Algorithmic High-Dimensional Robust Statistics

Ilias Diakonikolas (USC)

Simons Institute, UC Berkeley
August 2018

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 (e.g., biology). Hard to detect in several cases [Rosenberg *et al.*, Science'02; Li *et al.*, Science'08; Paschou *et al*., Journal of Medical Genetics'10]

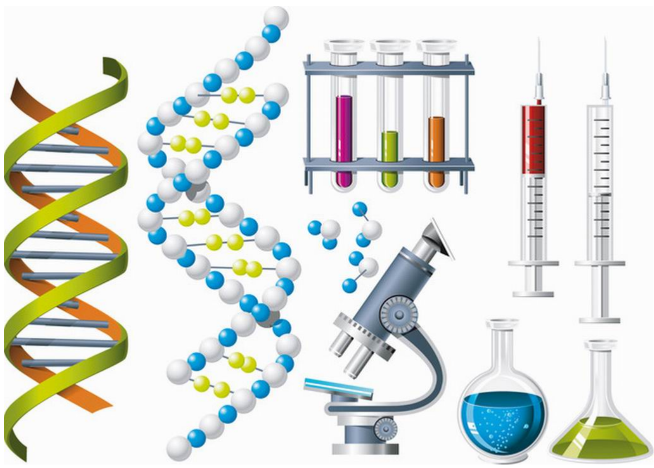
- **Reliable/Adversarial/Secure ML**: Data poisoning attacks (e.g., crowdsourcing) [Biggio *et al.* ICML'12, …]

DETECTING OUTLIERS IN REAL DATASETS

- High-dimensional datasets tend to be inherently noisy.

Biological Datasets: POPRES project,
HGDP datasets

[November *et al.*, Nature'08];
[Rosenberg *et al.*, Science'02];
[Li *et al.*, Science'08];
[Paschou *et al.*, Medical Genetics'10]

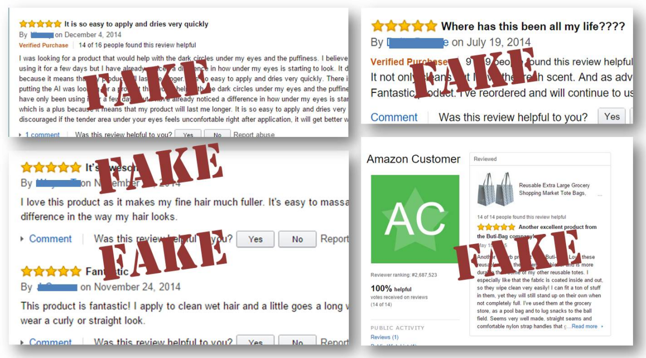


- Outliers: either interesting or can contaminate statistical analysis

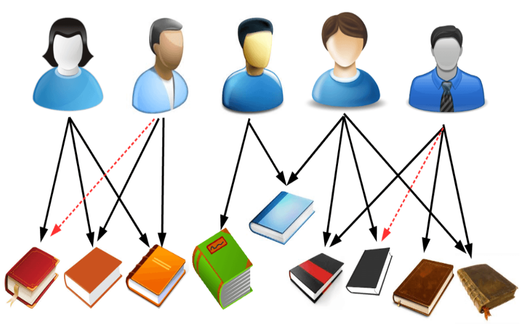
Data Poisoning

Fake Reviews [Mayzlin et al. ‘14]

So Many Misleading, “Fake” Reviews



Recommender Systems:



[Li et al. ‘16]

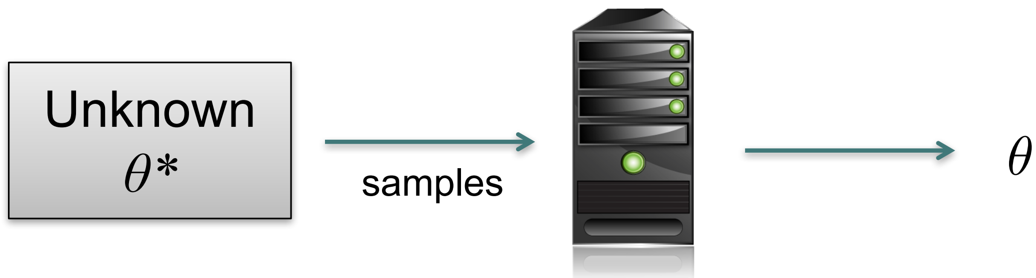
Crowdsourcing:

[Wang et al. ‘14]

Malware/spam:

[Nelson et al. ‘08]

The Statistical Learning Problem



- *Input*: sample generated by a **probabilistic 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?

Robustness in a Generative Model

Contamination Model:

Let $\mathcal{F}$ be a family of probabilistic 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]

MODELS OF ROBUSTNESS

- Oblivious/Adaptive Adversary
- Adversary can: add corrupted samples, subtract uncorrupted samples or both.
- Six Distinct Models:

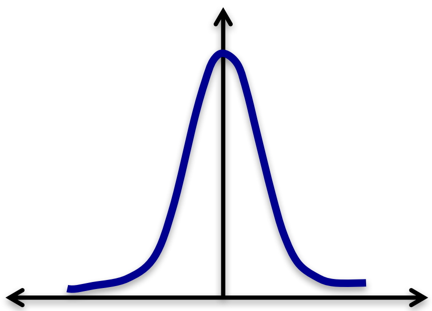
	Oblivious	Adaptive
Additive Errors	Huber's Contamination Model $P = (1-\epsilon)G + \epsilon B$	Additive Contamination ("Data Poisoning")
Subtractive Errors	$P = (1-\epsilon)G - \epsilon L$	Subtractive Contamination
Additive and Subtractive Errors	Hampel's Contamination $\mathrm{d}_{\mathrm{TV}}(P, G) \leq \epsilon$ $P = G - \epsilon L + \epsilon B$	Strong Contamination ("Nasty Learning Model")

Example: Parameter Estimation

Given 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:

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

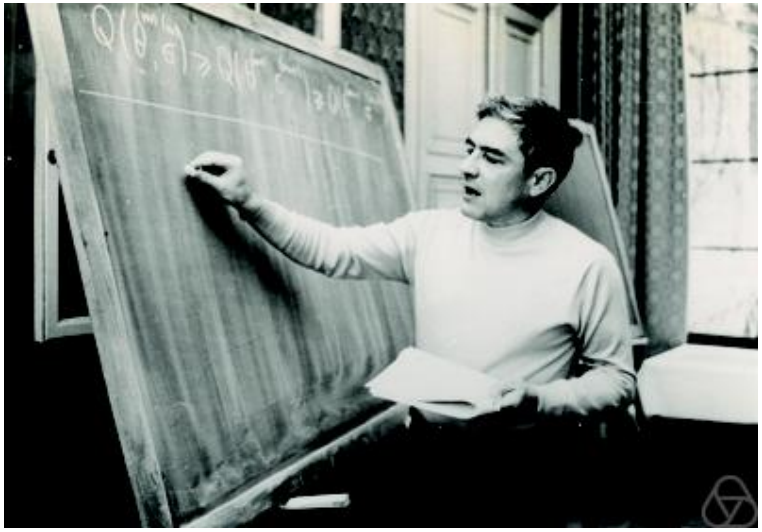
R. A. Fisher

The **maximum likelihood estimator** is asymptotically efficient (1910-1920)

J. W. Tukey

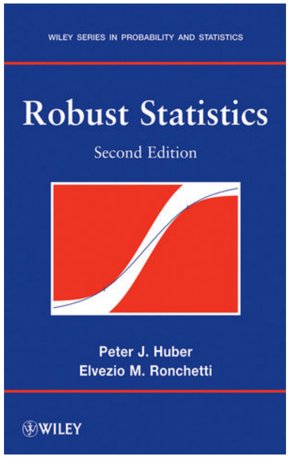
What about **errors** in the model itself? (1960)

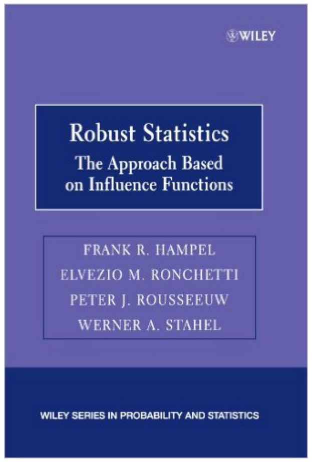
Peter J. Huber

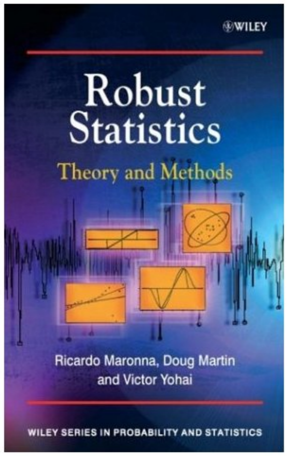


"Robust Estimation of a Location Parameter"
Annals of Mathematical Statistics, 1964.

