Outlier-Robust Learning of Geometric Concepts

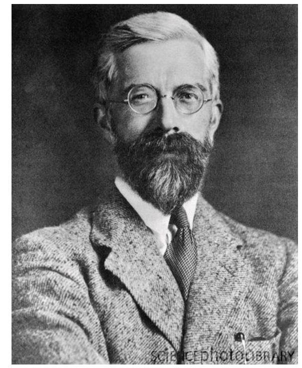
Ilias Diakonikolas (USC)

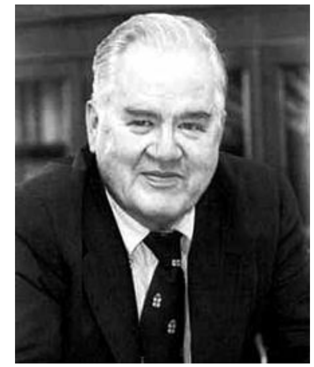
joint work with
Daniel Kane (UCSD) & Alistair Stewart (USC)

Can we develop learning algorithms that are ***robust*** to a ***constant*** fraction of ***corruptions*** in the data?

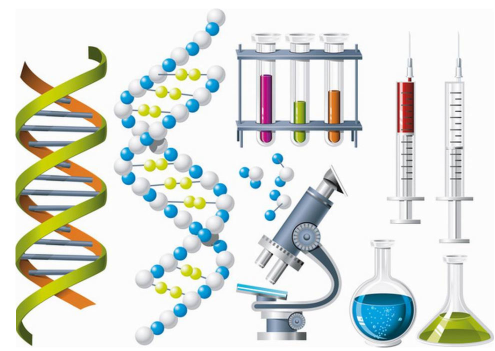
MOTIVATION

- **Model Misspecification/Robust Statistics**:
  Any model only approximately valid.
  Need *stable* estimators
  [Fisher 1920, Huber 1960s, Tukey 1960s]





- **Outlier Removal**: Natural outliers in real datasets.
  Hard to detect in several cases
  [Rosenberg *et al.*, Science'02; Li *et al.*, Science'08;
  Paschou *et al*., Journal of Medical Genetics'10]



- **Reliable/Adversarial/Secure ML**:
  Data poisoning attacks (e.g., crowdsourcing)
  [Biggio *et al.* ICML'12, …]

BACKGROUND: ALGORITHMIC HIGH-DIMENSIONAL ROBUST STATISTICS

Robust estimation in high-dimensions is algorithmically possible!

- Computationally efficient robust estimators that can tolerate a ***constant*** fraction of corruptions.
- General methodology to detect outliers in high dimensions.

Meta-Theorem (Informal): Can obtain *dimension-independent* error guarantees, as long as good data has nice concentration.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16]

Can tolerate a ***constant*** fraction of corruptions:

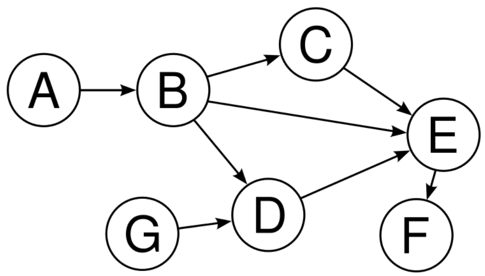
- Mean and Covariance Estimation
- Mixtures of Spherical Gaussians, Mixtures of Balanced Product Distributions

[**Lai-Rao-Vempala, FOCS'16]**

Can tolerate a ***mild sub-constant*** (*inverse logarithmic*) fraction of corruptions:

