Sever: A Robust Meta-Algorithm for Stochastic Optimization

Ilias Diakonikolas[1], Gautam Kamath[2], Daniel M. Kane[3], Jerry Li[4], Jacob Steinhardt[5], Alistair Stewart[1]

(alphabetical order)

[1]USC [2]Waterloo [3]UCSD [4]MSR AI [5]Berkeley

DEFENDING AGAINST DATA POISONING

Main question: can you learn a good classifier from poisoned training data?

DEFENDING AGAINST DATA POISONING

Main question: can you learn a good classifier from poisoned training data?

Given a labeled training set, where an (unknown) ε-fraction of them are adversarially corrupted, can we learn a model which **achieves good accuracy on a clean test set?**

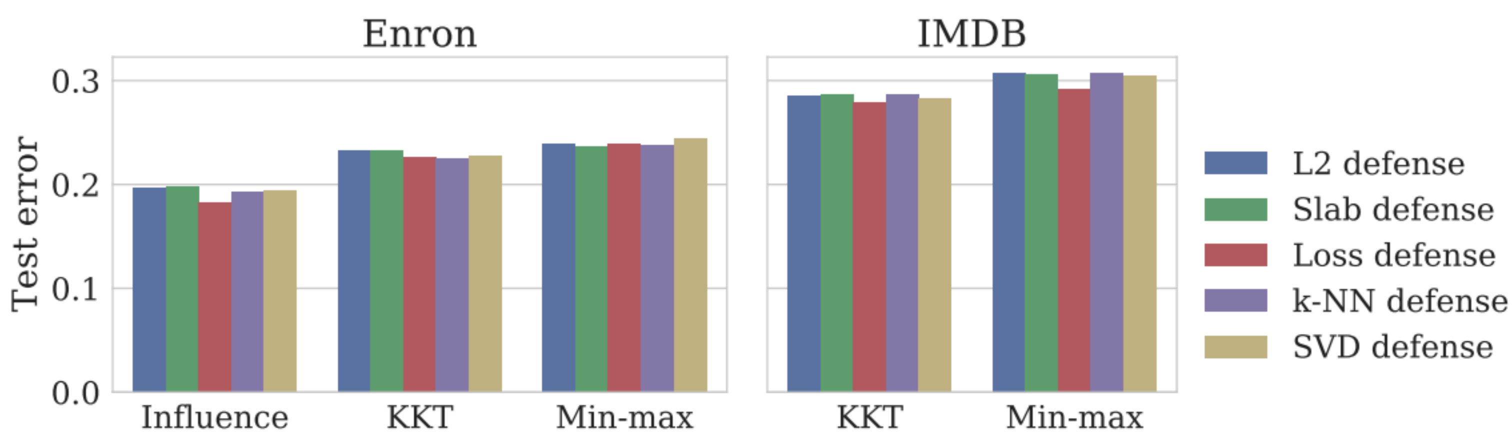
Main question: can you learn a good classifier from poisoned training data?

Example: Training an SVM with 3% poisoned data

DEFENDING AGAINST DATA POISONING

Main question: can you learn a good classifier from poisoned training data?

Example: Training an SVM with 3% poisoned data



[Koh-Steinhardt-Liang '18]

Against known defenses, the test error can go up to 30%!

DEFENDING AGAINST DATA POISONING

Main question: can you learn a good classifier from poisoned training data?

Example: Training an SVM with 3% poisoned data

Lots of work on related problems:

[Barreno-Nelson-Joseph-Tygar’10, Nasrabadi-Tran-Nguyen’11, Biggio-Nelson-Laskov’12, Nguyen-Tran’13, Newell-Potharaju-Xiang-Nita-Rotaru’14, Bhatia-Jain-Kar’15, Diakonikolas-Kamath-Kane-L-Moitra-Stewart’16, Bhatia-Jain-Kamalaruban-Kar’17, Balakrishnan-Du-L-Singh’17, Charikar-Steinhardt-Valiant’17, Steinhardt-Koh-Liang’17, Koh-Liang’17, Prasad-Suggala-Balakrishnan-Ravikumar’18, Diakonikolas-Kong-Stewart’18, Klivans-Kothari-Meka’18, Koh-Steinhardt-Liang’18…]

t-Liang ’18]

Agai … 0%!

