Algorithmic High-Dimensional Robust Statistics

Ilias Diakonikolas (UW Madison)

IFDS Summer School

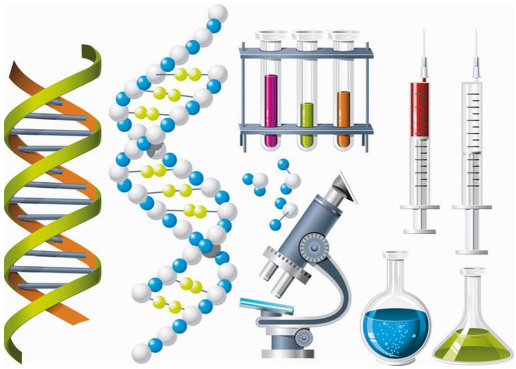
July 2021

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

PART I: INTRODUCTION

MOTIVATION

- **Model Misspecification/Robust Statistics**
  [Fisher 1920s, Tukey 1960s, Huber 1960s]

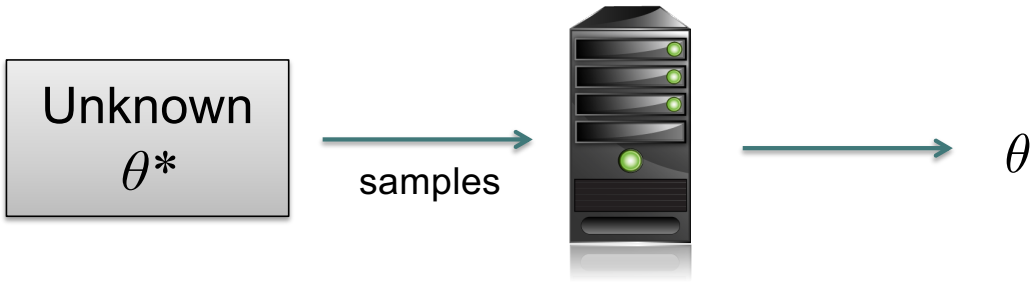


- **Outlier Detection/Removal**

- **Adversarial/Secure ML**

So Many Misleading, "Fake" Reviews

The Statistical Learning Problem



- *Input*: sample generated by a **statistical model** with unknown θ^*
- *Goal*: estimate parameters θ so that $\theta \approx \theta^*$

Question 1: Is there an *efficient* learning algorithm?

Main performance criteria:
- Sample size
- Running time
- **Robustness**

Question 2: Are there *tradeoffs* between these criteria?

(Outlier-) Robustness in a Generative Model

Strong Contamination Model:

Let $\mathcal{F}$ be a family of statistical models.
We say that a set of N samples is ϵ -corrupted from $\mathcal{F}$ if it is generated as follows:

- N samples are drawn from an unknown $F \in \mathcal{F}$
- An omniscient adversary inspects these samples and changes arbitrarily an ϵ-fraction of them.

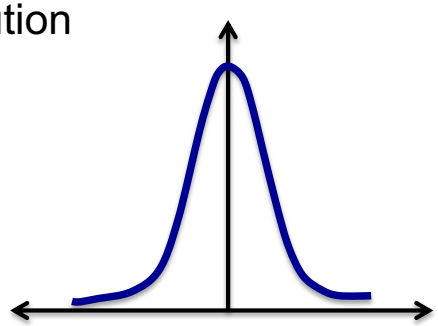
cf. Huber's contamination model [1964]

EXAMPLE: PARAMETER ESTIMATION

Given i.i.d. samples from an unknown distribution

e.g., a 1-D Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$



how do we accurately estimate its parameters?

empirical mean:

$$\frac{1}{N}\sum_{i=1}^{N} X_i \rightarrow \mu$$

empirical variance:

$$\frac{1}{N}\sum_{i=1}^{N} (X_i - \overline{X})^2 \rightarrow \sigma^2$$

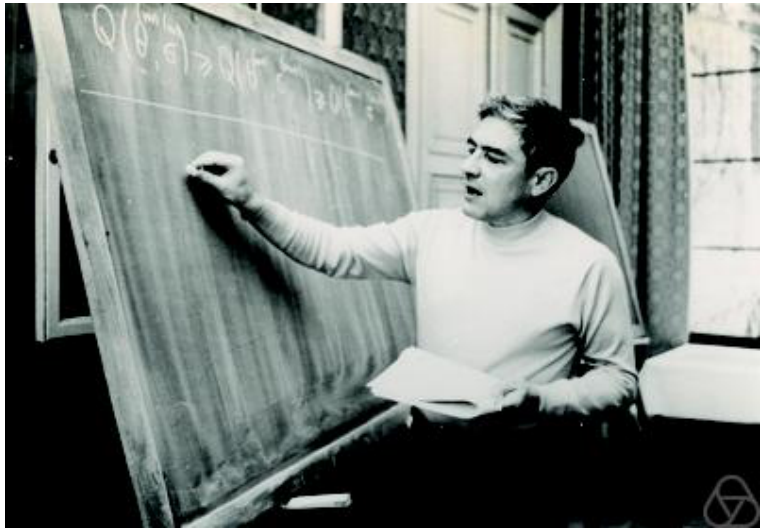
R. A. Fisher

Maximum Likelihood
(1920s)

J. W. Tukey

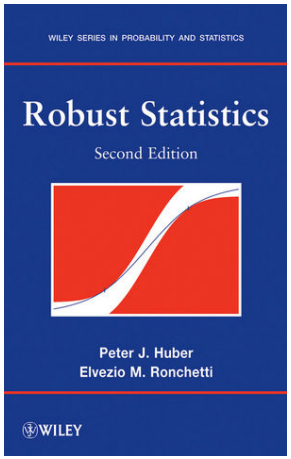
Model Misspecification ?
(1960s)

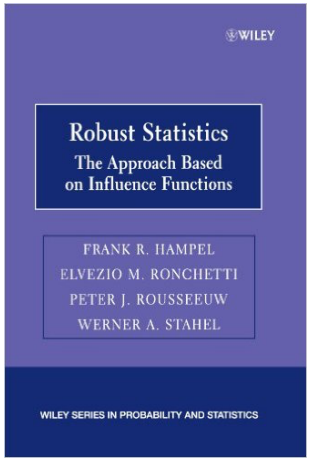
Peter J. Huber

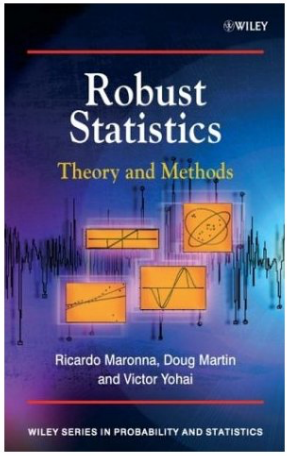


“Robust Estimation of a Location Parameter”
Annals of Mathematical Statistics, 1964.

Robust Statistics







What estimators behave well in the presence of outliers?

Robust Estimation: One Dimension

Given **corrupted** samples from a *one-dimensional* Gaussian, can we accurately estimate its parameters?

- A single corrupted sample can arbitrarily corrupt the empirical mean and variance
- But the **median** and **interquartile range** work

Fact [Folklore]: Given a set S of N ϵ-corrupted samples from a one-dimensional Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$

with high constant probability we have that:

$$|\widehat{\mu} - \mu| \leq O\left(\epsilon + \sqrt{1/N}\right) \cdot \sigma$$

where $\widehat{\mu} = \mathrm{median}(S)$.

What about robust estimation in *high-dimensions*?