Robust Statistics







What estimators behave well in a **neighborhood** around the model?

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?

Gaussian 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

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) \ .$$

Remark: Optimal rate of convergence with N samples is

$$O(\epsilon) + O\left(\sqrt{d/N}\right)$$

[Tukey'75, Donoho'82]

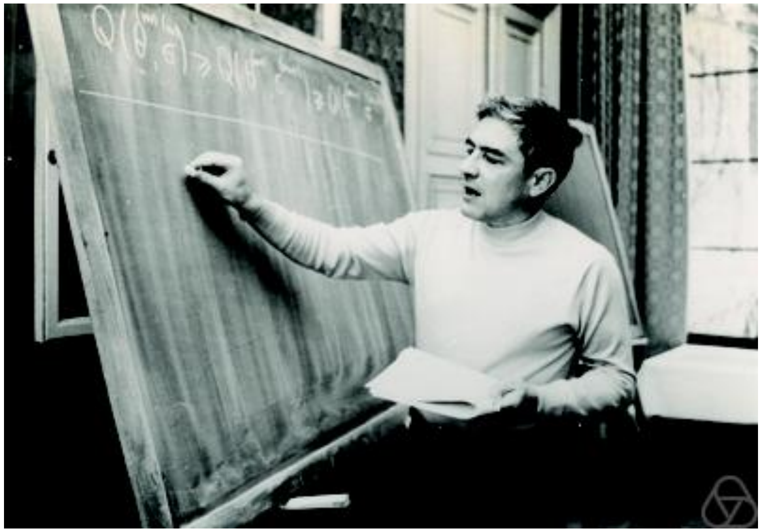
Previous Approaches: Robust Mean Estimation

Unknown Mean	Error Guarantee		Running Time	
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	$\Theta(\epsilon)$	✓	$N^{O(d)}$	✗

All known estimators are either **hard to compute** or can tolerate a **negligible fraction of corruptions**.

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.*

THIS TALK

Robust estimation in high-dimensions is algorithmically possible!

- First 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

This Talk: Robust Gaussian Mean Estimation

Theorem: There are polynomial time algorithms with the following behavior: Given $\epsilon > 0$ and a set of $N = \widetilde{O}(d/\epsilon^2)$ ϵ- corrupted samples from a d-dimensional Gaussian $\mathcal{N}(\mu, I)$, the algorithms find $\widehat{\mu} \in \mathbb{R}^d$ that with high probability satisfies:

- **[LRV'16]:**

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

  in *additive** contamination model.

- **[DKKLMS'16]:**

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

  in *strong* contamination model.

* Can be adapted to give error $O(\epsilon\sqrt{\log(1/\epsilon)}\sqrt{\log d})$ in strong contamination model as well.

OUTLINE

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- Extensions

Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

Outline

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- **Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal**
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- Extensions

Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

HIGH-DIMENSIONAL GAUSSIAN MEAN ESTIMATION (I)

Fact: Let $X_1, \ldots, X_N$ be IID samples from $\mathcal{N}(\mu, I)$. The empirical estimator $\widehat{\mu}$ satisfies $\|\widehat{\mu} - \mu\|_2 \leq \delta$ with probability at least 9/10 for $N = \Omega(d/\delta^2)$. Moreover, *any* estimator with this guarantee requires $\Omega(d/\delta^2)$ samples.

Proof:

By definition, $\widehat{\mu} = (1/N) \sum_{i=1}^{N} X_i$, where $X_i \sim \mathcal{N}(\mu, I)$.

Then,

$$\widehat{\mu} \sim \mathcal{N}\left(\mu, (1/N)I\right).$$

We have

$$\mathbf{E}[\|\widehat{\mu} - \mu\|_2^2] = \sum_{j=1}^{d} \mathbf{E}\left[(\widehat{\mu}_j - \mu_j)^2\right] = \sum_{j=1}^{d} \mathbf{Var}\left[\widehat{\mu}_j\right] = d/N$$

Therefore,

$$\mathbf{E}[\|\widehat{\mu} - \mu\|_2] \leq \mathbf{E}[\|\widehat{\mu} - \mu\|_2^2]^{1/2} = \sqrt{\frac{d}{N}}$$

and Markov's inequality gives the upper bound.

HIGH-DIMENSIONAL GAUSSIAN MEAN ESTIMATION (II)

Fact: Let $X_1, \ldots, X_N$ be IID samples from $\mathcal{N}(\mu, I)$. The empirical estimator $\widehat{\mu}$ satisfies $\|\widehat{\mu} - \mu\|_2 \leq \delta$ with probability at least 9/10 for $N = \Omega(d/\delta^2)$. Moreover, *any* estimator with this guarantee requires $\Omega(d/\delta^2)$ samples.

Proof:

For the lower bound, consider the following family of distributions:

$$\{\mathcal{N}(\mu, I)\}_{\mu \in \mathcal{M}}$$

where

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d : \mu_j = -\delta/\sqrt{d} \text{ or } \mu_j = \delta/\sqrt{d}, j \in [d]\right\} \; .$$

Apply Assouad's lemma to show that learning an unknown distribution in this family within error $\delta/2$ requires $\Omega(d/\delta^2)$ samples.

■

Information-Theoretic Limits On Robust Estimation (I)

Proposition: Any robust mean estimator for $\mathcal{N}(\mu, 1)$ has error $\Omega(\epsilon)$, even in Huber's model.

Claim: Let P_1, P_2 be such that $d_{TV}(P_1, P_2) = \epsilon/(1-\epsilon)$. There exist noise distributions B_1, B_2 such that $(1-\epsilon)P_1 + \epsilon B_1 = (1-\epsilon)P_2 + \epsilon B_2$.

Proof:

Can write

$$P_i = \left(1 - \frac{\epsilon}{1-\epsilon}\right) P + \frac{\epsilon}{1-\epsilon} Q_i$$

Take $B_1 = Q_2$ and $B_2 = Q_1$. In this case,

$$(1-\epsilon)P_1 + \epsilon B_1 = (1-\epsilon)P_2 + \epsilon B_2 = (1-2\epsilon)P + \epsilon(Q_1 + Q_2) .$$

■

Information-Theoretic Limits On Robust Estimation (II)

Proposition: Any robust mean estimator for $\mathcal{N}(\mu, 1)$ has error $\Omega(\epsilon)$, even in Huber's model.

Proof:
Need similar construction where P_1, P_2 are unit variance Gaussians.
Let $P_i = \mathcal{N}(\mu_i, 1)$ such that $\mathrm{d_{TV}}(P_1, P_2) = \epsilon/(1-\epsilon)$.

Since $\mathrm{d_{TV}}\left(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)\right) \leq |\mu_1 - \mu_2|/2$, this implies that

$$|\mu_1 - \mu_2| = \Omega(\epsilon) \ . \qquad \blacksquare$$

Remarks:
- More careful calculation shows that constant in $O\left(.\right)$ is $\sqrt{\pi/2} - o(1)$.
- Under different assumptions on good data, we obtain different functions of ϵ .

SAMPLE EFFICIENT ROBUST MEAN ESTIMATION (I)

Proposition: There is an algorithm that uses $N = O(d/\epsilon^2)$ ϵ- corrupted samples from $\mathcal{N}(\mu, I)$ and outputs $\widetilde{\mu} \in \mathbb{R}^d$ that with probability at least $9/10$ satisfies $\|\widetilde{\mu} - \mu\|_2 = O(\epsilon)$.

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of *all* its 1-dimensional projections (cf. Tukey median).

Basic Fact: $\|\widetilde{\mu} - \mu\|_2 = \max_{v:\|v\|_2=1} |v \cdot \widetilde{\mu} - v \cdot \mu|$

Claim 1: Suppose we can estimate $v \cdot \mu$ for each $v \in \mathbb{R}^d, \|v\|_2 = 1$, i.e., find $\{\widehat{\mu}_v\}_v$ such that for all $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ we have $|\widehat{\mu}_v - \mu \cdot v| \leq \delta$. Then, we can learn μ within error 2δ .

