# Online Robust Regression via SGD on the $\ell_1$ loss

*Scott Pesme (EPFL), Nicolas Flammarion (EPFL)*

EPFL — NEURAL INFORMATION PROCESSING SYSTEMS

## Problem setup and corruption model

Assume we are given a stream of i.i.d. datapoints $(x_i, y_i)_{i\geq 0}$ where the responses $(y_i)_{i\geq 0}$ have potentially been corrupted by an oblivious adversary:

Adversarial "sparse" noise $\mathbb{P}(b \neq 0) = \eta \in [0,1)$

$$y = \langle x, \theta^* \rangle + \underbrace{\varepsilon}_{\text{"Nice" noise}} + \overbrace{b}$$

In the paper we consider the following question:

***Can we efficiently recover the gold parameter $\theta^*$ ?***

Current methods rely on handling the entire dataset at once and are therefore inefficient in large-scale settings. We propose a different approach.

## Our approach

The $\ell_1$ loss is known for its robustness properties. Hence a natural approach is to consider the least absolute deviation (LAD) problem:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \mathbb{E}_{(x,y)}[\,|\,y - \langle x, \theta \rangle\,|\,]$$

Assuming $\mathbb{E}[x] = 0$ and $b \perp\!\!\!\perp (x, \varepsilon)$ then $\theta^* \in \text{argmin}_\theta\, f(\theta)$. Minimising the LAD problem therefore makes sense.

To solve this problem we propose to use the very simple and highly scalable stochastic gradient descent (SGD) algorithm:

$$\theta_n = \theta_{n-1} + \gamma_n \,\text{sgn}\big(y_n - \langle x_n, \theta_{n-1} \rangle\big) x_n$$

And we consider the averaged iterate $\bar{\theta}_n = n^{-1}\Sigma_{i=0}^{n-1}\theta_i$.

## Underlying challenges

Several technical manipulations are required in order to obtain the optimal rates:

- we cannot expect $f$ to be strongly convex over $\mathbb{R}^d$, hence simply applying the known SGD results leads to a suboptimal $O(n^{-1/2})$ rate
- the $\ell_1$ loss isn't smooth, therefore it isn't transparent that Polyak-Ruppert averaging will lead to a fast $O(1/n)$ rate
- ideally we want to obtain dominant convergence rate terms which are independent of the conditioning of the feature covariance matrix

## Main result

**Assumptions**:

- $x \sim \mathcal{N}(0, H)$ where $H$ is a $d \times d$ positive definite matrix
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and is independent of $x$
- the adversarial noise $b$ is independent of $(x, \varepsilon)$ and $\mathbb{P}(b \neq 0) = \eta \in [0,1)$

**Notations**: • $\mu = \lambda_{min}(H)$ • $\tilde{\eta} = \eta \cdot \underbrace{\left(1 - \mathbb{E}_b\left[\exp\left(-\frac{b^2}{2\sigma^2}\right) \mid b \neq 0\right]\right)}_{\textit{effective outlier proportion}} \in [0, \eta)$

• $R^2 = \text{trace}(H)$

**Theorem:**

Consider the SGD iterates on the $\ell_1$ loss. Assume $\gamma_n = \gamma_0\, n^{-1/2}$. Then for all $n \geq 1$ :

$$\underbrace{\mathbb{E}\left[\|\bar{\theta}_n - \theta^*\|_H^2\right]}_{\textit{prediction error}} = \underbrace{O\left(\frac{\sigma^2 d}{(1-\tilde{\eta})^2 n}\right)}_{\textit{optimal variance term}} + \underbrace{\tilde{O}\left(\frac{\|\theta_0 - \theta^*\|^4}{\gamma_0^2(1-\tilde{\eta})^2 n}\right)}_{\textit{bias term}} + \underbrace{\tilde{O}\left(\frac{\gamma_0^2 R^4}{(1-\tilde{\eta})^2 n}\right)}_{\textit{analysis by-product ?}} + \underbrace{\tilde{O}\left(\frac{1}{\mu^2 n^{3/2}}\right)}_{\textit{higher order herms}}$$

## Result analysis

**We highlight that :**

- the result is given in terms of the classical prediction error
- the overall $O(n^{-1})$ rate is unimprovable
- the variance term is statistically optimal with regards to $\sigma$, $d$ and $n$
- the bound depends on the effective outlier proportion $\tilde{\eta}$
- in the finite horizon framework with $N$ samples, the breakdown point is state of the art: $\tilde{\eta} = 1 - \tilde{\Omega}(N^{-1/2})$
- the dominant terms are independent of the condition constant $\mu$
- the algorithm is (nearly) parameter free