High-Dimensional Robust Mean Estimation

Robust Mean Estimation: Given an ϵ - corrupted set of samples from an **unknown mean**, identity covariance Gaussian $\mathcal{N}(\mu, I)$ in d dimensions, recover $\widehat{\mu}$ with

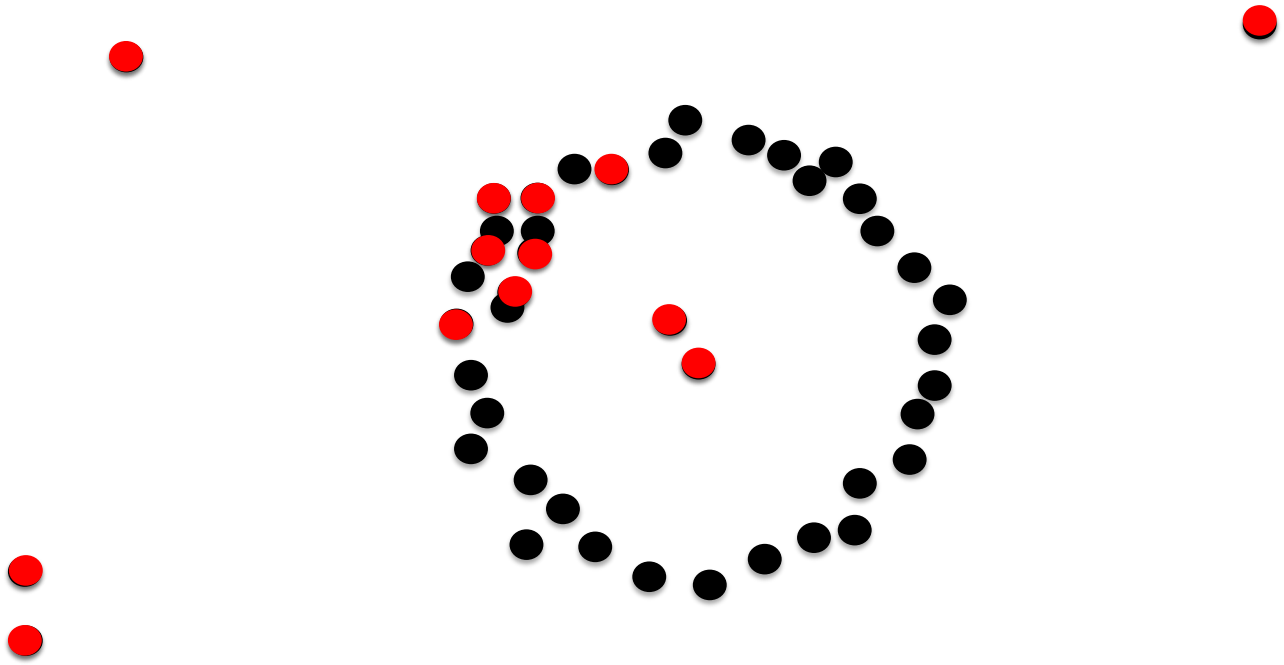
$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) + O(\sqrt{d/N})$$

Remark: Above convergence rate is optimal [Tukey'75, Donoho'82]

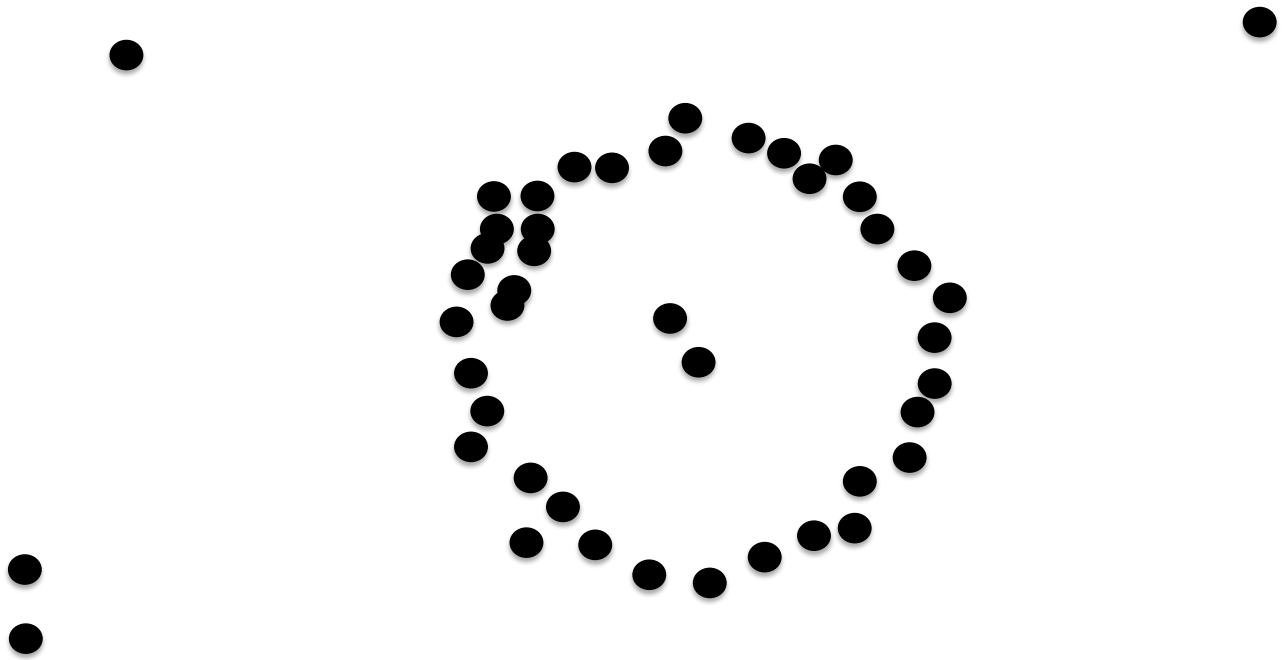
Previous Approaches: Robust Mean Estimation

Estimator	Error Rate	Running Time
Distance-Based Pruning	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Coordinate-wise Median	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Geometric Median	$\Theta(\epsilon\sqrt{d})$ ✗	$\mathrm{poly}(d, N)$ ✓
Tukey Median	$\Theta(\epsilon)$ ✓	NP-Hard ✗
Tournament	$\Theta(\epsilon)$ ✓	$N^{O(d)}$ ✗

DISTANCE-BASED PRUNING



Distance-Based Pruning = Naïve Outlier Removal

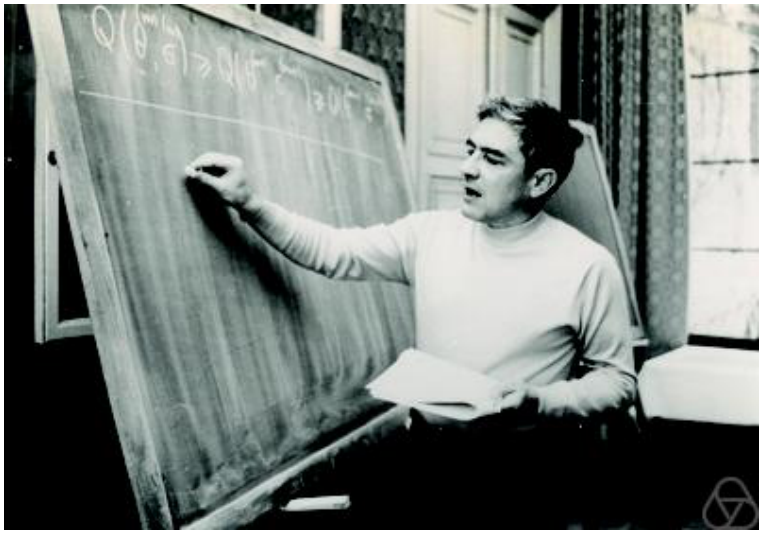


HIGH-DIMENSIONAL ROBUST STATISTICS: 1960-2016

All known estimators are either **require exponential time to compute** or can tolerate a **negligible fraction of outliers**.