Proof:

Consider *infinite size* LP: Find $x \in \mathbb{R}^d$ such that *for all* $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$: $|\widehat{\mu}_v - v \cdot x| \leq \delta$.

Let x^* be any feasible solution. Then

$$\|x^* - \mu\|_2 = \max_{v:\|v\|_2=1} |v \cdot x^* - v \cdot \mu| \leq \max_{v:\|v\|_2=1} |v \cdot x^* - \widehat{\mu}_v| + \max_{v:\|v\|_2=1} |v \cdot \mu - \widehat{\mu}_v| \leq 2\delta \;. \qquad \blacksquare$$

SAMPLE EFFICIENT ROBUST MEAN ESTIMATION (II)

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of "all" its 1-dimensional projections.

Claim 2: Suffices to consider a γ -net C over all directions, where γ is a small positive constant.

Proof:

This gives the following *finite* LP:

Find $x \in \mathbb{R}^d$ such that for all $v \in C$, we have $|\widehat{\mu}_v - v \cdot x| \leq \delta$.

Let x^* be any feasible solution. Let $u \in C$ such that $\|u - \frac{\mu - x^*}{\|\mu - x^*\|_2}\|_2 \leq \gamma$.

Then

$$\|x^* - \mu\|_2 = \left| \left(\left(\frac{\mu - x^*}{\|\mu - x^*\|_2} - u \right) + u \right) \cdot (x^* - \mu) \right| \leq \gamma \|x^* - \mu\|_2 + 2\delta$$

or

$$\|x^* - \mu\|_2 \leq \frac{2\delta}{1 - \gamma} \ .$$

■

Sample Efficient Robust Mean Estimation (III)

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of "all" its 1-dimensional projections.

So, for $\gamma = 1/2$, any feasible solution to the LP has $\|x^* - \mu\|_2 \leq 4\delta$.

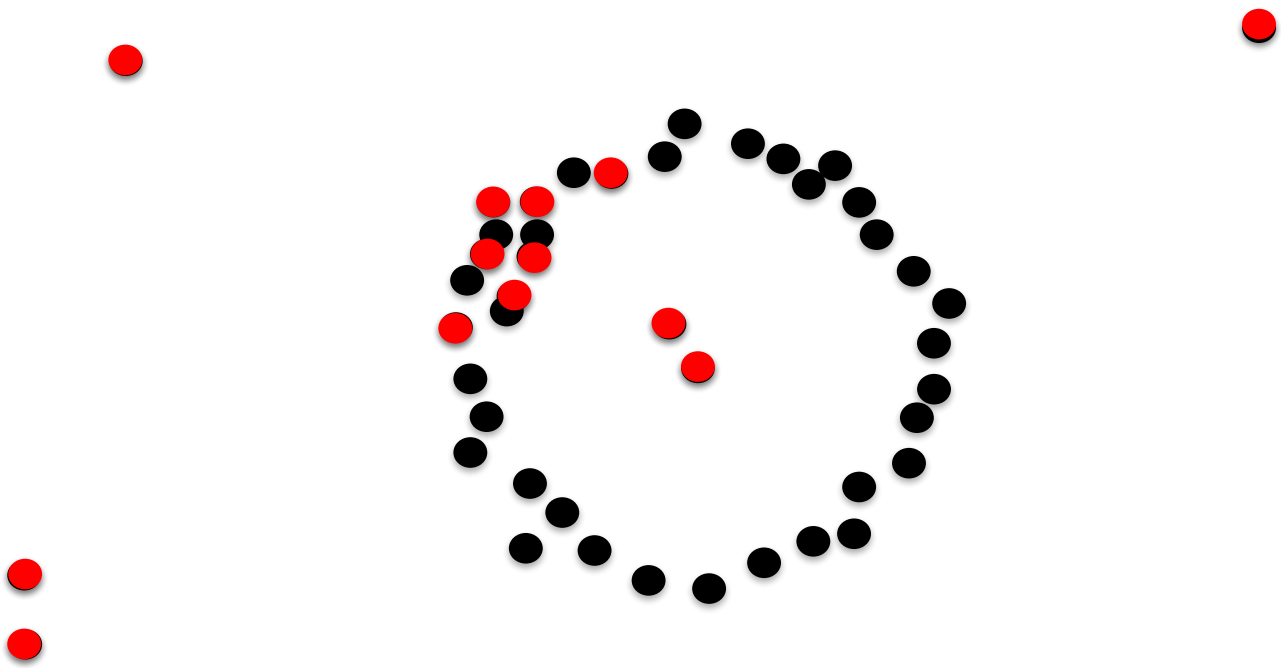
Sample Complexity: Note that the empirical median satisfies $\delta = O(\epsilon)$ with probability at least $1 - \tau$ after $O((1/\epsilon^2)\log(1/\tau))$ samples.

We need union bound over all $v \in C$. Since $|C| = (1/\gamma)^{O(d)} = 2^{O(d)}$, for $\tau = 1/(10|C|)$ our algorithm works with probability at least $9/10$.
Thus, sample complexity will be $N = O(d/\epsilon^2)$.

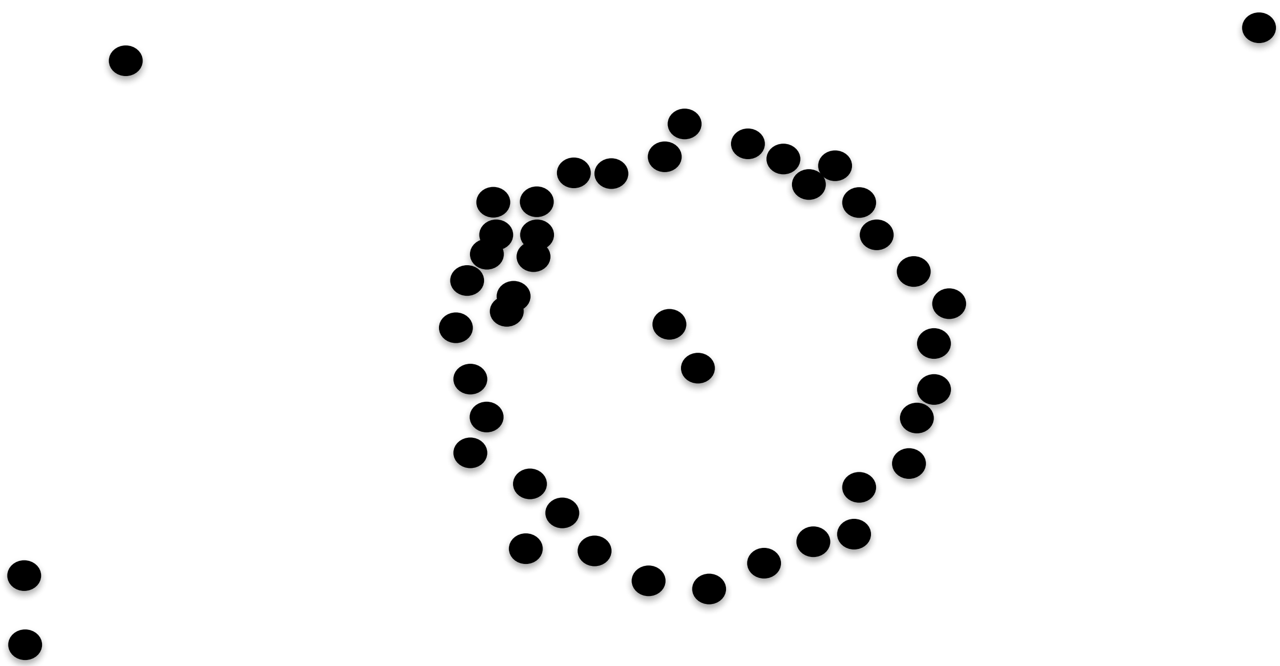
Runtime: $\mathrm{poly}(N, 2^d)$.

■

OUTLIER DETECTION ?



