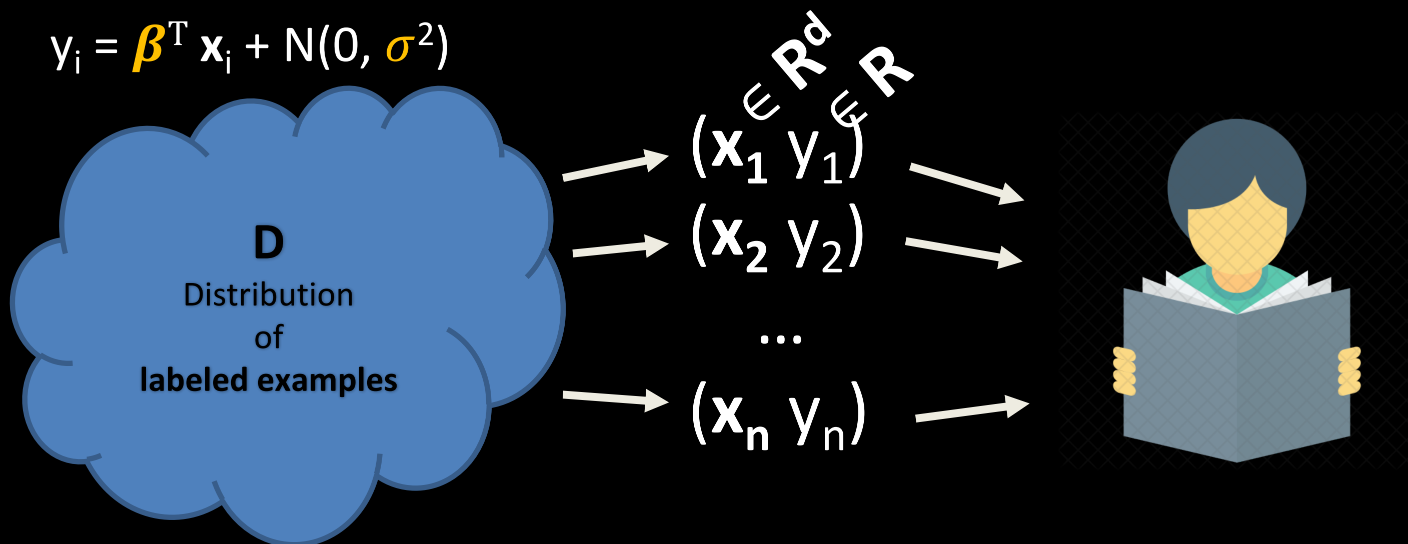
Efficient Algorithms and Lower Bounds for Robust Linear Regression

Ilias Diakonikolas(USC), **Weihao Kong**(Stanford) and Alistair Stewart(USC)