Is robust estimation *algorithmically* possible in high-dimensions?

Peter J. Huber, 1975



"[…] Only simple algorithms (i.e., with **a low degree of computational complexity**) will survive the onslaught of huge data sets. This runs counter to recent developments in computational robust statistics**. It appears to me that none of the above problems will be amenable to a treatment through theorems and proofs.** They will have to be attacked by heuristics and judgment, and by alternative "what if" analyses.[…]"

Robust Statistical Procedures, 1996*, Second Edition.*

Robust estimation in high-dimensions is algorithmically possible!

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

Meta-Theorem (Informal): Can obtain *dimension-independent* error guarantees, if distribution on inliers has nice concentration.

First Algorithmic Progress In Unsupervised Setting

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16/SICOMP'19/CACM'21]

Can tolerate ***constant*** fraction of corruptions.

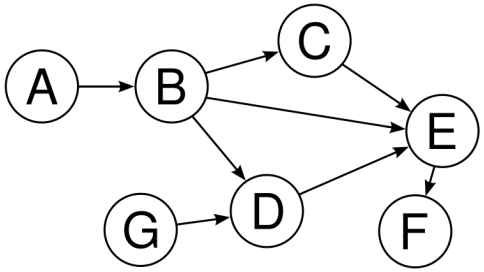
- Mean and Covariance Estimation
- Mixtures of Spherical Gaussians, Mixtures of Balanced Product Distributions

[Lai-Rao-Vempala, FOCS'16]

Can tolerate ***inverse logarithmic*** fraction of corruptions.

- Mean and Covariance Estimation
- Independent Component Analysis, SVD

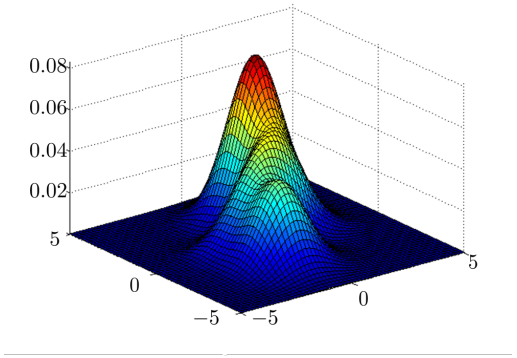
BEYOND ROBUST STATISTICS: ROBUST *UNSUPERVISED* LEARNING



Robustly Learning Graphical Models
[Cheng-**D**-Kane-Stewart'16,
D-Kane-Stewart'18,
D-KaneStewart-Sun'21]

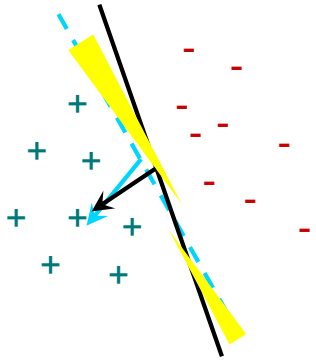
Computational/Statistical-Robustness Tradeoffs
[**D**-Kane-Stewart'17, **D**-Kong-Stewart'18, Hopkins-Li'19,
Brennan-Bresler'20, **D**-Kane-Pensia-Pittas'21, …]

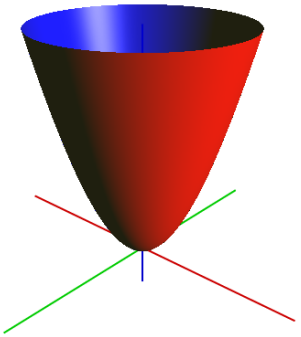
Robustly Learning Mixture Models
[Charikar-Steinhardt-Valiant'17, **D**-Kane-Stewart'18,
Hopkins-Li'18, Kothari-Steinhardt-Steurer'18,
Bakshi-Kothari'20, **D**-Hopkins-Kane-Kothari'20,
Liu-Moitra'20, Bakshi-**D**-Jia-Kane-Kothari-Vempala'20]

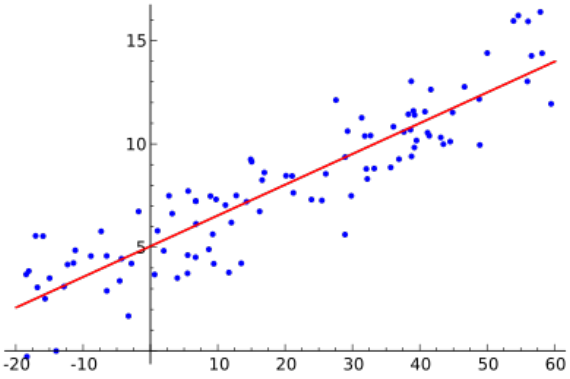


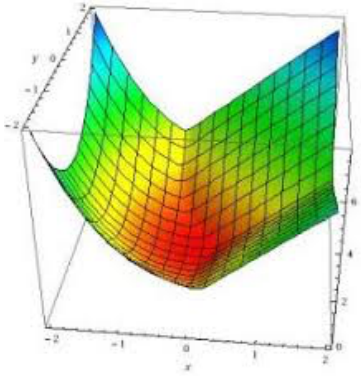
ROBUST *SUPERVISED* LEARNING

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









Robust Linear Regression
[Klivans-Kothari-Meka'18, **D-**Kong-Stewart'18, Bakshi-Prasad'20, …]

Stochastic Convex Optimization
[Prasad-Suggala-Balakrishnan-Ravikumar'19, **D**-Kamath-Kane-Li-Steinhardt-Stewart'19]

Subsequent Related Works

- Sparse Models [Balakrishan-Du-Li-Singh'17, D-Karmalkar-Kane-Price-Stewart'19, Liu-Shen-Li-Caramanis'19,…]
- Graphical Models [Cheng-D-Kane-Stewart'18, D-Kane-Stewart-Sun'21]
- Robust Regression/Classification [D-Kane-Stewart'18, Klivans-Kothari-Meka'18, D-Kong-Stewart'19 Bakshi-Prasad'21, …]
- Robust Stochastic Optimization [Prasad-Suggala-Balakrishnan-Ravikumar'18, D-Kamath-Kane-Li-Steinhard-Stewart'18, …]
- Robust Estimation via SoS [Hopkins-Li'18, Kothari-Steinhardt-Steurer'18, Karmalkar-Klivans-Kothari'19, Raghavendra-Yau'19, Bakshi-Kothari'20, D-Hopkins-Kane-Karmalkar'20, Liu-Moitra'20, Bakshi-D-Jia-Kane-Kothari-Vempala'20, …]
- Near-Linear Time Algorithms [Chen-D-Ge'18, Cheng-D-Ge-Woodruff'19, Depersin-Lecue'19, Dong-Hopkins-Li'19, Li-Ye'20, Cherapanamjeri-Mohanty-Yau'20, D-Kane-Koongsgard-Li-Tian'21, …]
- Computational-Statistical Tradeoffs [D-Kane-Stewart'17, D-Kong-Stewart'19, Hopkins-Li'19, …]
- Connections to Non-Convex Optimization [Chen-D-Ge-Soltanolkotabi'20, Zhu-Jiao-Steinhardt'20, …]
- List-Decodable Learning [Charikar-Steinhardt-Valiant'17, D-Kane-Stewart'18, Meister-Valiant'18, Karmalkar-Klivans-Kothari'19, Raghavendra-Yau'19, D-Kane-Koongsgard'20, D-Kane-Koongsgard-Li-Tian'21, …]
- Applications in Data Analysis [D-Kamath-Kane-Li-Moitra-Stewart'17, Tran-Li-Madry'18, D-Kamath-Kane-Li-Steinhardt-Stewart'19, Hayase-Kong-Somani-Oh'21, …]

High-Dimensional Robust Mean Estimation

Robust Mean Estimation: Gaussian Case

Problem: Given an ϵ-corrupted set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ from an unknown distribution D in a known family $\mathcal{F}$, estimate the mean μ of D.