NAÏVE OUTLIER REMOVAL (NAÏVE PRUNING)



Gaussian Annulus Theorem: $\mathbf{Pr}_{X\sim\mathcal{N}(\mu,I)}\left[\left|\|X\|_2^2 - d\right| > t\right] \leq 2e^{-\Omega\left(\min\left\{\frac{t^2}{d},t\right\}\right)}$

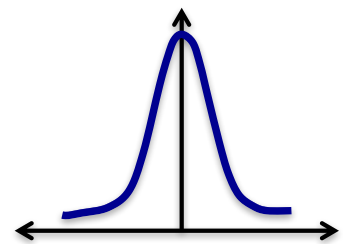
On The Effect of Corruptions

Question: What is the effect of additive and subtractive corruptions?

Let's study the simplest possible example of $\mathcal{N}(\mu, 1)$.

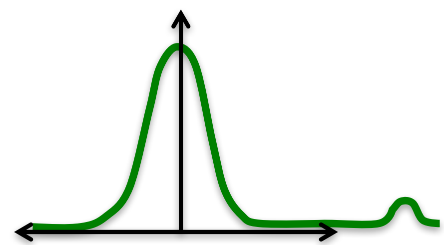
Subtractive errors at rate ϵ can:

- Move the mean by at most $O(\epsilon\sqrt{\log(1/\epsilon)})$
- Increase the variance by $O(\epsilon)$ and decrease it by at most $O(\epsilon \log(1/\epsilon))$



Additive errors at rate ϵ can:

- Move the mean arbitrarily
- Increase the variance arbitrarily and decrease it by at most $O(\epsilon)$



OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation

- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- **Overview of Algorithmic Approaches**
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- Extensions

Part III: Summary and Conclusions

- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

High-Level Goal: Reduce “structured” high-dimensional problem to a collection of “low-dimensional” problems.

Three Approaches: Overview And Comparison

Three Algorithmic Approaches:

- Recursive Dimension-Halving [LRV'16]
- Iterative Filtering [DKKLMS'16]
- Soft Outlier Removal [DKKLMS'16]

Commonalities:

- Rely on Spectrum of Empirical Covariance to Robustly Estimate the Mean
- Certificate of Robustness for the Empirical Estimator

Exploiting the Certificate:

- Recursive Dimension-Halving: Find "good" large subspace.
- Iterative Filtering: Check condition on entire space. If violated, filter outliers.
- Soft Outlier Removal: Convex optimization via approximate separation oracle.

OUTLINE

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- **Certificate of Robustness**
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- Extensions

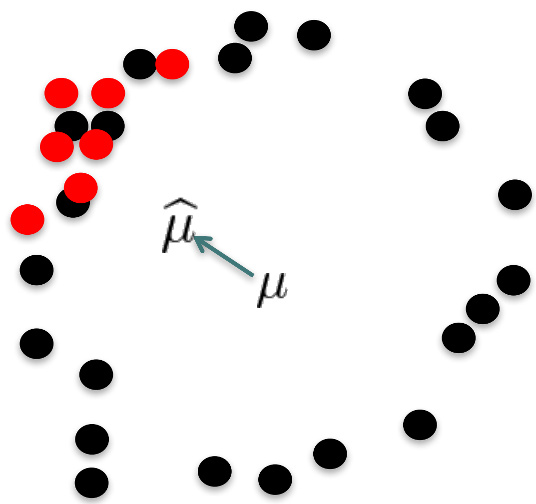
Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

Certificate Of Robustness For Empirical Estimator

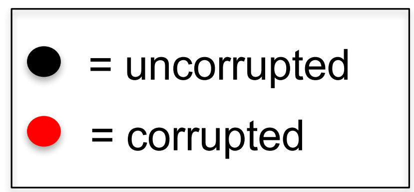
Idea #1 [DKKLMS'16, LRV'16]: If the empirical covariance is "close to what it should be", then the empirical mean works.

Certificate Of Robustness For Empirical Estimator

Detect when the empirical estimator *may* be compromised



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



There is *no* direction of large (> 1) 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

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

with high probability we have:

- **[LRV'16]:**

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

  in ***additive*** contamination model

- **[DKKLMS'16]:**

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

  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

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

with high probability we have:

- **[LRV'16]:**

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

  in ***additive*** contamination model

- **[DKKLMS'16]:**

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

  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:

- **[LRV'16]:**

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

in ***additive*** contamination model

PROOF OF KEY LEMMA: ADDITIVE CORRUPTIONS (I)

Let $S = \{X_1, \ldots, X_N\}$ be a multi-set of additively ϵ-corrupted samples from $\mathcal{N}(\mu, I)$. Can assume wlog that $\mu = \mathbf{0}$.

Note that $S = G \cup B$, where G is the uncorrupted set of samples and B is the set of added corrupted samples.

Express empirical mean and covariance as sum of terms, one depending on G and one on B.

Let $\widehat{\mu}_G = (1/|G|) \cdot \sum_{i \in I_G} X_i$, similarly define $\widehat{\mu}_B$.
We can write

$$\widehat{\mu} = (1-\epsilon)\widehat{\mu}_G + \epsilon\widehat{\mu}_B \ .$$

For simplicity, assume $N \to \infty$. Then have that $\widehat{\mu_G} = \mu = \mathbf{0}$.

Therefore, we obtain:

Claim 1: $\widehat{\mu} = \epsilon\widehat{\mu}_B$.

PROOF OF KEY LEMMA: ADDITIVE CORRUPTIONS (II)

Recall

Assumption: $\mu = \mathbf{0}$ **Claim 1:** $\widehat{\mu} = \epsilon \widehat{\mu}_B$.

Will express $\widehat{\Sigma}$ in similar form. By definition, $\widehat{\Sigma} = (1/N) \sum_{i \in [N]} X_i X_i^T - \widehat{\mu}\widehat{\mu}^T$

Define $\widehat{\Sigma}_G = (1/|G|) \sum_{i \in I_G} X_i X_i^T - \widehat{\mu}_G \widehat{\mu}_G^T$ and similarly $\widehat{\Sigma}_B$.

Since $N \to \infty$, we have $\widehat{\mu}_G = \mu = \mathbf{0}$ and $\widehat{\Sigma}_G = I$.
Will show:

Claim 2: $\widehat{\Sigma} = (1-\epsilon)I + \epsilon \widehat{\Sigma}_B + (\epsilon - \epsilon^2)\widehat{\mu}_B \widehat{\mu}_B^T$.

Proof: Note that

$$(1/N) \sum_{i \in I_G} X_i X_i^T = (1-\epsilon) I \quad \text{and} \quad (1/N) \sum_{i \in I_B} X_i X_i^T = \epsilon \widehat{\Sigma}_B + \epsilon \widehat{\mu}_B \widehat{\mu}_B^T .$$

Putting these together and using Claim 1 gives the claim. ■

PROOF OF KEY LEMMA: ADDITIVE CORRUPTIONS (III)

Recall

Assumption: $\mu = 0$ **Claim 1:** $\widehat{\mu} = \epsilon \widehat{\mu}_B$.
Claim 2: $\widehat{\Sigma} = (1-\epsilon)I + \epsilon \widehat{\Sigma}_B + (\epsilon - \epsilon^2)\widehat{\mu}_B \widehat{\mu}_B^T$.

Can now finish argument. Recall that $\|\widehat{\Sigma}\|_2 = \max_{v:\|v\|_2=1} v^T \widehat{\Sigma} v$.

Note that $v^T \widehat{\Sigma} v = (1-\epsilon) + \epsilon(v^T \widehat{\Sigma}_B v) + (\epsilon - \epsilon^2) v^T (\widehat{\mu}_B \widehat{\mu}_B^T) v$.

Choosing $v = \widehat{\mu}_B / \|\widehat{\mu}_B\|_2$ gives

$$\|\widehat{\Sigma}\|_2 \geq (1-\epsilon) + (\epsilon - \epsilon^2)\|\widehat{\mu}_B\|_2^2 \ .$$

In conclusion, if $\|\widehat{\Sigma}\|_2 \leq 1 + \delta$, then $\|\widehat{\mu}_B\|_2^2 \leq O(1 + \delta/\epsilon)$

Via Claim 1, we have shown the following implication:

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

Choosing $\delta = O(\epsilon)$ gives the lemma.

■

PROOF OF KEY LEMMA: ADDITIVE CORRUPTIONS (IV)

So far assumed we are in infinite sample regime.