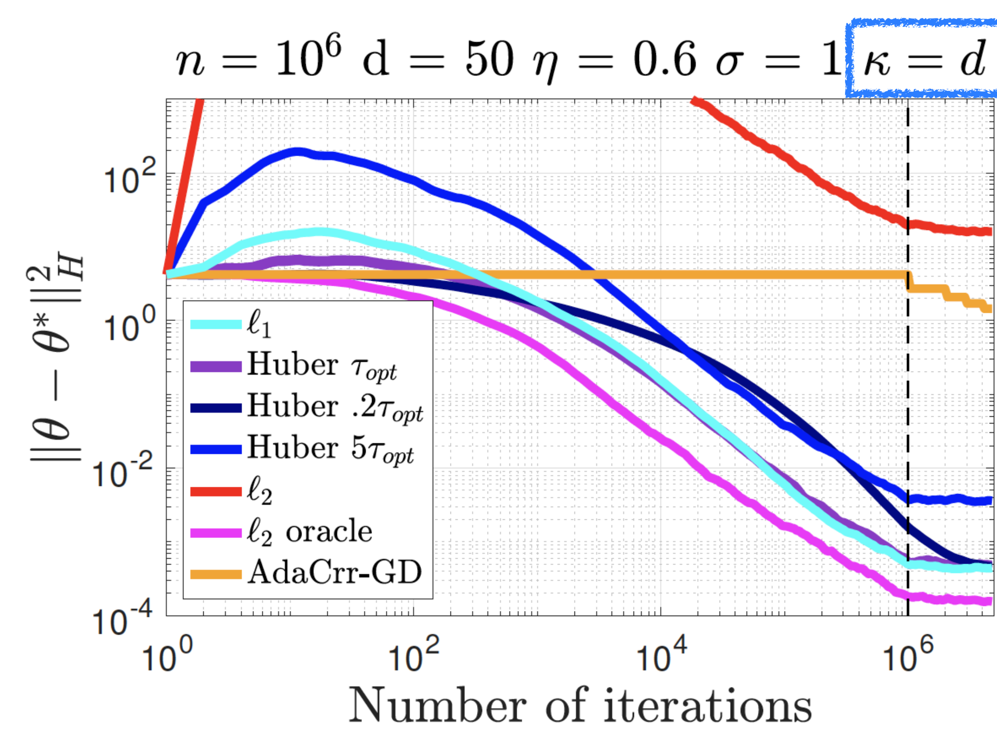
**Discussion and future work :**

- the Gaussian assumptions on $(x, \varepsilon)$ are quite strong, we believe they could be relaxed
- the optimal dependency on $\eta$ is still an open interesting question