Theorem 1: Let $\epsilon < 1/2$. If D is a spherical Gaussian, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) + O(\sqrt{d/N})$$

in the **additive contamination** model.

First-term of RHS Independent of d !

[D-Kamath-Kane-Li-Moitra-Stewart, SODA'18]

ROBUST MEAN ESTIMATION: *SUB*-GAUSSIAN CASE

Problem: Given an ϵ-corrupted set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ from an unknown distribution D in a known family $\mathcal{F}$, estimate the mean μ of D.

Theorem 2: Let $\epsilon < 1/2$. If D is a spherical *sub-Gaussian*, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}) + O(\sqrt{d/N}) \ .$$

in the **strong contamination** model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16, ICML'17]

Robust Mean Estimation: Bounded Covariance Case

Problem: Given an ϵ-corrupted set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ from an unknown distribution D in a known family $\mathcal{F}$, estimate the mean μ of D.

Theorem 3: Let $\epsilon < 1/2$. If D has covariance $\Sigma \preceq \sigma^2 \cdot I$, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\sigma\sqrt{\epsilon}) + O(\sqrt{d/N}) .$$

in the **strong contamination** model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, ICML'17; Steinhardt, Charikar, Valiant, ITCS'18]

Robust Mean Estimation: Summary

Assumptions on Inliers	Information-Theoretic Bound	Computationally Efficient Estimators	Reference
Gaussian with $\Sigma = I$	$\Theta(\epsilon)$	$O(\epsilon)$	Additive Contamination* [DKKLMS, SODA'18]
Subgaussian with $\Sigma = I$	$\Theta(\epsilon\sqrt{\log(1/\epsilon)})$	$O(\epsilon\sqrt{\log(1/\epsilon)})$	[DKKLMS, FOCS'16]
Bounded t-th Moments $\Sigma = I$	$\Theta(\epsilon^{1-1/t})$	$O(\epsilon^{1-1/t})$	Folklore (see, e.g., survey [DK19])
Unknown Covariance $\Sigma \preceq I$	$\Theta(\sqrt{\epsilon})$	$O(\sqrt{\epsilon})$	[DKKLMS, ICML'17; SCV, ITCS'18]
Bounded t-th Moments	$\Theta(\epsilon^{1-1/t})$	$O(\epsilon^{1-1/t})$	"Niceness" Assumption* [HL, STOC'18; KS, STOC'18]

PART II: BASIC ALGORITHMIC TECHNIQUES

OUTLINE

Part II

- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering

OUTLINE

Part II

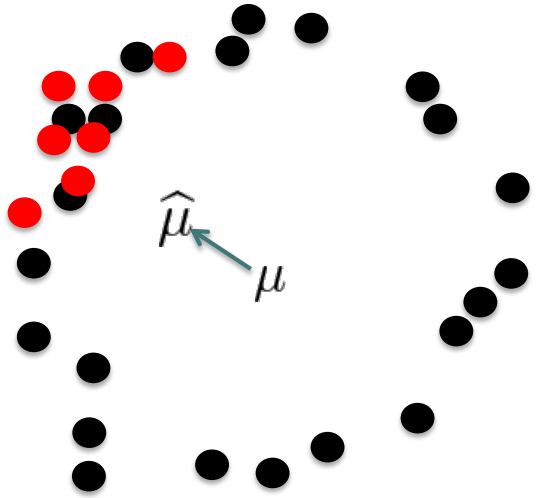
- **Certificate of Robustness**
- Recursive Dimension Halving
- Iterative Filtering

Certificate Of Robustness For Empirical Estimator

Idea #1 [DKKLMS'16, LRV'16]: If the empirical covariance is "close to what it should be", then the empirical mean works.

Certificate For Empirical Mean

Detect when the empirical estimator *may* be compromised



$$\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$$

= uncorrupted

= corrupted

There is *no* direction of large empirical variance

Key Lemma: Let $X_1, X_2, \ldots, X_N$ be an ϵ-corrupted set of samples from $\mathcal{N}(\mu, I)$ and $N = \Omega(d/\epsilon^2)$, then for

$$\text{(1)}\ \ \widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{(2)}\ \ \widehat{\Sigma} \triangleq \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$$

with high probability we have:

$$\|\widehat{\Sigma}\|_2 \leq 1 + O(\epsilon \log(1/\epsilon)) \rightarrow \|\widehat{\mu} - \mu\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$$

in **strong** contamination model.

Key Lemma: Let $X_1, X_2, \ldots, X_N$ be an ϵ-corrupted set of samples from $\mathcal{N}(\mu, I)$ and $N = \Omega(d/\epsilon^2)$, then for

(1) $\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$ (2) $\widehat{\Sigma} \triangleq \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$

with high probability we have:

$$\|\widehat{\Sigma}\|_2 \leq 1 + \delta \quad \rightarrow \quad \|\widehat{\mu} - \mu\|_2 \leq O(\sqrt{\delta\epsilon} + \epsilon\sqrt{\log(1/\epsilon)})$$

in **strong** contamination model.

Idea #2 [DKKLMS'16]: Removing *any* ϵ - fraction of good points does not move the empirical mean and covariance by much.

Remarks on Key Lemma

- Statement applies for spherical distributions with sub-Gaussian tails.
- Essentially same argument goes through if covariance is *approximately* known.

- Argument extends for distributions with known covariance and weaker concentration.

If D is isotropic with *sub-exponential* tails:

$$\|\widehat{\Sigma}\|_2 \leq 1+\delta \longrightarrow \|\widehat{\mu}-\mu\|_2 \leq O(\sqrt{\delta\epsilon}+\epsilon\log(1/\epsilon)) \ .$$

If D satisfies $\Sigma \preceq I$:

$$\|\widehat{\Sigma}\|_2 \leq 1+\delta \longrightarrow \|\widehat{\mu}-\mu\|_2 \leq O(\sqrt{\delta\epsilon}+\sqrt{\epsilon}) \ .$$

OUTLINE

Part II

- Certificate of Robustness
- **Recursive Dimension Halving**
- Iterative Filtering

Idea #3 [LRV'16]: Additive corruptions can move the covariance in *some* directions, but *not in all* directions simultaneously.

RECURSIVE DIMENSION-HALVING [LRV'16]

LRV Procedure:

Step #1: Find large subspace where "standard" estimator works.
Step #2: Recurse on complement.

Combine Results.

Can reduce dimension by factor of 2 in each recursive step.

FINDING A GOOD SUBSPACE (I)

"Good subspace ***G***" = one where the empirical mean works

By **Key Lemma**, sufficient condition is:

> Projection of empirical covariance on ***G*** has no large eigenvalues.

- Also want ***G*** to be "high-dimensional".

Question: How do we find such a subspace?

Finding a Good Subspace (II)

Good Subspace Lemma: Let $X_1, X_2, \ldots, X_N$ be an *additively* ϵ-corrupted set of $N = \Omega(d \log d/\epsilon^2)$ samples from $\mathcal{N}(\mu, I)$. ***After naïve pruning***, we have that

$$\lambda_{d/2}(\widehat{\Sigma}) \leq 1 + O(\epsilon)$$

Corollary: Let W be the span of the bottom $d/2$ eigenvalues of $\widehat{\Sigma}$. Then W is a good subspace.