OUR RESULTS

We present a framework for robust stochastic optimization

- **Strong theoretical guarantees** against strong adversarial models
- **Outperforms benchmark defenses** on state-of-the-art data poisoning attacks
- Works well in **high dimensions**
- Works with **black-box access** to any learner for any stochastic optimization task

SEVER

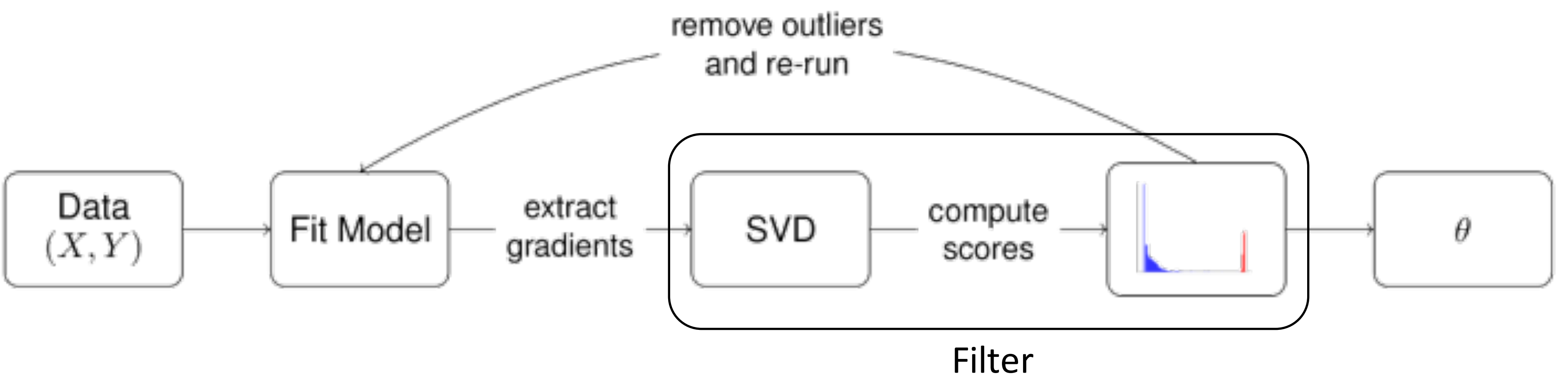
Idea: Until termination:

1. train black box learner to find approximate minima of empirical risk on corrupted training set,
2. then run outlier detection method on the gradients of the loss functions at ERM to remove suspected outliers

SEVER

Idea: Until termination:

1. train black box learner to find approximate minima of empirical risk on corrupted training set,
2. then run outlier detection method on the gradients of the loss functions at ERM to remove suspected outliers



FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

We exploit a novel connection to **robust mean estimation**

FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

We exploit a novel connection to **robust mean estimation**

Filtering [DKKLMS16, DKKLMS17]: Given a set of points $X_1, \ldots, X_n$ drawn from a "nice" distribution, but where an ε-fraction are corrupted, there is a linear time algorithm which either:

FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

We exploit a novel connection to **robust mean estimation**

Filtering [DKKLMS16, DKKLMS17]: Given a set of points $X_1, \ldots, X_n$ drawn from a "nice" distribution, but where an ε-fraction are corrupted, there is a linear time algorithm which either:

1. Certifies that the true mean is close to the empirical mean of the corrupted dataset

FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

We exploit a novel connection to **robust mean estimation**

Filtering [DKKLMS16, DKKLMS17]: Given a set of points $X_1, \ldots, X_n$ drawn from a "nice" distribution, but where an ε-fraction are corrupted, there is a linear time algorithm which either:

1. Certifies that the true mean is close to the empirical mean of the corrupted dataset
2. Removes more bad points than good points

FILTERING AND ROBUST MEAN ESTIMATION

How should we detect outliers from the gradients?

