the problem under standard RSC assumptions. Similar to the low-d case, we allow $1 - o(1)$ fraction of points to be corrupted, but the sample complexity requirement is only $n \geqslant k^* \log^2 p$, where k^* is the number of non-zero entries in the optimal sparse regression vector. Existing Huber-loss based estimators (Tsakonas et al., 2014) would be difficult to extend to this setting due to the additional non-smooth L_1 regularization of the regression vector. Existing hard-thresholding based consistent estimators (Bhatia et al., 2017) marginalize out the regression vector, which is possible only in low-d due to the closed form representation of the least squares solution, and hence, do not trivially extend to sparse regression.

Finally, we enhance and apply our technique to the problem of regression with heavy-tailed noise. By treating the tail as oblivious adversarial corruptions, we obtain consistent estimators for a large class of heavy-tailed noise distributions that might not even have well-defined first or second moments. Despite being a well-studied problem, to the best of our knowledge, this is the first such result in this domain of learning with heavy tailed noise. For example, our results provide consistent estimators with Cauchy noise, for which even the mean is not well defined, with rates which are very similar to that of standard sub-Gaussian distributions. In contrast, most existing results (Sun et al., 2018; Hsu and Sabato, 2016) do not even hold for Cauchy noise as they require the variance of the noise to be bounded. Furthermore, existing results mostly rely on *median of means* technique (Hsu and Sabato, 2016; Lecué and Lerasle, 2017; Prasad et al., 2018), while we present a novel but natural viewpoint of modeling the tail of noise as adversarial but oblivious corruptions.

2. Problem Setup and Main Results

We are given n independent data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim D$ sampled from a Gaussian distribution $D = \mathcal{N}(0, \Sigma)$ and their corrupted responses $y_1, \ldots, y_n$, where,

$$y_i = \mathbf{x}_i^T \mathbf{w}^* + \epsilon_i + b_i^*, \tag{1}$$

$\mathbf{w}^*$ is the true regression vector, ϵ_i - the white noise - is independent of $\mathbf{x}_i$ and is sampled from a sub-Gaussian distribution with parameter σ, and b_i^* is the corruption in the response of $\mathbf{x}_i$. $\{b_i^*\}_{i=1}^n$ is a sparse corruption set, i.e., $\|b^*\|_0 = |\{i,\ s.t.,\ b_i^* \neq 0\}| \leqslant \alpha \cdot n$ where $\alpha < 1$. Also, $\{b_i^*\}_{i=1}^n$ is *independent* of $\{\mathbf{x}_i, \epsilon_i\}_{i=1}^n$. Apart from this independence we do not impose any restrictions on the values of corruptions added by the adversary. Our goal is to robustly estimate $\mathbf{w}^*$ from the corrupted data $\{\mathbf{x}_i, y_i\}_{i=1}^n$. In particular, following are the key criteria in evaluating an estimator's performance:

- **Breakdown point:** It is the maximum fraction of corruption, α, above which the estimator is not guaranteed to recover $\mathbf{w}^*$ with small error, even as $n \to \infty$ (Hampel, 1971).
- **Statistical rates and sample complexity:** We are interested in the generalization error $(\mathbb{E}_{x\sim D}[(\langle \mathbf{x}, \mathbf{w}\rangle - \langle \mathbf{x}, \mathbf{w}^*\rangle)^2])$ of the estimator and its scaling with problem dependent quantities like n, p, noise variance σ^2 as well as the fraction of corruption α.
- **Computational complexity:** The number of computational steps taken to compute the estimator. The goal is to obtain nearly linear time estimators similar to the standard OLS solvers.

As discussed later in the section, our AdaCRR estimator is near optimal with respect to all three criteria above.

Heavy-tailed Regression. We also study the heavy-tailed regression problem where $y_i = \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i$ for all $\mathbf{x}_i \sim D$ and $i \in [n]$. Noise $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{E}$ where $\mathcal{E}$ is a heavy-tailed distribution, such as the Cauchy distribution which does not even have bounded first moment. The goal is to design an efficient estimator that provides nearly optimal statistical rates.

Notation. Let $X = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n]^T$ be the matrix whose i^{th} row is equal to $\mathbf{x}_i \in \mathbb{R}^p$. Let $\mathbf{y} = [y_1, y_2 \dots y_n]^T, \boldsymbol{\epsilon} = [\epsilon_1, \dots \epsilon_n]^T$, and $\mathbf{b}^* = [b_1^*, \dots b_n^*]^T$. $\|\mathbf{a}\|_\Sigma^2 := \mathbf{a}^T\Sigma\mathbf{a}$ for a positive definite matrix Σ. $\|\mathbf{a}\|_0$ denotes the L_0 norm of $\mathbf{a}$, i.e., the number of non-zero elements in $\mathbf{a}$. $b = \widetilde{O}(a)$ implies, $b \leqslant Ca\log a$ for a large enough constant $C > 0$ independent of a. We use $SG(\sigma^2)$ to denote the set of random variables whose Moment Generating Function (MGF) is less than the MGF of $\mathcal{N}(0, \sigma^2)$.