- Mean and Covariance Estimation
- Independent Component Analysis, SVD

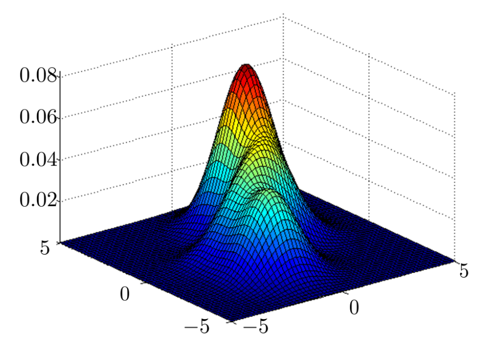
ROBUST *UNSUPERVISED* LEARNING



Robustly Learning Graphical Models
[Cheng-**D**-Kane-Stewart'16,
D-Kane-Stewart'18]

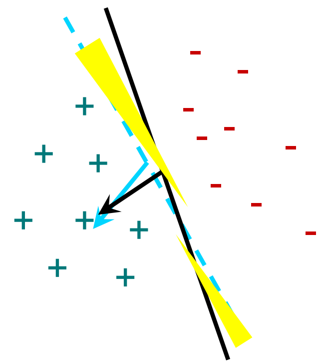
Computational/Statistical-Robustness Tradeoffs
[**D**-Kane-Stewart'17, **D**-Kong-Stewart'18]

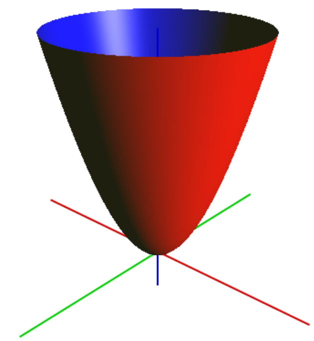
Clustering in Mixture Models
[Charikar-Steinhardt-Valiant'17,
D-Kane-Stewart'18,
Hopkins-Li'18,
Kothari-Steinhardt-Steurer'18]

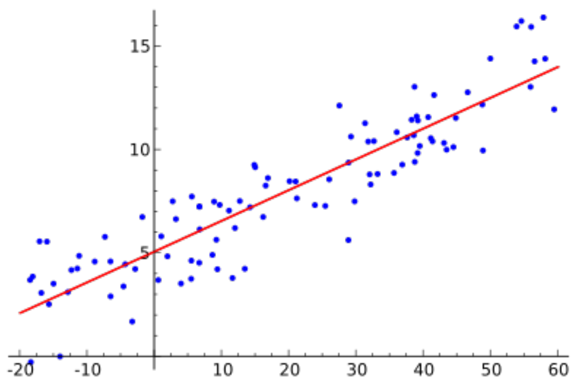


Robust *Supervised* Learning

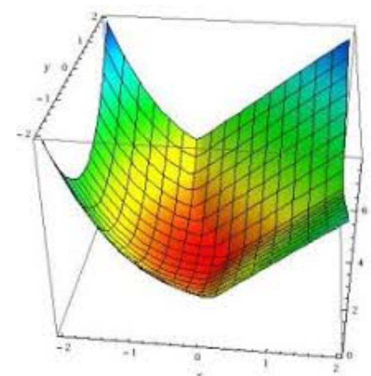
Malicious PAC Learning
[Klivans-Long-Servedio'10,
Awasthi-Balcan-Long'14,
D-Kane-Stewart'18]







Robust Linear Regression
[**D-**Kong-Stewart'18,
Klivans-Kothari-Meka'18]



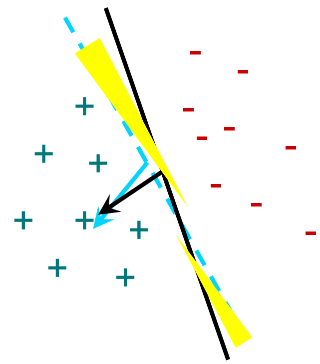
Stochastic (Convex) Optimization
[Prasad-Suggala-Balakrishnan-Ravikumar'18,
D-Kamath-Kane-Li-Steinhardt-Stewart'18]