We exploit a novel connection to **robust mean estimation**

Filtering [DKKLMS16, DKKLMS17]: Given a set of points $X_1, \ldots, X_n$ drawn from a "nice" distribution, but where an ε-fraction are corrupted, there is a linear time algorithm which either:

1. Certifies that the true gradient of the loss function is close to 0
2. Removes more bad points than good points

GUARANTEES

Theorem (informal): Suppose we have a distribution $\mathcal{D}$ over convex functions f, and $\mathrm{Cov}\left[\nabla f(\theta)\right] \preccurlyeq \sigma^2 I$. Suppose we have $f_1(\theta), f_2(\theta), \ldots, f_n(\theta)$ drawn from $\mathcal{D}$, where ε-fraction of them are adversarial. Under mild assumptions on $\mathcal{D}$, then given enough samples, SEVER outputs a $\hat{\theta}$ so that w.h.p.

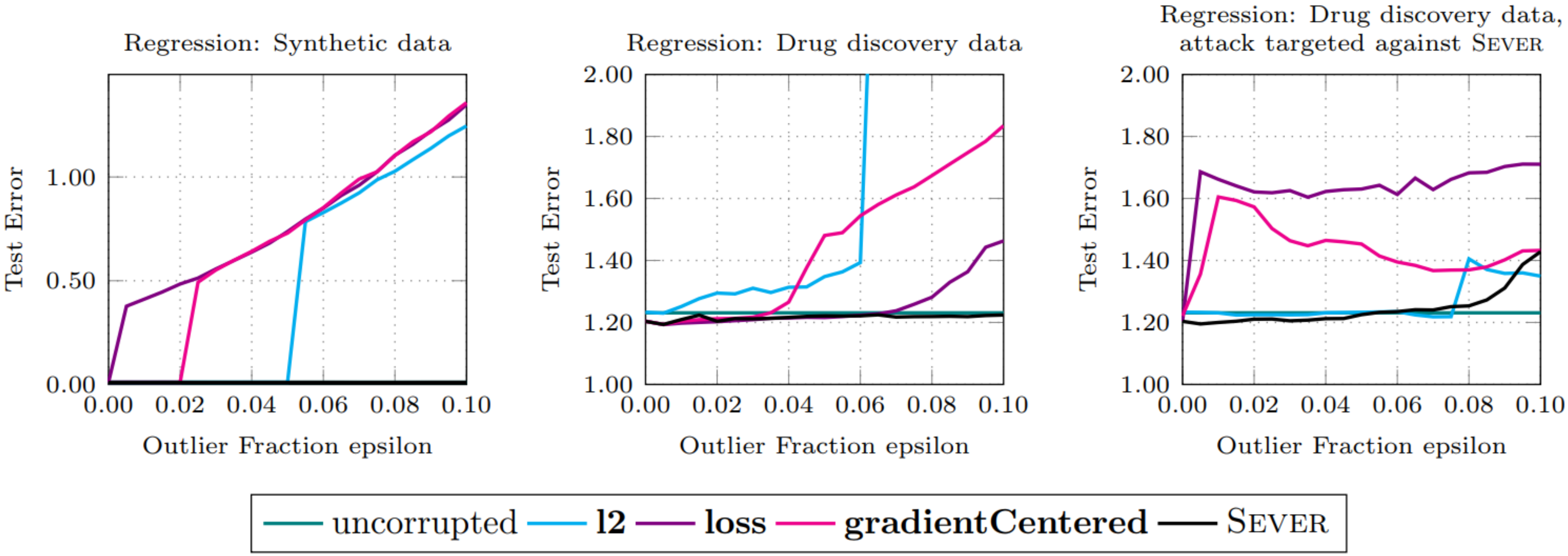
$$\bar{f}(\hat{\theta}) - \min_{\theta} f(\theta) < O\left(\sqrt{\sigma^2 \varepsilon}\right).$$