Essentially same argument holds in finite sample setting.
The following concentration inequalities suffice:

For $N = \Omega(d/\epsilon^2)$, with high probability we have that

$$\|\mu - \widehat{\mu}_G\|_2 \ll \epsilon$$

and

$$\|\widehat{\Sigma}_G - I\|_2 \ll \epsilon$$

∎

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:

- **[DKKLMS'16]:**

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

in ***strong*** contamination model

Handling Strong Corruptions

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

PROOF OF KEY LEMMA: STRONG CORRUPTIONS (I)

Let $S = \{X_1, \ldots, X_N\}$ be a multi-set of ϵ-corrupted samples from $\mathcal{N}(\mu, I)$. Can assume wlog that $\mu = \mathbf{0}$.

Note that $S = (G \setminus L) \cup B$, where G is the uncorrupted set of samples, B is the added corrupted samples, and $L \subset G$ is the subtracted set of samples.

Will express empirical mean and covariance as sum of three terms, depending on G , B, and L.

Let $\widehat{\mu}_G = (1/|G|) \cdot \sum_{i \in I_G} X_i$. Similarly define $\widehat{\mu}_B$ and $\widehat{\mu}_L$.
We can write

$$\widehat{\mu} = \widehat{\mu}_G - \epsilon\widehat{\mu}_L + \epsilon\widehat{\mu_B} \ .$$

When $N \to \infty$, we have that $\widehat{\mu_G} = \mu = \mathbf{0}$.

Therefore, we obtain

Claim 1: $\widehat{\mu} = \epsilon(\widehat{\mu_B} - \widehat{\mu_L})$.

PROOF OF KEY LEMMA: STRONG CORRUPTIONS (II)

Recall **Assumption:** $\mu = \mathbf{0}$ **Claim 1:** $\widehat{\mu} = \epsilon(\widehat{\mu_B} - \widehat{\mu_L})$.

Will express $\widehat{\Sigma}$ in similar form. By definition, $\widehat{\Sigma} = (1/N)\sum_{i\in[N]} X_i X_i^T - \widehat{\mu}\widehat{\mu}^T$

Define $\widehat{\Sigma}_G = (1/|G|)\sum_{i\in I_G} X_i X_i^T - \widehat{\mu}_G \widehat{\mu}_G^T$, similarly $\widehat{\Sigma}_B$ and $\widehat{M}_L = (1/|L|)\sum_{i\in I_L} X_i X_i^T$.

Since $N \to \infty$, we have $\widehat{\mu}_G = \mu = \mathbf{0}$ and $\widehat{\Sigma}_G = I$.
Will show:

Claim 2: $\widehat{\Sigma} = I + \epsilon\widehat{\Sigma}_B + \epsilon\widehat{\mu}_B\widehat{\mu}_B^T - \epsilon\widehat{M}_L - \epsilon^2(\widehat{\mu}_B - \widehat{\mu}_L)(\widehat{\mu}_B - \widehat{\mu}_L)^T$.

Proof: Note that

$$(1/N)\sum_{i\in I_G} X_i X_i^T = I \,,\ (1/N)\sum_{I\in I_B} X_i X_i^T = \epsilon\widehat{\Sigma}_B + \epsilon\widehat{\mu}_B\widehat{\mu}_B^T \ \text{ and } \ (1/N)\sum_{I\in I_L} X_i X_i^T = \epsilon\widehat{M}_L$$

Putting these together and using Claim 1 gives the claim. ∎

PROOF OF KEY LEMMA: STRONG CORRUPTIONS (III)

Recall

> **Assumption:** $\mu = 0$ **Claim 1:** $\widehat{\mu} = \epsilon(\widehat{\mu_B} - \widehat{\mu_L})$.
>
> **Claim 2:** $\widehat{\Sigma} = I + \epsilon\widehat{\Sigma}_B + \epsilon\widehat{\mu}_B\widehat{\mu}_B^T - \epsilon\widehat{M}_L - \epsilon^2(\widehat{\mu}_B - \widehat{\mu}_L)(\widehat{\mu}_B - \widehat{\mu}_L)^T$.

To finish argument, need to bound $\widehat{M}_L$ and $\widehat{\mu}_L$.

> **Claim 3**: Have $\|\widehat{M}_L\|_2 = O(\log(1/\epsilon))$ and $\|\widehat{\mu}_L\|_2 = O(\sqrt{\log(1/\epsilon)})$.

Assuming the claim holds, we get

$$\widehat{\Sigma} = I + \epsilon\widehat{\Sigma}_B + (\epsilon - \epsilon^2)\widehat{\mu}_B\widehat{\mu}_B^T + O(\epsilon\log(1/\epsilon)) .$$

This gives

$$\|\widehat{\Sigma}\|_2 \geq 1 + (\epsilon - \epsilon^2)\|\widehat{\mu}_B\|_2^2 - O(\epsilon\log(1/\epsilon)) .$$

PROOF OF KEY LEMMA: STRONG CORRUPTIONS (IV)

We can now finish the argument.
We have shown that

$$\|\widehat{\Sigma}\|_2 \geq 1 + (\epsilon - \epsilon^2)\|\widehat{\mu}_B\|_2^2 - O(\epsilon \log(1/\epsilon)) \ .$$

Suppose that $\|\widehat{\Sigma}\|_2 \leq 1 + \delta$. Then

$$\|\widehat{\mu}_B\|_2 \leq O\left(\sqrt{\delta/\epsilon} + \sqrt{\log(1/\epsilon)}\right)$$

Since $\widehat{\mu} = \epsilon(\widehat{\mu}_B - \widehat{\mu}_L)$, the final error is

$$\begin{aligned}\|\widehat{\mu}\|_2 &\leq \epsilon\|\widehat{\mu}_B\|_2 + \epsilon\|\widehat{\mu}_L\|_2 \\ &\leq O\left(\sqrt{\delta\epsilon} + \epsilon\sqrt{\log(1/\epsilon)}\right) \ .\end{aligned}$$

For $\delta = \Theta(\epsilon \log(1/\epsilon))$, the lemma follows.

∎

Proof of Key Lemma: Strong Corruptions (V)

Recall that $\widehat{M}_L := (1/|L|) \sum_{i \in I_L} X_i X_i^T = \mathbf{E}_{X \sim_U L}[X X^T]$. Remains to prove:

Claim 3: We have $\|\widehat{M}_L\|_2 = O(\log(1/\epsilon))$ and $\|\widehat{\mu}_L\|_2 = O(\sqrt{\log(1/\epsilon)})$.

Proof: By definition have $\|\widehat{M}_L\|_2 = \max_{v:\|v\|_2=1} |v^T \widehat{M}_L v| = \max_{v:\|v\|_2=1} \mathbf{E}_{X \sim_U L}[(v \cdot X)^2]$.

Since $L \subset G$, for any event, $|L| \cdot \mathbf{Pr}_{X \sim_U L}[X \in \mathcal{E}] \leq |S| \cdot \mathbf{Pr}_{X \sim_U G}[X \in \mathcal{E}]$.

For any unit vector v:

$$
\begin{aligned}
\mathbf{E}_{X \sim_U L}[(v \cdot X)^2] &= 2 \int_0^{O(\sqrt{d})} \mathbf{Pr}_{X \sim_U L}[|v \cdot X| > T] T dT \\
&\leq 2 \int_0^{O(\sqrt{d})} \min\left\{1, (1/\epsilon) \cdot \mathbf{Pr}_{X \sim_U G}[|v \cdot X| > T]\right\} T dT \\
&\leq 2 \int_0^{O(\sqrt{\log(1/\epsilon)})} T dT + (1/\epsilon) \cdot \int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d})} e^{-T^2/2} T dT \\
&= O(\log(1/\epsilon)) + O(1) \; .
\end{aligned}
$$

Finally, by definition we have that $\|\widehat{\mu}_L\|_2^2 \leq \|\widehat{M}_L\|_2$.

■

OUTLINE

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- **Recursive Dimension Halving**
- Iterative Filtering, Soft Outlier Removal
- Extensions

Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

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

RECURSIVE DIMENSION-HALVING [LRV'16]

Recursive Procedure:

Step #1: Find large subspace where "standard" estimator works.
Step #2: Recurse on complement.
(If dimension is small, use brute-force.)

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.

Proof of Good Subspace Lemma (I)

Let $S = \{X_1, \ldots, X_N\}$ be a multi-set of additively ϵ -corrupted samples from $\mathcal{N}(\mu, I)$. Can assume wlog that $\mu = \mathbf{0}$.

Note that $S = G \cup B$, where G is the uncorrupted set of samples and B is the added corrupted samples. Let S' be the subset of S obtained after naïve pruning. We know that $S' = G \cup B'$, where $B' \subseteq B$, and each $x \in S'$ satisfies $\|x\|_2 = O(\sqrt{d})$.