Linear regression

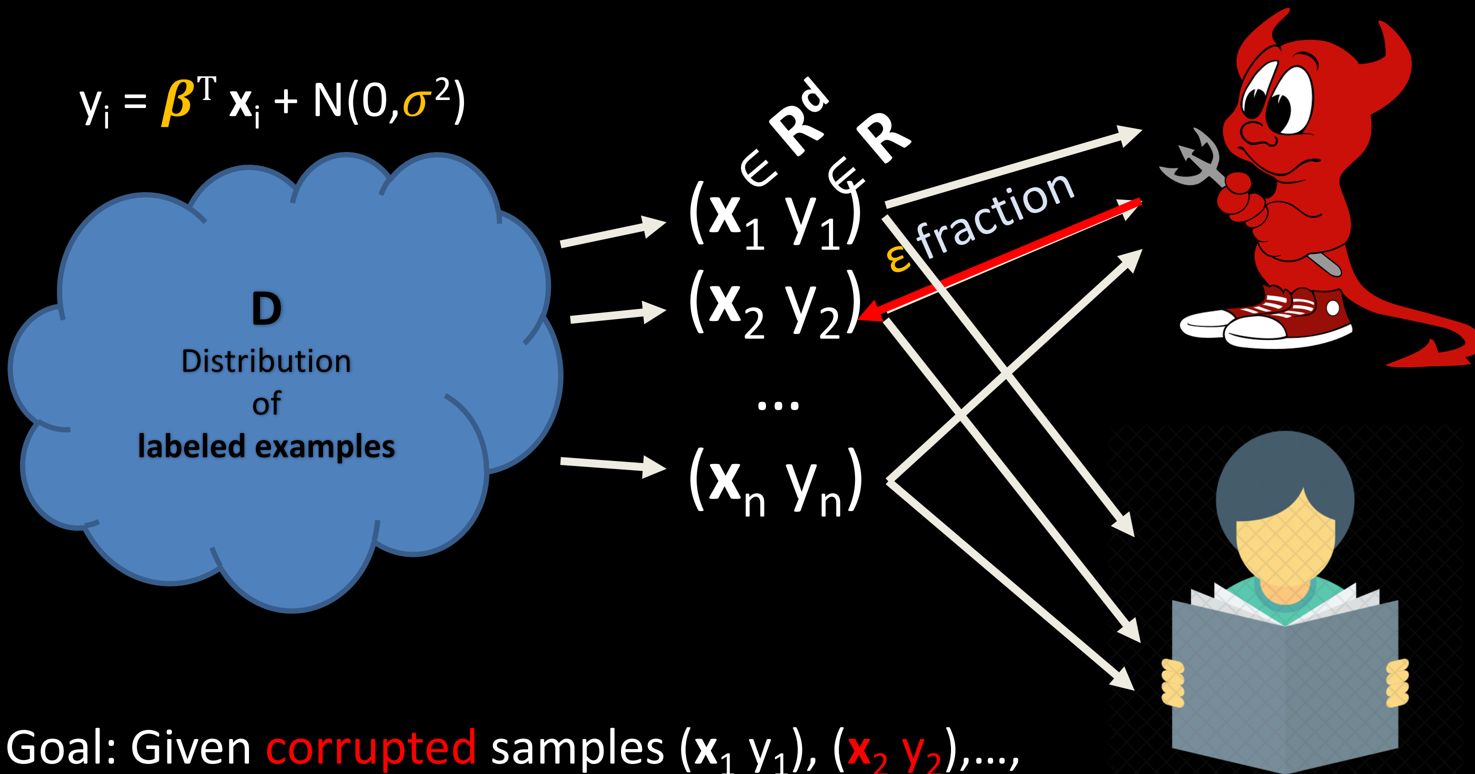
$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + N(0, \sigma^2)$



Goal: Given n iid samples$(\mathbf{x}_1\ y_1), (\mathbf{x}_2\ y_2), \ldots, (\mathbf{x}_n\ y_n)$, estimate $\boldsymbol{\beta}$

Linear regression with Corruption

$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + N(0,\sigma^2)$



Goal: Given corrupted samples $(\mathbf{x}_1\ y_1)$, $(\mathbf{x}_2\ y_2)$,…, $(\mathbf{x}_n\ y_n)$, estimate $\boldsymbol{\beta}$

Related Work and Our Contribution

1. No corruption: $\mathbf{x} \sim N(0, \Sigma)$, $\mathrm{y} = \boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{x} + \eta$, $\eta \sim N(0, \sigma^2)$, Σ unknown.

   Easy fact: For any accuracy parameter ϵ>0, Ordinary Least Square estimator achieves $\|(\beta - \hat{\beta})\|_{\Sigma} \leq \sigma\epsilon$ with $\Omega(d/\epsilon^2)$ samples.

2. Response variable y corrupted: $\mathbf{x} \sim N(0, \Sigma)$, $\mathrm{y} = \boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{x} + \eta$, $\eta \sim N(0, \sigma^2)$, ϵ fraction of corruption, Σ unknown. [Bhatia Jain Kar 15] [Bhatia Jain Kamalaruban Kar 17]

Related Work and Our Contribution

3. Corruption on **x** and y: $\mathbf{x} \sim N(0, \Sigma), y = \boldsymbol{\beta}^T \mathbf{x} + \eta, \eta \sim N(0, \sigma^2)$, ϵ fraction of corruption, Σ unknown.