RECURSIVE DIMENSION-HALVING ALGORITHM [LRV'16]

Algorithm works as follows:

- Remove gross outliers (e.g., naïve pruning).
- Let W, V be the span of bottom $d/2$ and upper $d/2$ eigenvalues of $\widehat{\Sigma}$ respectively .
- Use empirical mean on W.
- Recurse on V (If the dimension is one, use median).

Error Analysis:

$O(\log d)$ levels of the recursion → final error of $O(\epsilon\sqrt{\log d})$

OUTLINE

Part II

- Certificate of Robustness
- Recursive Dimension Halving
- **Iterative Filtering**

Idea #4 [DKKLMS'16]: Iteratively "remove outliers" in order to "fix" the empirical covariance.

ITERATIVE FILTERING [DKKLMS'16]

Iterative Two-Step Procedure:

Step #1: Test certificate of robustness of "standard" estimator

Step #2: If certificate is violated, detect and remove outliers

Iterate on "cleaner" dataset.

General recipe that works in general settings.

Let's see how this works for robust mean estimation.

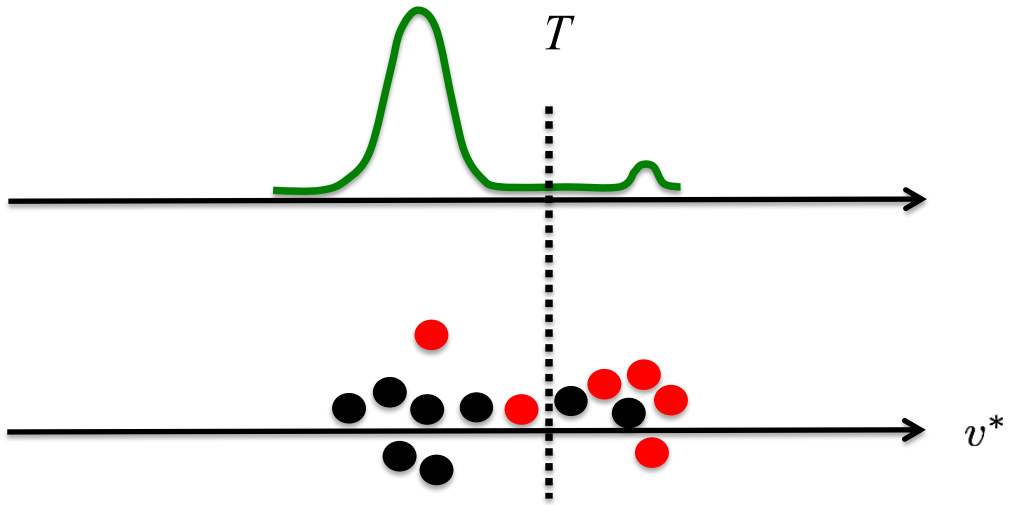
Filtering Subroutine

Either output empirical mean or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Let v^* be the direction of maximum variance.



Filtering Subroutine

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Let v^* be the direction of maximum variance.

- Project all the points on the direction of v^*.
- Find a threshold T such that

$$\mathbf{Pr}_{X \sim_U S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] \geq 8 \cdot e^{-T^2/2}.$$

- Throw away all points x such that

$$|v^* \cdot x - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1$$

- Iterate on new dataset.

Filtering Subroutine: Analysis Sketch

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Claim: In each iteration, we remove more outliers than inliers.

After a bounded number of iterations, we stop removing points.

Eventually the empirical mean works

Runtime: $\tilde{O}(Nd^2)$

FILTERING PSEUDO-CODE

Input: ϵ-corrupted set S from $\mathcal{N}(\mu, I)$

Output: Set $S' \subseteq S$ that is ϵ'-corrupted, for some $\epsilon' < \epsilon$

OR robust estimate of the unknown mean μ

1. Let $\widehat{\mu}_S, \widehat{\Sigma}_S$ be the empirical mean and covariance of the set S.
2. **If** $\|\widehat{\Sigma}_S\|_2 \leq 1 + C\epsilon \log(1/\epsilon)$, for an appropriate constant $C > 0$:
   Output $\widehat{\mu}_S$
3. **Otherwise**, let (λ^*, v^*) be the top eigenvalue-eigenvector pair of $\widehat{\Sigma}_S$.
4. Find $T > 0$ such that

$$\mathbf{Pr}_{X \sim_U S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] \geq 8 \cdot e^{-T^2/2}.$$

5. **Return**

$$S' = \{x \in S : |v^* \cdot x - \text{median}(\{v^* \cdot x, x \in S\})| \leq T + 1\}.$$

Remarks on Filtering Method(s)

- For known covariance sub-Gaussian case, filter relied on violation of concentration.

- This extends to weaker concentration, as long as covariance is (approximately) known.
- For example, for *sub-exponential* concentration, filter would be:

> Find $T > 0$ such that $\mathbf{Pr}_{X \sim_U S}[|v^* \cdot (X - \widehat{\mu})| > T] \geq 8 \cdot e^{-T}$.

- For *the bounded covariance* setting, *randomized* filtering / down-weighting.

> Remove point x with probability proportional to $(v^* \cdot (x - \widehat{\mu}))^2$.

- Analogue of Claim 1: Remove more outliers than inliers *in expectation*.

PART III: EXTENSIONS

OUTLINE

Part III

- General Framework for Robust Mean Estimation
- Robust Stochastic Optimization
- Learning with Majority of Outliers

OUTLINE

Part III

- **General Framework for Robust Mean Estimation**
- Robust Stochastic Optimization
- Learning with Majority of Outliers

Non-Convex Optimization Formulation (I)

Optimization Formulation:

Assign *weights* to the samples so that weighted empirical mean works.

Let $$\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \right\}$$

$$\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i \quad \text{and} \quad \widehat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \widehat{\mu}_w)(X_i - \widehat{\mu}_w)^T .$$

Generalization of Key Lemma: For any $w \in \Delta_{N,2\epsilon}$

$$\|\widehat{\Sigma}_w\|_2 \le 1 + O(\epsilon \log(1/\epsilon)) \quad \rightarrow \quad \|\widehat{\mu}_w - \mu\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$$

Non-Convex Optimization Formulation (II)

Notation: $\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \right\}$

$$\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i \qquad \widehat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \widehat{\mu}_w)(X_i - \widehat{\mu}_w)^T .$$

Generalization of Key Lemma

$$\|\widehat{\Sigma}_w\|_2 \le 1 + O(\epsilon \log(1/\epsilon)) \quad \rightarrow \quad \|\widehat{\mu}_w - \mu\|_2 = O(\epsilon \sqrt{\log(1/\epsilon)})$$

Non-Convex Formulation:

$$\min_w \|\widehat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}$$

Non-Convex Optimization Formulation (III)

Notation: $\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \right\}$

$$\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i \qquad \widehat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \widehat{\mu}_w)(X_i - \widehat{\mu}_w)^T .$$

Non-Convex Formulation:

$$\min_w \|\widehat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}$$

Algorithmic Approaches:

- This is what filtering does!
- Ellipsoid Method [DKKLMS'16]
- Bi-level optimization [Cheng-D-Ge'18] (near-linear time!)
- Gradient Descent [Cheng-D-Ge-Soltanolkotabi, ICML'20]

Robust Mean Estimation via Gradient-Descent

Notation: $\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \right\}$

$$\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i \qquad\qquad \widehat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \widehat{\mu}_w)(X_i - \widehat{\mu}_w)^T .$$

Non-Convex Formulation:

$$\min_w \|\widehat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}$$

Theorem [Cheng-D-Ge-Soltanolkotabi, ICML'20]

Any approximate stationary point w defines $\widehat{\mu}_w$ that is close to μ .