Let $\widehat{\Sigma}_{S'} = (1/|S'|) \sum_{i \in I_{S'}} X_i X_i^T - \widehat{\mu}_{S'} \widehat{\mu}_{S'}^T$ be the empirical covariance of S' and $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ be its spectrum.

Want to show that $\lambda_{d/2} \leq 1 + O(\epsilon)$.

This follows from the following claims:

Claim 1: $\lambda_1 \geq 1 - O(\epsilon)$.

Claim 2: $\mathrm{Tr}(\widehat{\Sigma}_{S'}) \leq d(1 + O(\epsilon))$.

Proof of Good Subspace Lemma (II)

Let $\widehat{\Sigma}_{S'} = (1/|S'|) \sum_{i \in I_{S'}} X_i X_i^T - \widehat{\mu}_{S'} \widehat{\mu}_{S'}^T$ be the empirical covariance of S' and $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ be its spectrum.

Claim 1: $\lambda_1 \geq 1 - O(\epsilon)$.

Claim 2: $\mathrm{Tr}(\widehat{\Sigma}_{S'}) \leq d(1 + O(\epsilon))$.

By Claim 1, $A = \sum_{i=1}^{d/2} \lambda_i \geq (d/2)(1 - O(\epsilon))$

Moreover, $B = \sum_{i=d/2+1}^{d} \lambda_i \geq (d/2)\lambda_{d/2}$

By Claim 2, $A + B \leq d(1 + O(\epsilon))$

Therefore, $B \leq (d/2)(1 + O(\epsilon))$

which gives $\lambda_{d/2} \leq 1 + O(\epsilon)$. ∎

PROOF OF GOOD SUBSPACE LEMMA (III)

Let $\widehat{\Sigma}_{S'} = (1/|S'|) \sum_{i \in I_{S'}} X_i X_i^T - \widehat{\mu}_{S'} \widehat{\mu}_{S'}^T$ be the empirical covariance of S' and $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ be its spectrum.

Claim 1: $\lambda_1 \geq 1 - O(\epsilon)$.

Proof: Recall that $S' = G \cup B'$, where G is the uncorrupted set of samples and B' is a subset of the added corrupted samples. Therefore,

$$\widehat{\Sigma}_{S'} = (1-\epsilon)I + \epsilon \widehat{\Sigma}_{B'} + (\epsilon - \epsilon^2) \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T$$

Denoting $M = \epsilon \widehat{\Sigma}_{B'} + (\epsilon - \epsilon') \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T$, we have that

$$\lambda_{\min}(\widehat{\Sigma}_{S'}) \geq (1-\epsilon) + \min_{v : \|v\|_2 = 1} v^T M v \geq 1 - \epsilon \, .$$

■

Proof of Good Subspace Lemma (IV)

Let $\widehat{\Sigma}_{S'} = (1/|S'|) \sum_{i \in I_{S'}} X_i X_i^T - \widehat{\mu}_{S'} \widehat{\mu}_{S'}^T$ be the empirical covariance of S' and $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ be its spectrum.

Claim 2: $\mathrm{Tr}(\widehat{\Sigma}_{S'}) \leq d(1 + O(\epsilon))$.

Proof: Recall that

$$\widehat{\Sigma}_{S'} = (1-\epsilon)I + \epsilon \widehat{\Sigma}_{B'} + (\epsilon - \epsilon^2) \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T$$

Thus,

$$\mathrm{Tr}(\widehat{\Sigma}_{S'}) \leq d(1-\epsilon) + \epsilon \mathrm{Tr}(\widehat{\Sigma}_{B'} + \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T)$$

Note that

$$\widehat{\Sigma}_{B'} + \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T = (1/|B'|) \sum_{i \in I_{B'}} X_i X_i^T$$

Moreover, for every $x \in B' \subseteq S'$ we have $\|x\|_2 = O(\sqrt{d})$.
Thus,

$$\mathrm{Tr}(\widehat{\Sigma}_{B'} + \widehat{\mu}_{B'} \widehat{\mu}_{B'}^T) = O(d) \ .$$

■

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 I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation

- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- **Iterative Filtering, Soft Outlier Removal**
- Extensions

Part III: Summary and Conclusions

- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

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

ITERATIVE FILTERING [DKKLMS'16]

Iterative Two-Step Procedure:

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

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

Iterate on "cleaner" dataset.

General recipe that works for fairly general settings.

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

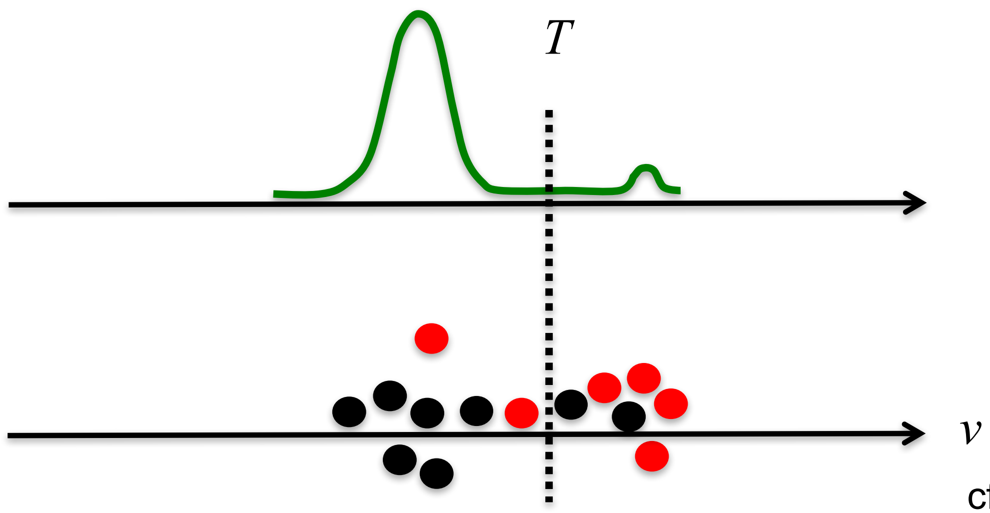
Filtering Subroutine

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.



cf. [Klivans-Long-Servedio'09, Lai-Rao-Vempala'16]

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

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

- Throw away all points x such that

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

- 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 1: In each iteration, we remove more corrupted than uncorrupted points.

After a number of iterations, we have removed all corrupted points.

Eventually the empirical mean works

Filtering Subroutine: 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

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

5. **Return**

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

SKETCH OF CORRECTNESS (I)

Claim 2: Can always find a threshold satisfying the Condition of Step 4.
Proof:
By contradiction. Suppose that for all $T > 0$ we have

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

Will use this to show that $\lambda^* = \|\widehat{\Sigma}_S\|_2$ is smaller than it was assumed to be.

Since the median is a robust estimator of the mean, it follows that for all $T > 0$

$$\mathbf{Pr}_{X\sim_U S}[|v^* \cdot X - \mu| > T + 2] < 8 \cdot e^{-T^2/2}.$$

Since $B \subset S$, for any event $\mathcal{E}$, $|B| \cdot \mathbf{Pr}_{X\sim_U B}[X \in \mathcal{E}] \leq |S| \cdot \mathbf{Pr}_{X\sim_U S}[X \in \mathcal{E}]$

Therefore,

$$\mathbf{Pr}_{X\sim_U B}[|v^* \cdot (X - \mu)| > T] \leq (1/\epsilon) \cdot \mathbf{Pr}_{X\sim_U S}[|v^* \cdot (X - \mu)| > T]$$

SKETCH OF CORRECTNESS (II)

Assume wlog $\mu = 0$. Recall that

$$\widehat{\Sigma} = I + \epsilon\widehat{\Sigma}_B + (\epsilon - \epsilon^2)\widehat{\mu}_B\widehat{\mu}_B^T + O(\epsilon\log(1/\epsilon)) \ .$$

So, it suffices to show that $\widehat{M}_B := \widehat{\Sigma}_B + \widehat{\mu}_B\widehat{\mu}_B^T = \mathbf{E}_{X\sim_U B}[XX^T]$ has small v^*-variance, i.e., that $\mathbf{E}_{X\sim_U B}[(v^*\cdot X)^2]$ is small.
We have

$$\begin{aligned}
\mathbf{E}_{X\sim_U B}[(v^*\cdot X)^2] &= 2\int_0^{O(\sqrt{d})} \mathbf{Pr}_{X\sim_U B}[|v^*\cdot X| > T]TdT \\
&\leq O(1) + 2\int_2^{O(\sqrt{d})} \mathbf{Pr}_{X\sim_U B}[|v^*\cdot X| > T]TdT \\
&\leq O(1) + 2\int_2^{O(\sqrt{d})} \min\left\{1, (1/\epsilon)\cdot \mathbf{Pr}_{X\sim_U S}[|v^*\cdot X| > T]\right\} TdT \\
&\leq O(1) + 2\int_2^{O(\sqrt{\log(1/\epsilon)})} TdT + 16\int_{O(\sqrt{\log(1/\epsilon)})}^{O(\sqrt{d})} Te^{-(T-2)^2/2}dT \\
&= O(\log(1/\epsilon)) + O(1) \ . \qquad \blacksquare
\end{aligned}$$

SUMMARY: ROBUST MEAN ESTIMATION VIA FILTERING

Certificate of Robustness:

"Spectral norm of empirical covariance is what it should be."