[Gao 17] *even with infinite sample, can not estimate better than $\sigma\epsilon$.*

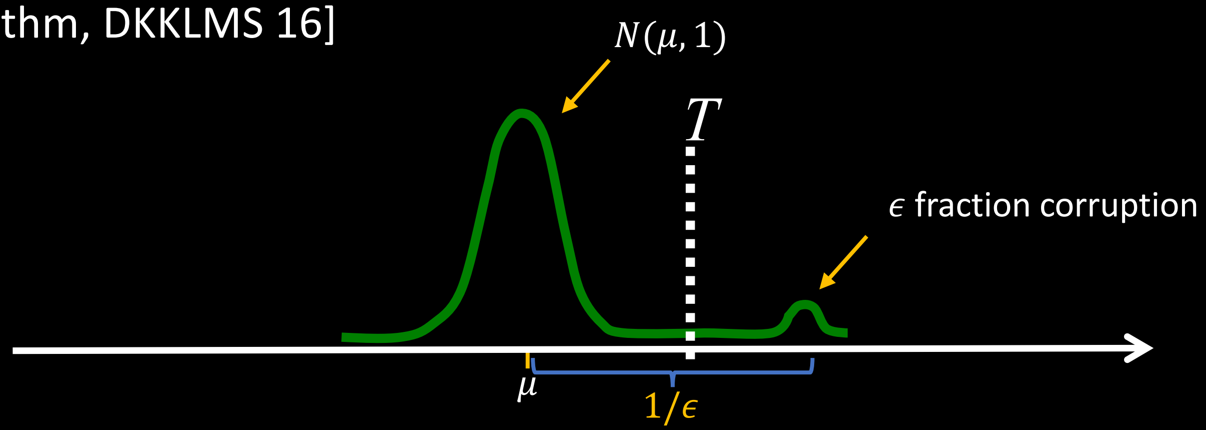
Result	Error $\|(\beta - \hat{\beta})\|_\Sigma$	Sample Complexity
[Prasad-Suggala-Balakrishnan-Ravikumar 18]	$\sigma\sqrt{\epsilon \log(d)}$	$\tilde{O}(d^2/\epsilon^{4/3})$
[Diakonikolas-Kamath-Kane-Li-Steinhardt-Stewart 18]	$\sigma\sqrt{\epsilon}$	$\tilde{O}(d^5/\epsilon^2)$
[Klivans-Kothari-Meka 18]	$\sigma\sqrt{\epsilon}$	Poly$(d, 1/\epsilon)$
Our algorithm	$\sigma\epsilon \log(1/\epsilon)$	$\tilde{O}(d^2/\epsilon^2)$
Info-theory LB [Gao 17]	$\sigma\epsilon$	$\Omega(d/\epsilon^2)$
Our Statistical Query LB	$\sigma\sqrt{\epsilon}$	$\Omega_\epsilon(d^2)$

4. Same setting except Σ is known.

[Balakrishnan-Du-Li-Singh 17]	$\sigma\sqrt{1 + \|\beta\|^2}\epsilon \log(1/\epsilon)^2$	$\tilde{O}(d^2/\epsilon^2)$
Our algorithm	$\sigma\epsilon\log(1/\epsilon)$	$\tilde{O}(d/\epsilon^2)$

Preliminary

Toy example: how to detect corruption in mean estimation?
[Filter Algorithm, DKKLMS 16]



Observation: to change the mean by a constant, the corrupted samples must be put $1/\epsilon$ far away from μ, simply because there is only ϵ fraction of corruption! Variance will increase by $(1/\epsilon)^2 \cdot \epsilon = 1/\epsilon$. For small ϵ, will be able to find the variance is abnormally large.

Proposition: Variance large ➡ *Mean **could** be corrupted*

➡ *$\exists$ T, s.t. thresholding at T throws away more bad samples than good samples*

Proposition: Variance normal ➡ *Mean is **NOT** corrupted*

➡ *Output sample mean*

Preliminary

How does this intuition generalize to high dimension?

Filter Algorithm for robust mean estimation with identity covariance.
Input: Set of samples S = $\{x_1, x_2, \ldots, x_n\}$.

1. Compute sample mean $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$.
2. If $\|\hat{\Sigma}\|_{op}$ is close to 1, output $\hat{\mu}$.
3. Otherwise:
   - Find top eigenvector $\mathbf{v}$ of $\hat{\Sigma}$ and threshold T.
   - Throw away $|\mathbf{v}^{\mathrm{T}}(\mathbf{x} - \hat{\mu})| > T$.
   - Goto step 1.

Algorithm Idea

1. Unknown covariance setting:

First robustly estimate Σ using d^2/ϵ^2 unlabeled examples, then reduce to identity covariance setting by scaling $\mathbf{x}$ by $\Sigma^{-1/2}$.

2. Identity covariance setting:

Observe that $\mathrm{E}[y\mathbf{x}] = \mathrm{E}[\mathbf{x}(\mathbf{x}^T\beta + \eta)] = \beta$. Suffices to robustly estimate the distribution mean of $y\mathbf{x}$.