See also [Zhu et al., Arxiv, May 2020]

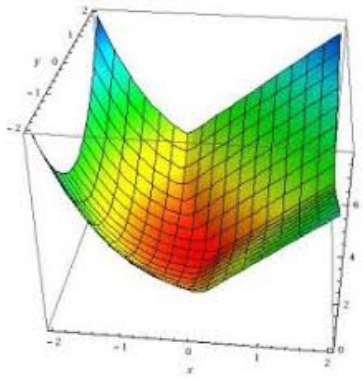
OUTLINE

Part III

- General Framework for Robust Mean Estimation
- **Robust Stochastic Optimization**
- Learning with Majority of Outliers

Robust Stochastic Optimization

Sever: A Robust Meta-Algorithm for Stochastic Optimization.
[D-Kamath-Kane-Li-Steinhardt-Stewart, ICML'19]



Robust Stochastic Convex Optimization

Problem: Given loss function $\ell(X, w)$ and ϵ- corrupted samples from a distribution $\mathcal{D}$ over X, minimize $f(w) = \mathbb{E}_{X \sim \mathcal{D}}[\ell(X, w)]$

Difficulty: Corrupted data can move the gradients.

Theorem: Suppose ℓ is convex and $\mathrm{Cov}_{\mathrm{X} \sim \mathcal{D}}[\nabla \ell(\mathrm{X}, \mathrm{w})] \preceq \sigma^2 \cdot I$.
Under mild assumptions on $\mathcal{D}$, can recover a point such that

$$f(\hat{w}) - \min_{w} f(w) \leq O(\sigma\sqrt{\epsilon}) \ .$$

Main Idea: Filter at minimizer of empirical risk.

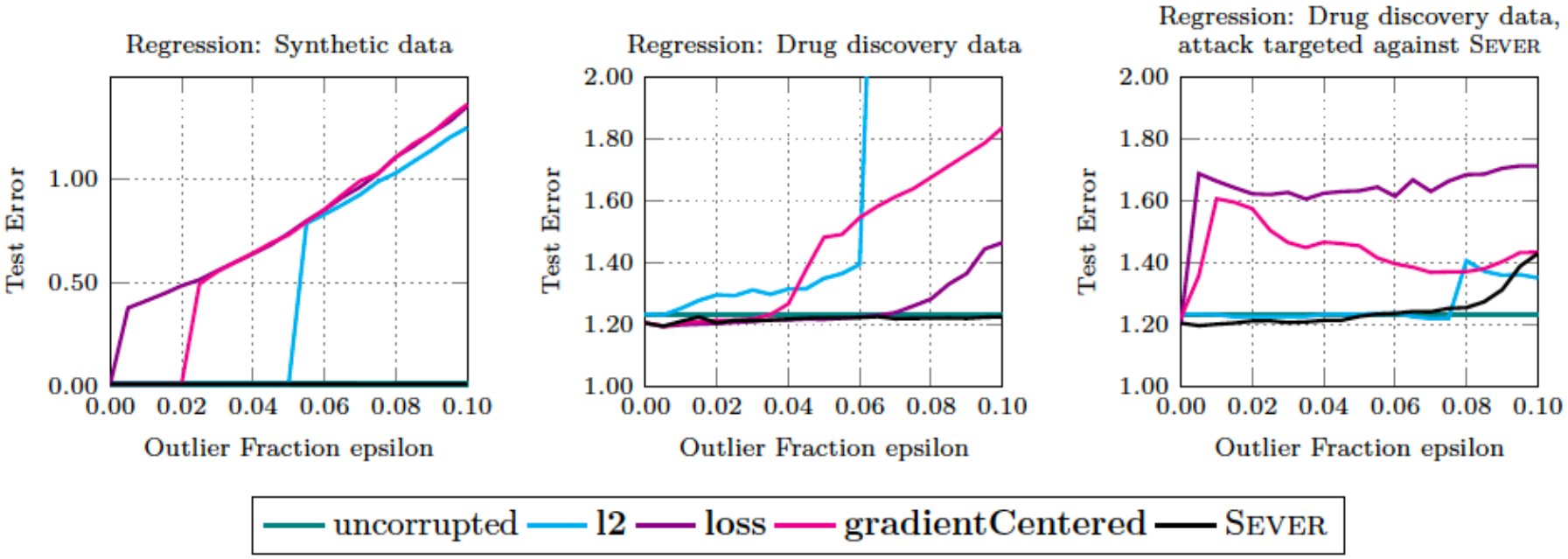
SPECIFIC APPLICATIONS

Corollary: Outlier-robust learning algorithms with dimension-independent error guarantees for:

- SVMs
- Linear Regression
- Logistic Regression
- GLMs
- Experimental Performance Against Data Poisoning Attacks.

Concurrent works obtained tighter guarantees in terms of either sample complexity or error, by focusing on specific tasks and distributional assumptions [Klivans-Kothari-Meka'18, Diakonikolas-Kong-Stewart'18, …].

EXPERIMENTS: RIDGE REGRESSION



OUTLINE

Part III

- General Framework for Robust Mean Estimation
- Robust Stochastic Optimization
- **Learning with Majority of Outliers**

Learning with A Majority of Outliers

- So far focused on setting where $\epsilon < 1/2$.
- What can we learn from a dataset in which the ***majority*** of points are corrupted?

Problem: Given a set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ and $0 < \alpha \leq 1/2$ such that:

- An unknown subset of αN points are drawn from an unknown $D \in \mathcal{F}$, and
- The remaining $(1-\alpha)N$ points are arbitrary,

approximate the mean μ of D.

Which is the "real" D?

List-Decodable Learning

- Return *several hypotheses* with the guarantee that at least one is close.

List-Decodable Mean Estimation:
Given a set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ and $0 < \alpha \leq 1/2$ such that:
- An unknown subset of αN points are drawn from an unknown $D \in \mathcal{F}$, and
- The remaining $(1-\alpha)N$ points are arbitrary,

output a small list of s hypotheses vectors such that one is close to the mean μ of D.

- Model defined in [Balcan-Blum-Vempala'08]
- First studied for mean estimation [Charikar-Steinhardt-Valiant'17]
- Application: Learning Mixture Models

List-Decodable Mean Estimation

Theorem [Charikar-Steinhardt-Valiant'17] Let $0 < \alpha \leq 1/2$. If D has covariance $\Sigma \preceq I$ there is an efficient algorithm that uses $N \geq d/\alpha$ corrupted points, and outputs a list of $s = O(1/\alpha)$ vectors $\widehat{\mu}_1, \ldots, \widehat{\mu}_s$ such that with high probability

$$\min_i \|\widehat{\mu}_i - \mu\|_2 = \tilde{O}(1/\sqrt{\alpha}) \ .$$

Theorem [Diakonikolas-Kane-Stewart'18] *Any* list-decodable mean estimator for bounded covariance distributions must have error $\Omega(1/\sqrt{\alpha})$ as long as the list size is any function of α.

- Initial algorithm [CSV'17] based on ellipsoid method.
- Generalization of filtering ("multi-filtering") works for list-decodable setting [DKS'18].