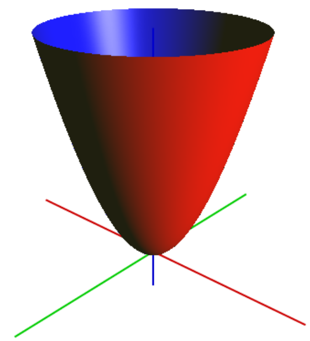
Subsequent Related Works

- Graphical Models [Cheng-D-Kane-Stewart’16, D-Kane-Stewart’18]
- Sparse models (e.g., sparse PCA, sparse regression) [Li’17, Du-Balakrishan-Singh’17, Liu-Shen-Li-Caramanis’18, …]
- List-Decodable Learning [Charikar-Steinhardt-Valiant ’17, Meister-Valiant’18, D-Kane-Stewart’18]
- Robust PAC Learning [Klivans-Long-Servedio’10, Awasthi-Balcan-Long’14, D-Kane-Stewart’18]
- “Robust estimation via SoS” (higher moments, learning mixture models) [Hopkins-Li’18, Kothari-Steinhardt-Steurer’18, …]
- “SoS Free” learning of mixture models [D-Kane-Stewart’18]
- Robust Regression [Klivans-Kothari-Meka’18, D-Kong-Stewart’18, …]
- Robust Stochastic Optimization [Prasad-Suggala-Balakrishnan-Ravikumar’18, D-Kamath-Kane-Li-Steinhard-Stewart’18]
- Near-Linear Time Algorithms [Cheng-D-Ge’19, Cheng-D-Ge-Woodruff’19, …]

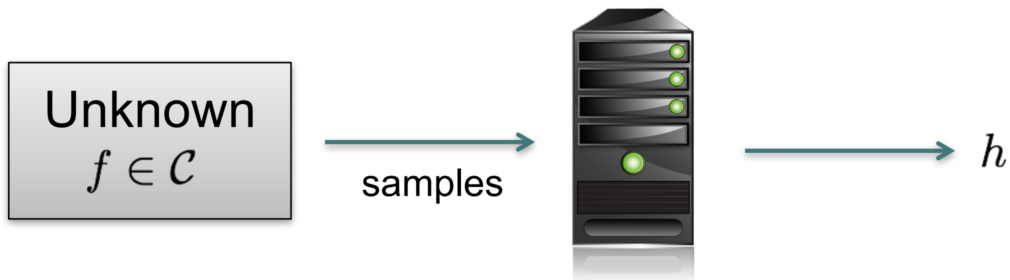
THIS TALK

Malicious PAC Learning
[**D-**Kane-Stewart'18]





THE PAC LEARNING PROBLEM [VALIANT'84]



$\mathcal{C}$: known class of Boolean-valued functions on $\mathbb{R}^n$
D : fixed (unknown) distribution on $\mathbb{R}^n$

- *Input*: labeled sample $\{(x^{(i)}, y_i)\}_{i=1}^m$ where $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$
- *Goal*: compute hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such that $\Pr_{x \sim D}[h(x) \neq f(x)]$ is small

Question: Is there an ***efficient*** **learning algorithm?**

PAC Learning With Adversarial Noise

"nasty" PAC learning [Bshouty-Eiron-Khusilevitz'02]

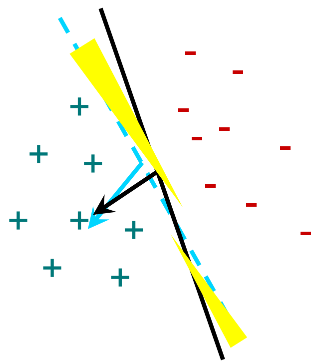
Contamination Model:

Fix $0 < \epsilon < 1/2$. We say that a set of m samples is ϵ-corrupted from $\mathcal{C}$ if it is generated as follows:

- m samples $\{(x^{(i)}, y_i)\}_{i=1}^{m}$ are drawn, where $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$ for some unknown $f \in \mathcal{C}$
- An omniscient adversary inspects these samples and changes arbitrarily an ϵ - fraction of them.