Exploiting the Certificate:

- Check if certificate is satisfied.
- If violated, find "subspace" where behavior of outliers different than behavior of inliers.
- Use it to detect and remove outliers.
- Iterate on "cleaner" dataset.

Soft Outlier Removal

Let

$$S_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : 0 \leq w_i \leq \frac{1}{(1-2\epsilon)N} \right\}$$

Let $\delta = \Theta(\epsilon \log(1/\epsilon))$. Consider the convex set

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\epsilon} : \left\| \sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \delta \right\}$$

Algorithm:

- Find $w^* \in \mathcal{C}_\delta$
- Output $\widehat{\mu}_{w^*} = \sum_{i=1}^{N} w_i^* X_i$.

Main Issue: μ unknown.

Soft Outlier Removal

Let

$$S_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : 0 \leq w_i \leq \frac{1}{(1-2\epsilon)N} \right\}$$

Let $\delta = \Theta(\epsilon \log(1/\epsilon))$. Consider the convex set

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\epsilon} : \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \delta \right\}$$

Algorithm:

- Find $w^* \in \mathcal{C}_\delta$
- Output $\widehat{\mu}_{w^*} = \sum_{i=1}^N w_i^* X_i$.

- Adaptation of key lemma gives: For all $w \in \mathcal{C}_\delta$, we have:

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

APPROXIMATE SEPARATION ORACLE

Input: ϵ -corrupted set S and weight vector w

Output: Separation oracle for $\mathcal{C}_\delta$

- Let $\delta = \Theta(\epsilon \log(1/\epsilon))$
- Let $\widehat{\mu}_w = \sum_{i=1}^N w_i X_i$ and $\widehat{\Sigma}_w = \sum_{i=1}^N w_i X_i X_i^T - \widehat{\mu}_w \widehat{\mu}_w^T$
- Let (λ^*, v^*) be the top eigenvalue-eigenvector pair of $\widehat{\Sigma}_w$.
- If $\lambda^* \leq 1 + \delta$, return "YES".
- Otherwise, return the hyperplane $L : \mathbb{R}^d \to \mathbb{R}$ with

$$L(u) = \sum_{i=1}^N u_i((X_i - \widehat{\mu}_w) \cdot v^*)^2 - \lambda^* .$$

Deterministic Regularity Conditions

Convex program only requires the following conditions:

- For all $w \in S_{N,\epsilon}$, the following hold:

$$\left\| \sum_{i \in I_G} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \delta_1 := \Theta(\epsilon \log(1/\epsilon))$$

$$\left\| \sum_{i \in I_G} w_i (X_i - \mu) \right\|_2 \leq \delta_2 := \Theta(\epsilon \sqrt{\log(1/\epsilon)})$$

OUTLINE

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- **Extensions**

Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

ROBUST MEAN ESTIMATION: *SUB*-GAUSSIAN CASE

What we have *really* shown:

Theorem [DKKLMS, ICML'17]: There is a polynomial time algorithm with the following behavior: Given $\epsilon > 0$ and a set of $N = \widetilde{O}(d/\epsilon^2)$ ϵ- corrupted samples from a d-dimensional **sub-Gaussian distribution with identity covariance**, the algorithm finds $\widehat{\mu} \in \mathbb{R}^d$ that with high probability satisfies:

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

in *strong* contamination model.

Information-theoretically optimal error, even in one-dimension.

Optimal Gaussian Robust Mean Estimation?

Recall [DKKLMS'16]: There is a $\mathrm{poly}(d/\epsilon)$ time algorithm for robustly learning $\mathcal{N}(\mu, I)$ within error

$$O\left(\epsilon\sqrt{\log(1/\epsilon)}\right) .$$

(Open) Question: Is there a $\mathrm{poly}(d/\epsilon)$ time algorithm for robustly learning $\mathcal{N}(\mu, I)$ within error $O(\epsilon)$?

How about

$$o(\epsilon\sqrt{\log(1/\epsilon)}) \ ?$$

GAUSSIAN ROBUST MEAN ESTIMATION: ADDITIVE ERRORS

Theorem [DKKLMS, SODA'18] There is a polynomial time algorithm with the following behavior: Given $\epsilon > 0$ and $N = \mathrm{poly}(d/\epsilon)$ corrupted samples from an unknown mean, identity covariance Gaussian distribution on $\mathbb{R}^d$, the algorithm finds a hypothesis mean $\widehat{\mu}$ that satisfies

$$\|\mu - \widehat{\mu}\|_2 \leq \sqrt{\pi} \cdot \epsilon + o(\epsilon)$$

in ***additive*** contamination model.

- Robustness guarantee optimal up to $\sqrt{2}$ factor!
- For any univariate projection, mean robustly estimated by median.

Generalized Filtering: Additive Corruptions

- *Univariate* filtering based on tails not sufficient to remove the incurred $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ error, even for additive errors.

- **Generalized Filtering Idea**: Filter using *top - k eigenvectors* of empirical covariance.

- **Key Observation**: Suppose that $\|\mu - \widehat{\mu}\|_2 \geq \epsilon$. Then either

(1) $\widehat{\Sigma}$ has k eigenvalues at least $1 + \Omega(\epsilon)$, or

(2) The error comes from a k-dimensional subspace.

- Choose $k = \Theta(\log(1/\epsilon))$.

COMPUTATIONAL LIMITATIONS TO ROBUST MEAN ESTIMATION

Theorem [DKS, FOCS'17] Suppose $d \geq \mathrm{polylog}(1/\epsilon)$. Any *Statistical Query** algorithm that learns an ϵ- corrupted Gaussian $\mathcal{N}(\mu, I)$ in the ***strong*** contamination model within distance

$$o(\epsilon\sqrt{\log(1/\epsilon)})$$

requires runtime

$$d^{\omega(1)} .$$

*Instead of accessing samples from distribution D, a Statistical Query algorithm can adaptively query $\mathbb{E}_{x\sim D}[f(x)]$, for any $f : \mathbb{R}^d \rightarrow [0, 1]$

Take-away: Any asymptotic improvement in error guarantee over [DKKLMS'16] algorithms may require super-polynomial time.

Power of SQ 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, …

***Only known exception**: Gaussian elimination over finite fields (e.g., learning parities).

- For all problems in this talk, strongest known algorithms are SQ.

Methodology For SQ Lower Bounds

- **Statistical Query Dimension**:
- Fixed-distribution PAC Learning
  [Blum-Furst-Jackson-Kearns-Mansour-Rudich'95; …]
- General Statistical Problems
  [Feldman-Grigorescu-Reyzin-Vempala-Xiao'13, …, Feldman'16]
- Pairwise correlation between D_1 and D_2 with respect to D:

$$\chi_D(D_1, D_2) := \int_{\mathbb{R}^d} D_1(x) D_2(x) / D(x) dx - 1$$

- **Fact**: Suffices to construct a large set of distributions that are *nearly* uncorrelated.

GENERIC LOWER BOUND CONSTRUCTION

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

Theorem [DKS, FOCS'17] : For a unit vector v and a univariate distribution with density A, let $\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}$.

Any SQ algorithm that finds the hidden direction v 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 $\mathcal{N}(0,I)$, need *exponentially many* random projections.

Proof uses *Ornstein-Uhlenbeck* (Gaussian noise) operator.