Part IV: Computational-Statistical Tradeoffs & Future Directions

OUTLINE

Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

OUTLINE

Computational Limits to Robust Estimation

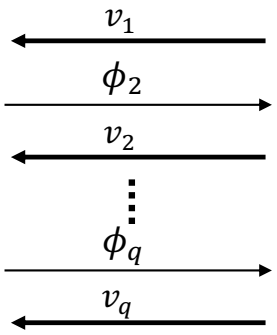
- **Statistical Query Learning Model**
- Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

STATISTICAL QUERIES [KEARNS'93]

$x_1, x_2, \ldots, x_m \sim D$ over X

STATISTICAL QUERIES [KEARNS'93]

SQ algorithm



$\mathrm{STAT}_D(\tau)$ oracle

$$\phi_1 : X \to [-1,1] \qquad |v_1 - \mathbf{E}_{x\sim D}[\phi_1(x)]| \le \tau$$

τ is tolerance of the query; $\tau = 1/\sqrt{m}$

Problem $P \in \mathrm{SQCompl}(q, m)$:
If exists a SQ algorithm that solves P using q queries to $\mathrm{STAT}_D(\tau = 1/\sqrt{m})$

POWER OF SQ LEARNING ALGORITHMS

- **Restricted Model**: Hope to prove unconditional computational lower bounds.
- **Powerful Model**: Wide range of algorithmic techniques in ML are implementable using SQs*:
- PAC Learning: AC^0, decision trees, linear separators, boosting.
- Unsupervised Learning: stochastic convex optimization, moment-based methods, *k*-means clustering, EM, …
  [Feldman-Grigorescu-Reyzin-Vempala-Xiao/JACM'17]
- **Only known exception**: Gaussian elimination over finite fields (e.g., learning parities).
- For all problems in this talk, strongest known algorithms are SQ.

OUTLINE

Computational Limits to Robust Estimation

- Statistical Query Learning Model
- **Results**
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

GENERIC SQ LOWER BOUND CONSTRUCTION

General Technique for SQ Lower Bounds:
Leads to Tight Lower Bounds
for a range of High-dimensional Estimation Tasks

Concrete Applications of Technique:

- Robustly Learning Mean and Covariance
- Learning Gaussian Mixture Models (GMMs)
- Statistical-Computational Tradeoffs (e.g., sparsity)
- Robustly Testing a Gaussian

SQ Lower Bound for Robust Mean Estimation

Theorem: Suppose $d \geq \mathrm{polylog}(1/\epsilon)$. Any SQ algorithm that learns an ϵ - corrupted Gaussian $\mathcal{N}(\mu, I)$ in the strong contamination model within error

$$O(\epsilon\sqrt{\log(1/\epsilon)}/M)$$

requires either:

- SQ queries of accuracy $d^{-M/6}$

or

- At least $d^{\Omega(M^{1/2})}$ many SQ queries.

Take-away: Any asymptotic improvement in error guarantee over filtering algorithm requires super-polynomial time.

SQ Lower Bounds for Learning *Separated* GMMs

Theorem: Suppose that $d \geq \mathrm{poly}(k)$. Any SQ algorithm that learns *separated* k-GMMs over $\mathbb{R}^d$ to constant error requires either:

- SQ queries of accuracy $d^{-k/6}$

or

- At least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

Take-away: Computational complexity of learning GMMs is inherently exponential in **number of components**.

Applications: Concrete SQ Lower Bounds

Learning Problem	Upper Bound	SQ Lower Bound
Robust Gaussian Mean Estimation	Error: $O(\epsilon \log^{1/2}(1/\epsilon))$ [DKKLMS'16]	Runtime Lower Bound: $d^{\mathrm{poly}(M)}$ for factor M improvement in error.
Robust Gaussian Covariance Estimation	Error: $O(\epsilon \log(1/\epsilon))$ [DKKLMS'16]	
Learning k-GMMs (without noise)	Runtime: $d^{g(k)}$ [MV'10, BS'10]	Runtime Lower Bound: $d^{\Omega(k)}$
Robust k-Sparse Mean Estimation	Sample size: $O(k^2 \log d)$ [BDLS'17]	If sample size is $O(k^{1.99})$ runtime lower bound: $d^{k^{\Omega(1)}}$
Robust Covariance Estimation in Spectral Norm	Sample size: $\tilde{O}(d^2)$ [DKKLMS'16]	If sample size is $O(d^{1.99})$ runtime lower bound: $2^{d^{\Omega(1)}}$

OUTLINE

Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- **Generic Lower Bound Technique**
- Applications: Robust Mean Estimation & Learning GMMs

General Recipe For SQ Lower Bounds

- **Step #1:** Construct distribution $\mathbf{P}_v$ that is standard Gaussian in all directions except v.
- **Step #2:** Construct the univariate projection in the v direction so that it matches the first m moments of $\mathcal{N}(0,1)$
- **Step #3:** Consider the family of instances $\mathcal{D} = \{\mathbf{P}_v\}_v$

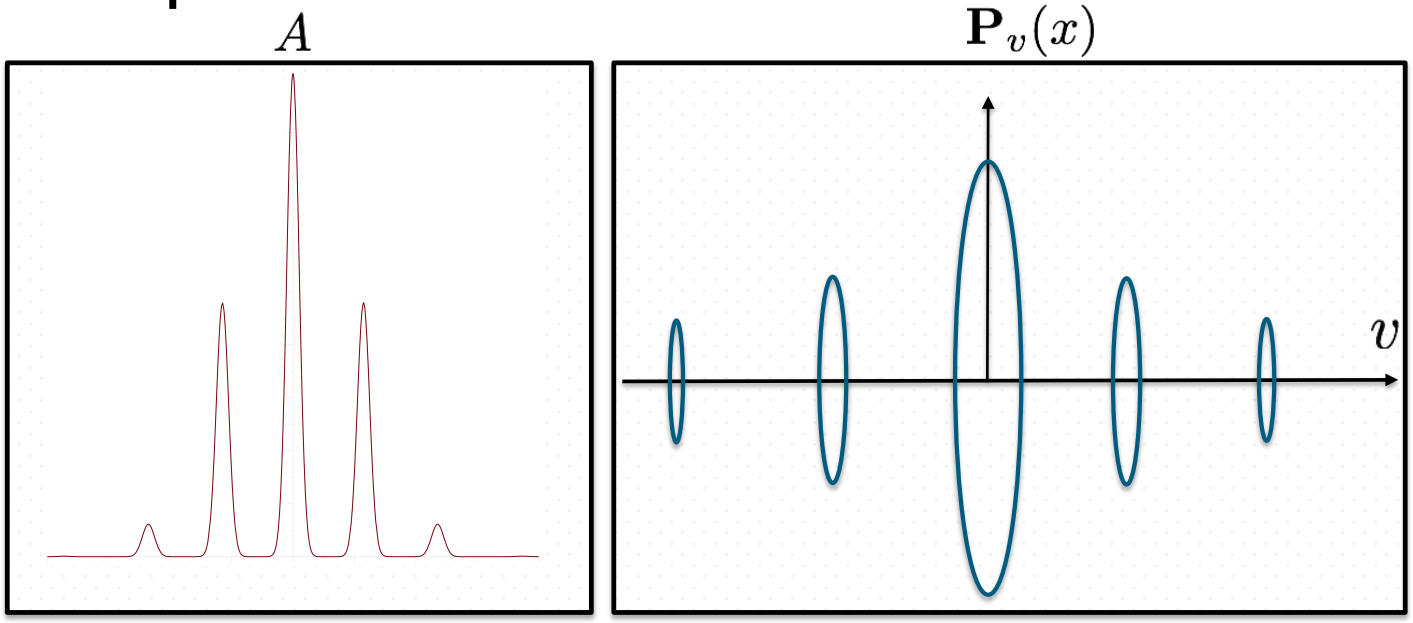
Non-Gaussian Component Analysis [Blanchard *et al.* 2006]

Hidden Direction Distribution

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Example:



Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Why is Finding a Hidden Direction Hard

Observation: Low-Degree Moments do not help.