cf. malicious PAC learning [Valiant'85, Kearns-Li'93]
agnostic PAC learning [Haussler'92, Kearns-Shapire-Sellie'94]

This Talk: Geometric Concept Classes

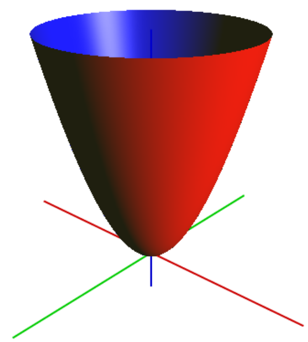


Polynomial Threshold Functions (PTFs)

$f : \mathbb{R}^n \to \{\pm 1\}$ such that

$$f(x) = \mathrm{sgn}(p(x))$$

where $p : \mathbb{R}^n \to \mathbb{R}$ is a degree-d real polynomial.



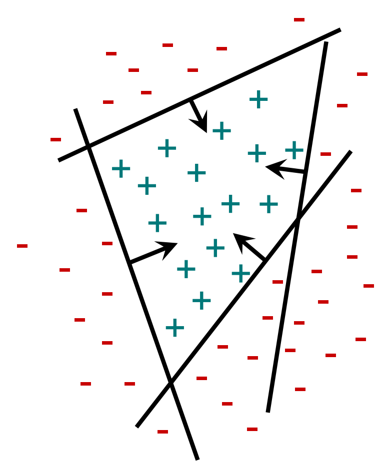
Linear Threshold Functions (Halfspaces)

$f : \mathbb{R}^n \to \{\pm 1\}$ such that

$$f(x) = \mathrm{sgn}(w \cdot x - \theta)$$

where $w \in \mathbb{R}^n, \theta \in \mathbb{R}$

Intersections of LTFs



Previous Work: PAC Learning (No Corruptions)

- Low-degree PTFs efficiently PAC learnable under **any** distribution [Blumer et al. '89]:

  "For *all* $\gamma > 0$, can achieve accuracy γ with $\mathrm{poly}(n^d, 1/\gamma)$ samples and time."

- Intersection of 2 Halfspaces under **any** distribution:

  ?

- Intersection of any constant number of Halfspaces efficiently PAC learnable under "**well-behaved**" distributions e.g., [Baum'91, Blum-Kannan'96, Klivans-O'Donnell-Servedio'02, Vempala'10].

Previous Work: "Robust" PAC Learning

If $0 < \epsilon < 1/2$ is fraction of corruptions, information-theoretic optimal error is $\Theta(\epsilon)$.

Distribution D is arbitrary:

- Can efficiently achieve error $\epsilon \cdot n$ [Kearns-Li'93].
- "Hard" to get **dimension-independent** error, even for LTFs [Daniely'16].

Distribution D is "well-behaved":

- Agnostic learning model (label corruptions): "L_1-regression" algorithm [KKMS'05] can get error $\epsilon + \gamma$ with samples and time $n^{\mathrm{poly}(1/\gamma)}$
- Malicious learning model: $\mathrm{poly}(n, 1/\epsilon)$ time algorithms for *origin-centered* LTFs [Klivans-Long-Servedio'09], [Awasthi-Balcan-Long'14/'17], [Daniely'15].

Origin-centered LTFs only concept class for which efficient malicious PAC learning algorithms known.

What about efficient robust estimation for *more general concept classes*?

Goal: Dimension-independent error guarantees.

THIS TALK: OUR CONTRIBUTION

First efficient robust learning algorithms with **dimension-independent** error guarantees for more general *geometric* concept classes.

- Efficient PAC learning algorithm in nasty noise model for that can tolerate a ***constant*** fraction of corruptions for:
    - low-degree PTFs
    - intersections of constantly many LTFs

  under Gaussian distribution.

- Near-optimal error guarantee for all LTFs.