Challenges comparing to [DKKLMS 17] :

1. Previous work on (sub-)Gaussian, but the distribution of $y\mathbf{x}$ is generalized Chi-square.
2. The covariance of $y\mathbf{x}$ is not known, depends on the mean (Cov($y\mathbf{x}$) = $(\sigma^2 + \beta^T\beta)I + \beta\beta^T$).

Proposition[Basic Algorithm]:Given an ε-corrupted set of labeled samples of size (d/ε^2)polylog(d), there exists an efficient algorithm that returns a candidate vector $\hat{\beta}$ s.t. $\|\beta - \hat{\beta}\|_2$ = O($\sqrt{\sigma^2 + \|\beta\|^2}\ \varepsilon$log(1/$\varepsilon$)).

However, the error bound has a dependency on $\|\beta\|_2$ which is not information theoretically necessary($\sigma\epsilon$ by [Gao 17]).

Algorithm Idea

1. Let $\tilde{\beta}$ be the ordinary least square estimator.
2. We run the filter algorithm to robustly estimate the mean of $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})\mathbf{x}$.
   1. If filter algorithm returns sample mean. Notice that sample mean is 0. Hence $\mathrm{E}[(\mathbf{y} - \tilde{\beta}^T\mathbf{x})\mathbf{x}] \leq \sigma\epsilon\log(1/\epsilon) \Rightarrow \tilde{\beta}^T \approx \beta$ Done!
   2. If the filter algorithm returns a set of cleaner samples. Goto step 1.

Filter algorithm either

1. Returns the sample mean.
2. Returns a set of cleaner samples.

How to robustly estimate the mean of $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})\mathbf{x}$?

No sample covariance concentration from the uncorrupted samples*.

If ignore samples with large $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})$, do have concentration!

*need concentration of uncorrupted samples to claim covariance abnormal/normal.

Lowerbound Construction

Regression setting:

Pick $\beta = \sqrt{\epsilon}\mathbf{v}$, where **v** is an uniformly randomly unit vector.

$\mathbf{x} \sim N\left(0, I - \frac{1}{3}\mathbf{v}\mathbf{v}^T\right)$.

Pick σ^2 such that the variance of y is 1.

Corruption scheme:

Corrupt the conditional distribution **x**|y.

Proposition: After ϵ fraction of additive corruption(on the **v** direction), it's hard for SQ algorithm to find the **v** direction.

Summary

Theorem[Main Algorithm]: In the setting where $\mathbf{x} \sim N(0, I), \eta \sim N(0, \sigma^2), y = \beta^T \mathbf{x} + \eta$*, given an* ε*-corrupted set of labeled samples of size* (d/ε^2)*polylog(*d*), there exists an efficient algorithm that returns a candidate vector* $\hat{\beta}$ *s.t.* $\|\beta - \hat{\beta}\|_2 = O(\sigma \varepsilon \log(1/\varepsilon))$*.*

Theorem[Unknown Covariance]: In the setting where $\mathbf{x} \sim N(0, \Sigma), \eta \sim N(0, \sigma^2), y = \beta^T \mathbf{x} + \eta$*, given an* ε*-corrupted set of labeled samples of size* d^2/ε^2*, there exists an efficient algorithm that returns a candidate vector* $\hat{\beta}$ *s.t.* $\|\beta - \hat{\beta}\|_\Sigma = O(\sigma \varepsilon \log(1/\varepsilon))$*.*

Theorem[SQ Lowerbound]: No SQ algorithm for robust linear regression for Gaussian covariates with unknown bounded covariance and random noise with $\sigma^2 \leq 1$ *can output a candidate* $\hat{\beta}$ *with* $\|\hat{\beta} - \beta\|_\Sigma = o(\sqrt{\epsilon})$ *on all instances unless it uses* $2^{\Omega(d)}$ *statistical queries or each query requires* $\Omega(d^2)$ *samples to be simulated.*

Algorithm Idea

How to robustly estimate the mean of $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})\mathbf{x}$?

If ignore samples with large $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})$, do have 
concentration!

What do we do with the samples with large $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})$?

We first run filter algorithm on $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})$, after which the tail of $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})$ is small enough and won't cause trouble for $(\mathbf{y} - \tilde{\beta}^T\mathbf{x})\mathbf{x}$.