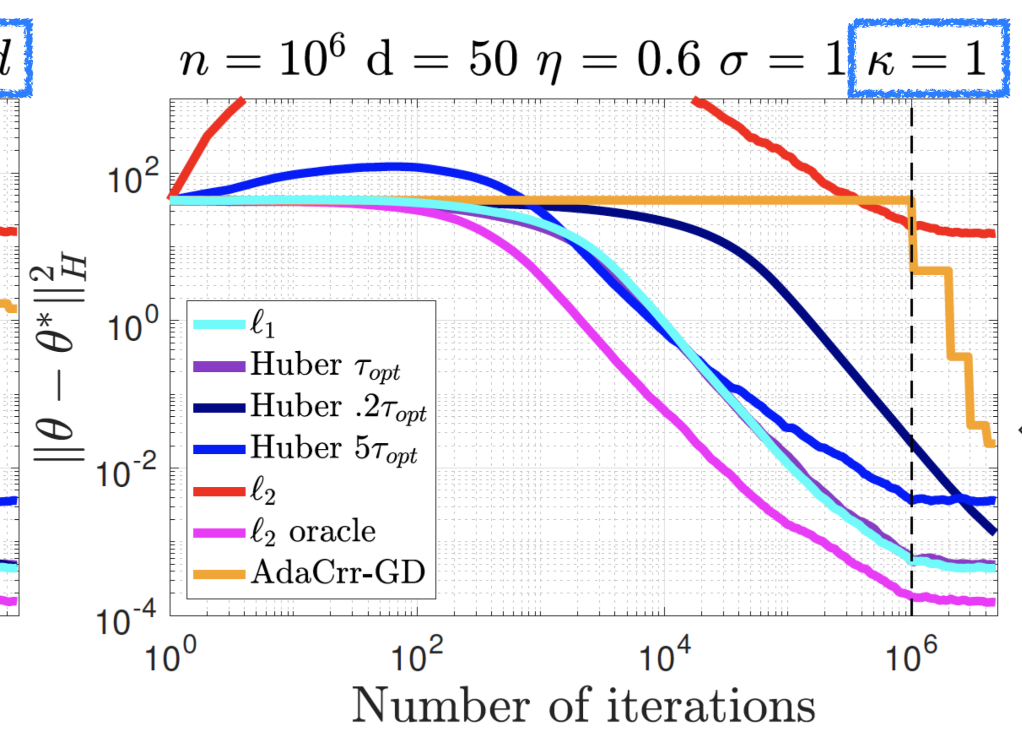
## Experiments

**Experimental setup:**

- i.i.d. inputs $x_i \sim \mathcal{N}(0, H)$ where $H$ is either identity or positive semi definite with eigenvalues $(1/k)_{1\leq k \leq d}$
- the outputs are generated using i.i.d. noises $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $b_i$ following a toy contamination model (see paper)
- we compare averaged SGD on the $\ell_1$, $\ell_2$ , Huber losses and to the state of the art AdaCRR-GD algorithm from (Suggala et al. 2019)




*Convergence rate with bad conditioning*




*Convergence rate with good conditioning*

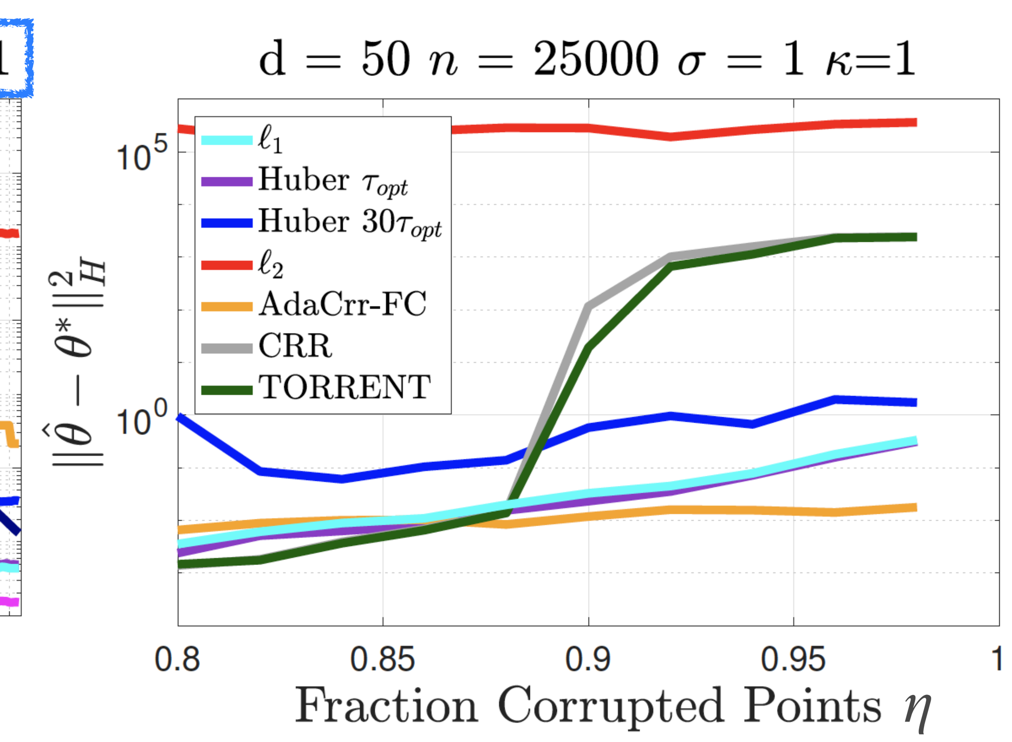



*Illustration of the breakdown points*

**Notice that :**

- averaged SGD on the $\ell_1$ loss exhibits a clear $O(n^{-1})$ convergence rate
- AdaCRR-GD is very sensitive to the conditioning of the covariance matrix $H$, this is not the case for our algorithm
- averaged SGD on the Huber loss does not lead to better performances and requires an extra parameter to tune