2.1. Main Results

Robust Regression: For robust regression with oblivious response variable corruptions, we propose the first efficient consistent estimator with break-down point of 1. That is,

Theorem 1 (Robust Regression) *Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be n observations generated from the oblivious adversary model, i.e., $\mathbf{y} = X\mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{b}^*$ where $\epsilon_i \in SG(\sigma^2)$, $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$, $\|\mathbf{b}^*\|_0 \leqslant \alpha \cdot n$ and $\mathbf{b}^*$ is selected independently of $X, \boldsymbol{\epsilon}$. Suppose AdaCRR-FC is run for T iterations with approparite choice of hyperparameters. Then with probability at least $1 - T/n^6$, the T-th iterate $\mathbf{w}_T$ produced by the AdaCRR-FC algorithm satisfies:*

$$\|\mathbf{w}_T - \mathbf{w}^*\|_\Sigma \leqslant \widetilde{O}\left(\frac{\sigma}{1-\alpha}\sqrt{\frac{p\log^2 n + (\log n)^3}{n}}\right),$$

for any $\alpha \leqslant 1 - \frac{\Theta(1)}{\log\log n}$, where the number of iterations $T = \widetilde{O}\left(\log\left(\frac{n}{p} \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_\Sigma}{\sigma}\right)\right)$.

Remarks: a) AdaCRR-FC solves an OLS problem in each iteration and the number of iterations is $\approx \log n$, so the overall time complexity of the algorithm is still nearly linear in n. In contrast, standard Huber-loss or L_1 loss based methods (Tsakonas et al., 2014; Nasrabadi et al., 2011) have iteration complexity of $1/\sqrt{\epsilon}$ for ϵ-suboptimality and require $\epsilon \approx 1/\sqrt{n}$, which implies super-linear $O(n^{1.25})$ time complexity.

b) Break-down point α of AdaCRR-FC satisfies: $\alpha \to 1$ for $n \to \infty$. In contrast, similar consistent estimator by Bhatia et al. (2017) requires $\alpha < 1/100$. In fact, we show that fixed hard thresholding operators like the ones used by (Bhatia et al., 2015, 2017) *cannot* provide consistent estimator for $\alpha \to 1$; instead, we propose and analyze a randomized and adaptive thresholding operator to avoid sub-optimal fixed-points.

c) Generalization error of AdaCRR-FC is $O(\sigma^2 \cdot p\log^2 n/n)$, which is information theoretically optimal up to $\log^2 n$ factors. In contrast, most of the existing analysis for L_1-loss do not guarantee such consistent estimators (Nasrabadi et al., 2011; Wright and Ma, 2010; Nguyen and Tran, 2013).

e) Sample complexity of AdaCRR-FC is nearly optimal $n = O(p\log^2 p)$ and can be improved to $n = O(k^*\log^2 p)$ for k^*-sparse estimators with the data that satisfies Restricted Strong Convexity

Paper	Breadkdown Point	Consistent	Optimal Sample Comlexity	Computational Rates
Wright and Ma (2010)	$\alpha \to 1$	No	Yes	$O(1/\sqrt{\epsilon})$
Nasrabadi et al. (2011)	$\alpha \to 1$	No	Yes	$O(1/\sqrt{\epsilon})$
Tsakonas et al. (2014)	$\alpha \to 1$	Yes	No	$O(1/\sqrt{\epsilon})$
Bhatia et al. (2017)	$\alpha = \Theta(1)$	Yes	Yes	$O(\log(1/\epsilon))$
This paper	$\alpha \to 1$	Yes	Yes	$O(\log(1/\epsilon))$

Table 1: Comparison of various approaches for regression under oblivious adversary model. The computational rates represents the time taken by estimator to compute an ϵ-approximate solution.

and Restricted Strong Smoothness.

See Table 1 for a detailed comparison with the existing works.

Regression with Heavy-tailed Noise: We now present our result for regression with heavy-tailed noise.

Theorem 2 (Heavy-tailed Regression) *Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be n observations generated from the linear model, i.e., $y_i = \mathbf{x}_i^T \mathbf{w}^* + \epsilon_i$ where $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$, ϵ_i's are sampled i.i.d. from a distribution s.t. $\mathbb{E}[|\epsilon|^\delta] \leqslant C$ for a constant $\delta > 0$ and are independent of $\mathbf{x}_i$. Then, for $T = \widetilde{O}\left(\log\left(\frac{n}{p} \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_\Sigma}{\sigma}\right)\right)$, the $\mathbf{w}_T$-th iterate of AdaCRR-FC guarantees the following with probability $\geqslant 1 - T/n^6$:*

$$\|\mathbf{w}_T - \mathbf{w}^*\|_\Sigma \leqslant O\left(C^{1/\delta}\sqrt{\frac{p\log n + \log^2 n}{n}}\right).$$

Remarks: a) Note that our technique does not even require the first moment to exist. In contrast, existing results hold only when the variance is bounded (Hsu and Sabato, 2016). In fact, the general requirement on distribution of ϵ is significantly weaker and holds for almost every distribution whose parameters are independent of n. Also, we present a similar result for mean estimation with symmetric noise ϵ.

b) For Cauchy noise (Johnson et al., 2005) with location parameter 0, and scale parameter σ, we can guarantee error rate of $\approx \sigma\sqrt{\frac{p\log^2 n}{n}}$, i.e., we can obtain almost same rate as sub-Gaussian noise despite unbounded variance which precludes most of the existing results.

c) Similar to robust regression, the estimator is nearly linear in n, p. Moreover, we can extend our analysis to sparse linear regression with heavy-tailed response noise.

Acknowledgements

A.S., P.R. acknowledge the support of NSF via IIS-1661755, IIS-1664720, and Rakuten Inc.

References

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate Discrete Distributions*, volume 444. John Wiley & Sons, 2005.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

Nasser M Nasrabadi, Trac D Tran, and Nam Nguyen. Robust lasso with missing and grossly corrupted observations. In *Advances in Neural Information Processing Systems*, pages 1881–1889, 2011.

Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, (just-accepted):1–35, 2018.

Efthymios Tsakonas, Joakim Jaldén, Nicholas D Sidiropoulos, and Björn Ottersten. Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, 2014.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

John Wright and Yi Ma. Dense error correction via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.