- A matches the first m moments of $\mathcal{N}(0,1)$
- The first m moments of $\mathbf{P}_v$ are identical to those of $\mathcal{N}(0,I)$
- Degree-$(m+1)$ moment tensor has $\Omega(d^m)$ entries.

Claim: Random projections do not help.

- To distinguish between $\mathbf{P}_v$ and $\mathcal{N}(0,I)$, would need exponentially many random projections.

Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within δ error requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

OUTLINE

Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- **Applications: Robust Mean Estimation & Learning GMMs**

SQ Lower Bound for Robust Mean Estimation (I)

Want to show:

Theorem: Any SQ algorithm that learns an ϵ- corrupted Gaussian in the strong contamination model within error $\epsilon\sqrt{\log(1/\epsilon)/M}$ requires either SQ queries of accuracy $d^{-M/6}$ or at least $d^{\Omega(M^{1/2})}$ many SQ queries.

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ Lower Bound for Robust Mean Estimation (II)

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Lemma: There exists a univariate distribution A that is ϵ - close to $\mathcal{N}(\mu, 1)$ such that:

- A agrees with $\mathcal{N}(0,1)$ on the first M moments.
- We have that $\mu = \Omega(\epsilon\sqrt{\log(1/\epsilon)/M^2})$
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) = \Omega(\mu)$.

SQ Lower Bound for Learning GMMs (I)

Want to show:

Theorem: Any SQ algorithm that learns separated k-GMMs over $\mathbb{R}^d$ to constant error requires either SQ queries of accuracy $d^{-k/6}$ or at least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ Lower Bound for Learning GMMs (II)

Proposition: Suppose that:

- *A* matches the first *m* moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

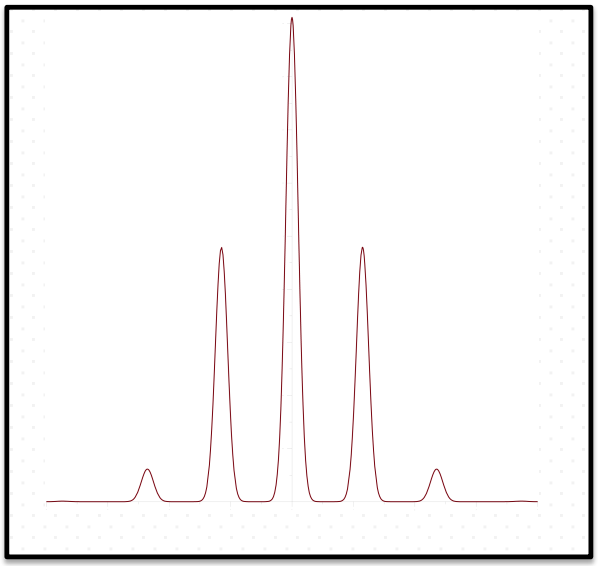
- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

SQ Lower Bound for Learning GMMs (III)

Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

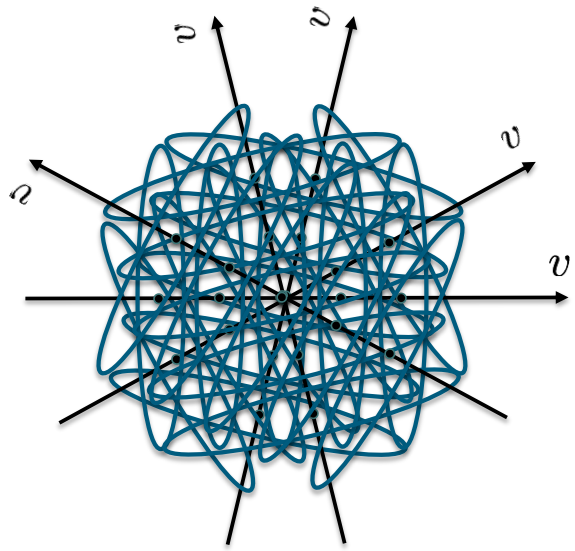
- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

A



SQ Lower Bound for Learning GMMs (IV)

High-Dimensional Distributions $\mathbf{P}_v$ look like "parallel pancakes":



Efficiently learnable for k=2. [Brubaker-Vempala'08]

Future Directions: Computational Lower Bounds

Robustness can make high-dimensional estimation harder computationally and information-theoretically.

Subsequent Work & Future Directions:

- Further Applications of this Framework
  - List-Decodable Mean Estimation [D-Kane-Stewart, STOC'18]
  - Robust Regression [D-Kong-Stewart, SODA'19]
  - Adversarial Examples [Bubeck-Price- Razenshteyn, ICML'19]
  - Agnostic Learning of LTFs/ReLUs [D-Kane-Zarifis, NeurIPS'20, D-Kane-Pitttas-Zarifis, COLT'21]
  - List-Decodable Linear Regression [D-Kane-Pensia-Pittas'21]

- Alternative Evidence of Computational Hardness?
  - ❖ Reductions from Average-Case Problems [Brennan-Bresler, COLT'20]
  - ❖ Reductions from Worst-case Problems? [Hopkins-Li, COLT'19; Bruna-Regev-Song-Tang, STOC'21]
  - ❖ SoS Lower Bounds ?

Future Directions: Algorithms

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

NEAR-LINEAR TIME ALGORITHMS

Filtering for robust mean estimation is practical, but runtime is *super-linear* $\tilde{\Theta}(Nd^2)$.

Question: Can we design near-linear time algorithms?

- Robust Mean Estimation
  - [Cheng-D-Ge, SODA'19; Depersin-Lecue, 2019; Dong-Hopkins-Li, NeurIPS'19, …]

- How about more general estimation tasks?
  - ❖ Robust Covariance Estimation
    - ❖ [Cheng-D-Ge-Woodruff, COLT'19; Li-Ye, NeurIPS'20]
  - ❖ List-Decodable Learning / Learning Mixture Models
    - ❖ [Cherapanamjeri-Mohanty-Yau, FOCS'20; D-Kane-Koongsgard, NeurIPS'20, D-Kane-Koongsgard-Li-Tian'20, D-Kane-Koongsgard-Li-Tian'21]
  - ❖ Robust *Sparse* Estimation?

Non-Convex Optimization & Robust Statistics

Question: Can we design robust estimators using first-order methods?

- Robust Mean Estimation
  - [Cheng-D-Ge-Soltanolkotabi, ICML'20; Zhu et al. '20]
- Robust Sparse Estimation
  - [Cheng-D-Ge-Gupta-Kane-Soltanolkotabi'21]
- How about more general estimation tasks?

Broader Research Directions

General Algorithmic Theory of Robustness

How can we robustly learn rich representations of data, based on natural hypotheses about the structure in data?

Broader Challenges:

- Richer Families of Problems and Models
- Connections to RL, Adversarial Examples, GANs, …
- Relation to Related Notions of Algorithmic Stability
  (Differential Privacy, Adaptive Data Analysis)
- Further Applications (ML Security, Computer Vision, …)
- Other notions of robustness?
  See, e.g., [D-Gouleakis-Tzamos'19]

Thank you!
Questions?

Related Materials (I)

- **Simons Institute Bootcamp Tutorial:**

  www.iliasdiakonikolas.org/simons-tutorial-robust.html

- **TTI-Chicago Summer Workshop Program**

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

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

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

Related Materials (II)

- **STOC'19 Tutorial** (June 2019)
  - http://www.iliasdiakonikolas.org/stoc-robust-tutorial.html
- **ISIT'19 Tutorial (**July 2019**)**
  - http://www.iliasdiakonikolas.org/isit-robust-tutorial.html
- **ICML'20 Tutorial** (July 2020)
  - http://www.iliasdiakonikolas.org/icml-robust-tutorial.html
- **Survey article:**
  - https://arxiv.org/abs/1911.05911