Further Applications of Generic Construction

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 (no corruptions)	Runtime: $d^{g(k)}$ [MV'10, BS'10]	Runtime Lower Bound: $d^{\Omega(k)}$
Robust k-Sparse Mean Estimation	Sample size: $\tilde{O}(k^2 \log d)$ [Li'17, DBS'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: $2^{d^{\Omega(1)}}$

Robust Mean Estimation: General Case

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate mean μ of D.

Theorem [DKKLMS-ICML'17, CSV-ITCS'18] If $N = \Omega(d/\epsilon)$, and D has covariance $\Sigma \preceq \sigma^2 \cdot I$, then we can efficiently recover $\widehat{\mu}$ with ,

$$\|\widehat{\mu} - \mu\|_2 = O(\sigma \cdot \sqrt{\epsilon}) \ .$$

- Sample-optimal, even without corruptions.
- Information-theoretically optimal error, even in one-dimension.
- Adaptation of Iterative Filtering.

Robust Covariance Estimation

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate covariance Σ of D.

Theorem: Let $\epsilon < 1/2$. If $N = \Omega(d^2/\epsilon^2)$, then can efficiently recover $\widehat{\Sigma}$ such that

$$\|\Sigma^{-1/2}(\widehat{\Sigma} - \Sigma)\Sigma^{-1/2}\|_F = f(\epsilon) ,$$

where f depends on the concentration of D.

Main Idea: Use *fourth-order moment tensors*

OUTLINE

Part I: Introduction
- Motivation
- Robust Statistics in Low and High Dimensions
- This Talk

Part II: High-Dimensional Robust Mean Estimation
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering, Soft Outlier Removal
- Extensions

Part III: Summary and Conclusions
- Beyond Robust Statistics: Unsupervised and Supervised Learning
- Conclusions & Future Directions

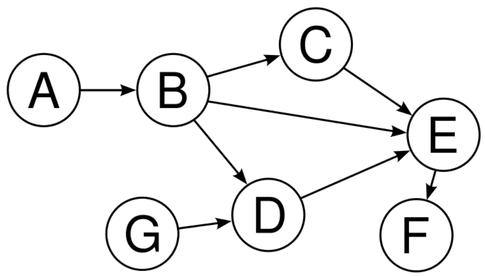
Summary and Conclusions

- High-Dimensional Computationally Efficient Robust Estimation is Possible!

- First Computationally Efficient Robust Estimators with **Dimension-Independent** Error Guarantees.

- General Methodologies for High-Dimensional Estimation Problems.

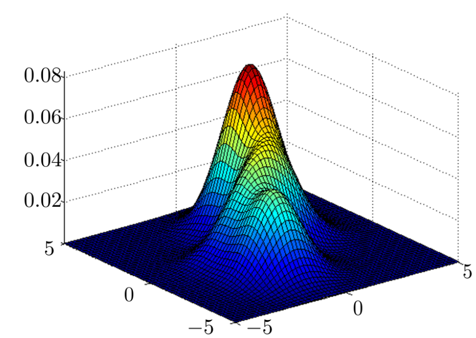
BEYOND ROBUST STATISTICS: ROBUST *UNSUPERVISED* LEARNING



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

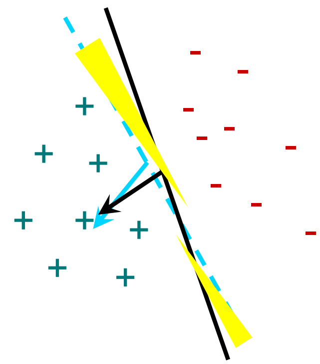
Computational/Statistical-Robustness Tradeoffs
[**D-**Kane-Stewart'17, **D**-Kong-Stewart'18]

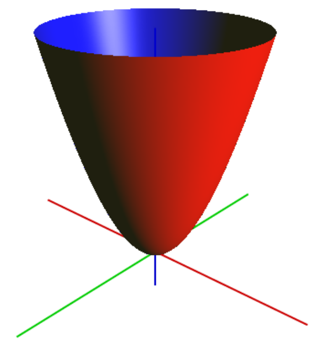
Clustering in Mixture Models
[Charikar-Steinhardt-Valiant'17,
D-Kane-Stewart'18,
Hopkins-Li'18,
Kothari-Steinhardt-Steurer'18]

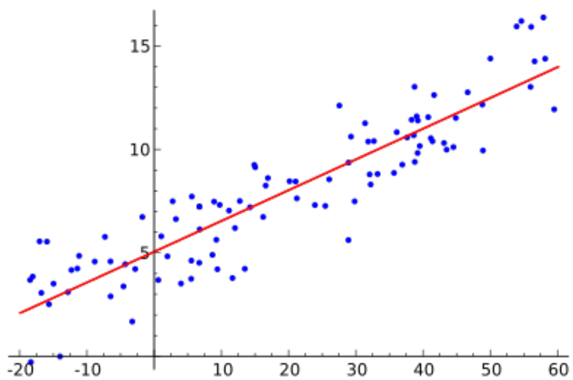


BEYOND ROBUST STATISTICS: ROBUST *SUPERVISED* LEARNING

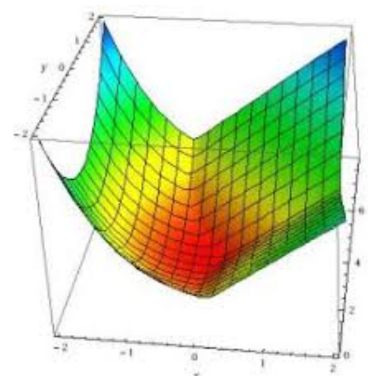
Malicious PAC Learning
[Klivans-Long-Servedio'10,
Awasthi-Balcan-Long'14,
D-Kane-Stewart'18]







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

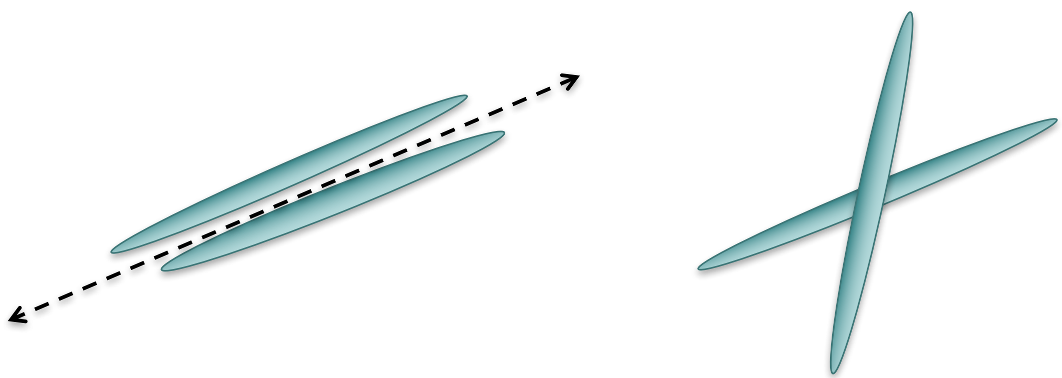
- Graphical Models [Cheng-D-Kane-Stewart'16, D-Kane-Stewart'18]
- Sparse models (e.g., sparse PCA, sparse regression) [Li'17, Du-Balakrishan-Singh'17, Liu-Shen-Li-Caramanis'18, …]
- List-Decodable Learning [Charikar-Steinhardt-Valiant '17, Meister-Valiant'18, D-Kane-Stewart'18]
- Robust PAC Learning [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]
- …

OPEN QUESTIONS

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

Concrete Open Problem:

Robustly Learn a Mixture of 2 *Arbitrary* Gaussians



Future Directions

General Algorithmic Theory of Robustness

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

Can we robustly *test* our hypotheses about structure in data before learning?

Broader Challenges:

- Richer Families of Problems and Models
- Connections to Non-convex Optimization, Adversarial Examples, GANs, …
- Relation to Related Notions of Algorithmic Stability
  (Differential Privacy, Adaptive Data Analysis)
- Practical / Near-Linear Time Algorithms?
  [D-Kamath-Kane-Moitra-Lee-Stewart, ICML'17] [D-KKL-Steinhardt-S'18]
  [Cheng-D-Ge'18]
- Further Applications (ML Security, Computer Vision, …)
- Other models of robustness?

Thank you!
Questions?

Related Workshops:

- **TTI-Chicago Summer Workshop Program**

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

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

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

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

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