Can also give results for non-convex objectives

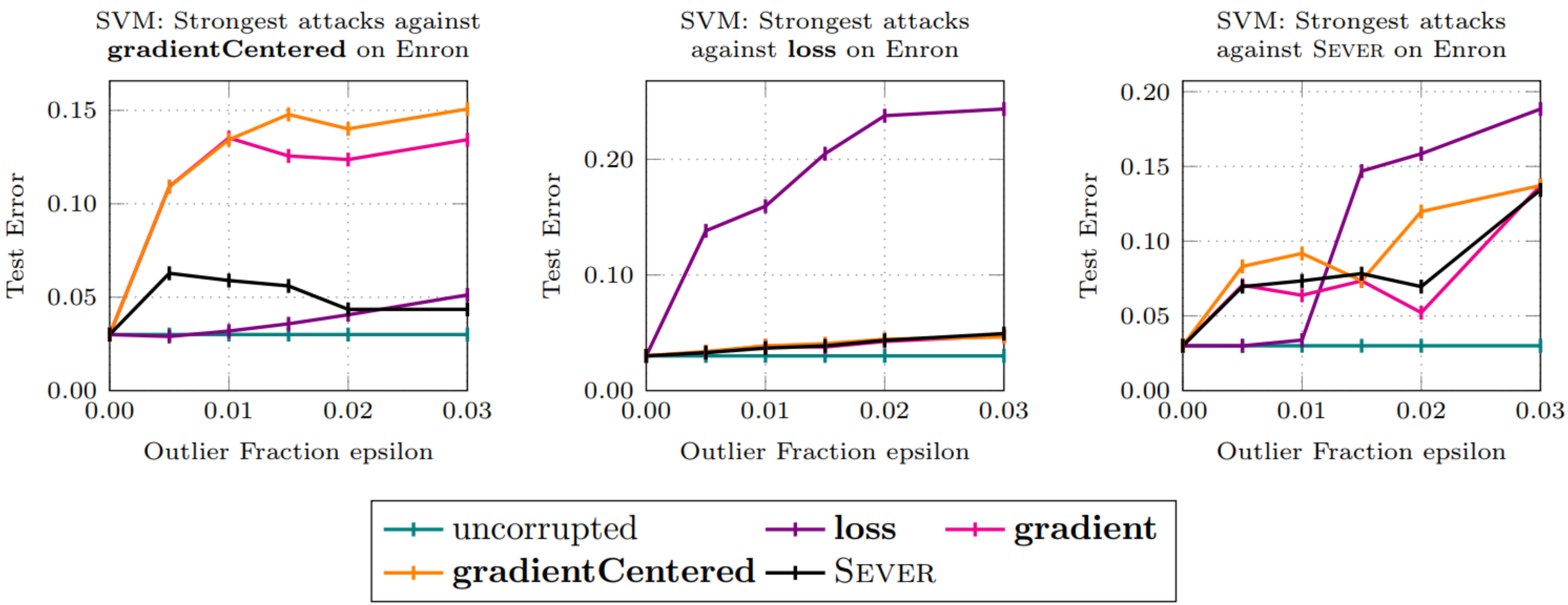
Sample complexity / runtime are polynomial but not super tight

For GLMs (e.g. SVM, regression), we obtain tight(er) bounds

EMPIRICAL EVALUATION: REGRESSION



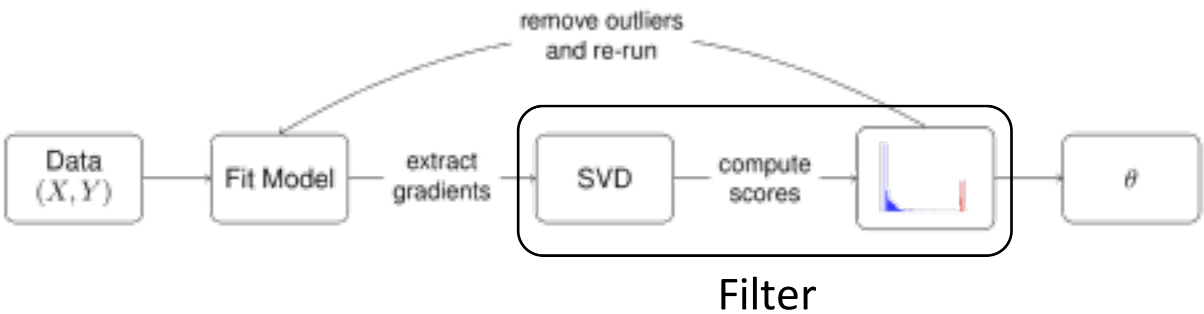
EMPIRICAL EVALUATION: SVM



CONCLUSIONS

Main question: can you learn a good classifier from poisoned data?

Sever is a meta-algorithm for robust stochastic optimization



Based on connections to robust mean estimation

Interested? See poster #143 this evening!