OUTLINE

Part I: Introduction

- Motivation
- PAC Learning, Geometric Concepts, Robustness
- Prior Work
- Our Contribution

Part II: Robust Learning of Geometric Concepts

- Statements of Results
- Overview of Algorithmic Ideas

Part III: Future Directions

OUTLINE

Part I: Introduction
- Motivation
- PAC Learning, Geometric Concepts, Robustness
- Prior Work
- Our Contribution

Part II: Robust Learning of Geometric Concepts
- Statements of Results
- Overview of Algorithmic Ideas

Part III: Future Directions

Robust PAC Learning of Low-Degree PTFs

Problem: Given m samples $\{(x^{(i)}, y_i)\}_{i=1}^m$ of which $(1-\epsilon)m$ satisfy $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$, for an unknown **degree-*d* PTF** f, compute hypothesis h such that $\Pr_{x\sim D}[h(x) \neq f(x)]$ is small.

Theorem: Let D be any log-concave distribution with known moments up to degree $2d$. There is a $\mathrm{poly}(n^d, 1/\epsilon)$ time algorithm that outputs a degree-d PTF h such that

$$\Pr_{x\sim D}[h(x) \neq f(x)] \leq \epsilon^{\Omega(1/d)}$$

Error Guarantee Independent of n !

- For d=1 under $N(0, I)$, error is $O(\epsilon\sqrt{\log(1/\epsilon)})$
- For d=1, get dimension-independent error for uniform distribution on $\{\pm 1\}^n$

NEAR-OPTIMAL ROBUST PAC LEARNING OF LTFS

Problem: Given m samples $\{(x^{(i)}, y_i)\}_{i=1}^m$ of which $(1-\epsilon)m$ satisfy $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$, for an unknown **LTF** f, compute hypothesis h such that $\Pr_{x \sim D}[h(x) \neq f(x)]$ is small.

Theorem: Let D be $N(0, I)$. There is a $\mathrm{poly}(n, 1/\epsilon)$ time algorithm that outputs an LTF h such that

$$\Pr_{x \sim D}[h(x) \neq f(x)] \leq O(\epsilon)$$

Error guarantee optimal, up to constant factor

cf. [DKS'17] SQ lower bound for robust mean estimation within $o(\epsilon\sqrt{\log(1/\epsilon)})$.

Robust PAC Learning of Polytopes

Problem: Given m samples $\{(x^{(i)}, y_i)\}_{i=1}^m$ of which $(1-\epsilon)m$ satisfy $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$, for an unknown **intersection of k LTFs** f, compute hypothesis h such that $\Pr_{x\sim D}[h(x) \neq f(x)]$ is small.

Theorem: Let D be $N(0, I)$. There is an algorithm that draws $\mathrm{poly}(n, k, 1/\epsilon)$ corrupted labeled examples, runs in time $\mathrm{poly}_k(n, 1/\epsilon)$, and outputs an intersection of k LTFs h such that

$$\Pr_{x\sim D}[h(x) \neq f(x)] \leq k^{O(1)} \cdot \epsilon^{\Omega(1)}$$

Error Guarantee Independent of n !

No non-trivial robust learning algorithm previously known even for k=2.

OUTLINE

Part I: Introduction

- Motivation
- PAC Learning, Geometric Concepts, Robustness
- Prior Work
- Our Contributions

Part II: Robust Learning of Geometric Concepts

- Statements of Results
- Overview of Algorithmic Ideas

Part III: Future Directions

Robust Learning Algorithm for Low-Degree PTFs

Def: Let $f : \mathbb{R}^n \to [-1, 1]$ and D a distribution on $\mathbb{R}^n$. The degree-d Chow parameters of f with respect to D are $\mathbf{E}_{x \sim D}[f(x) m_i(x)]$ for all degree at most d monomials $m_i(x)$

Two-step Procedure:

Step 1: Robustly estimate the degree at most d "Chow parameters" of f.

Step 2: Find a degree-d PTF h with (approximately) these Chow parameters.

Output h.

Robust Estimation of *Low-Degree Chow Parameters* (I)

Def: Let $f : \mathbb{R}^n \to [-1, 1]$ and D a distribution on $\mathbb{R}^n$. The degree-d Chow parameters of f with respect to D are $\mathbf{E}_{x \sim D}[f(x) m_i(x)]$ for all degree at most d monomials $m_i(x)$

Problem: Given m samples $\{(x^{(i)}, y_i)\}_{i=1}^m$ of which $(1-\epsilon)m$ satisfy $x^{(i)} \sim D$ and $y_i = f(x^{(i)})$, for an unknown $f : \mathbb{R}^n \to [-1, 1]$, compute an approximation to the degree-d Chow parameters of f in l_2 - norm.

Theorem: Let D be $N(0, I)$, uniform on $\{\pm 1\}^n$ or any log-concave distribution with known moments up to degree $2d$. There is a $\mathrm{poly}(n^d, 1/\epsilon)$ time algorithm that outputs an approximation with l_2 – error

$$O_d(\epsilon \cdot \log(1/\epsilon)^d) \ .$$

ROBUST ESTIMATION OF *LOW-DEGREE CHOW PARAMETERS* (II)

- Let S be a set of samples from D. Then $\mathbf{E}_{x\sim_u S}[f(x)m_i(x)] \approx \mathbf{E}_{x\sim D}[f(x)m_i(x)]$

- Let S be an ϵ- corrupted set of samples from D.

  $\mathbf{E}_{x\sim D}[f(x)p(x)]$ can be very far from $\mathbf{E}_{x\sim_u S}[f(x)p(x)]$ for some degree-d polynomials p.

Main Idea: "Fix the moments" by iterative filtering
(inspired by [D-Kamath-Kane-Lee-Moitra-Stewart'16])

- Detect whether there is a degree-d polynomial whose *empirical* variance is much larger than its variance under D.
- If no such polynomial exists, use empirical.
- Otherwise, can detect and remove outliers.

Robust Learning Algorithm for Polytopes

Two-step Procedure:

Step 1: Robustly estimate the degree at most 2 "Chow parameters" of f.

Step 2: Project to an approximate $k+1$ dimensional subspace V and solve the problem by using a cover on V. Let g be the output.

Output $h(x) = g(\pi_V(x))$.

Main challenge: Analysis of Correctness

OUTLINE

Part I: Introduction

- Motivation
- PAC Learning, Geometric Concepts, Robustness
- Prior Work
- Our Contribution

Part II: Robust Learning of Geometric Concepts

- Statements of Results
- Overview of Algorithmic Ideas

Part III: Future Directions

SUMMARY AND CONCLUSIONS

- First computationally efficient robust PAC learners with **dimension-independent** error guarantees for low-degree PTFs and intersections of LTFs.

- Near-optimal error guarantees for robust PAC learning of LTFs.

- General procedure for robustly learning low-degree Chow parameters.

Future Directions

General Algorithmic Theory of Robustness

- Pick your favorite high-dimensional learning problem for which a (non-robust) efficient algorithm is known.
- Make it robust!

Concrete Open Questions:

- Near-optimal error guarantees, e.g., $O_d(\epsilon)$ error for degree-d PTFs
- More general classes of distributions
- Practical Algorithms?
  [D-Kamath-Kane-Moitra-Lee-Stewart, ICML'17] [DKKLSS'18]
- Alternate models of robustness?

Thank you!
Questions?

Related Materials:

- **TTI-Chicago Summer Workshop Program**

  http://www.ttic.edu/summer-workshop-2018/

  (Aug. 13-17 2018, co-organized with Daniel Kane)

- **Simons Institute, Foundations of Data Science Program**

  https://simons.berkeley.edu/data-science-2018-2

  (Oct. 29-Nov. 2 2018, co-organized with Montanari, Candes, Vempala)