Recent Advances in Algorithmic High-Dimensional Robust Statistics

Ilias Diakonikolas (USC) and Daniel Kane (UCSD)

STOC 2019 Tutorial
June 2019

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

Outline Of This Tutorial (I)

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

Outline Of This Tutorial (II)

Part III: Extensions – Beyond Robust Mean Estimation

- Robust Covariance Estimation
- Robust Sparse Estimation Tasks
- List-Decodable Learning
- Robust Estimation of Higher Moments

Part IV: Computational Limits and Future Directions

- Computational—Statistical Tradeoffs in Robust Learning
- Open Problems and Research Directions

OUTLINE

Part I: Introduction

- **Motivation**
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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]

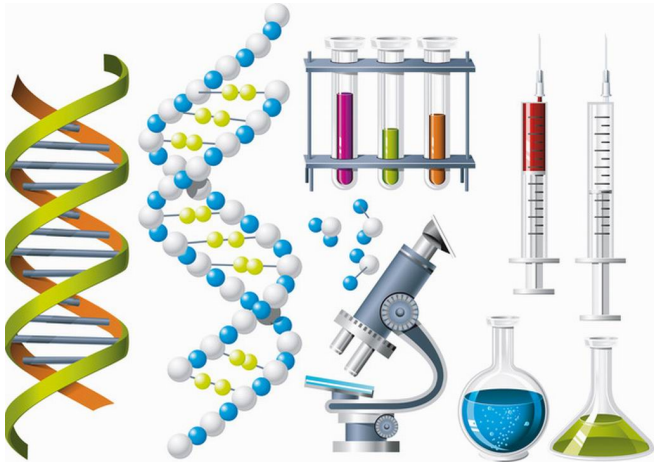
- **Reliable/Adversarial/Secure ML**: Data poisoning attacks (e.g., crowdsourcing) [Biggio *et al.* ICML'12, …]

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]

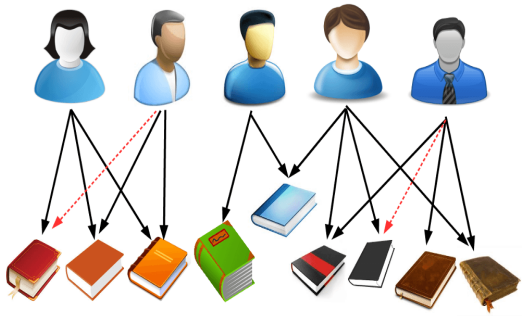


- Outliers: either interesting or can contaminate statistical analysis

DATA POISONING

Fake Reviews [Mayzlin et al. ‘14]

Recommender Systems:



[Li et al. ‘16]

Crowdsourcing:

[Wang et al. ‘14]

Malware/spam:

[Nelson et al. ‘08]

OUTLINE

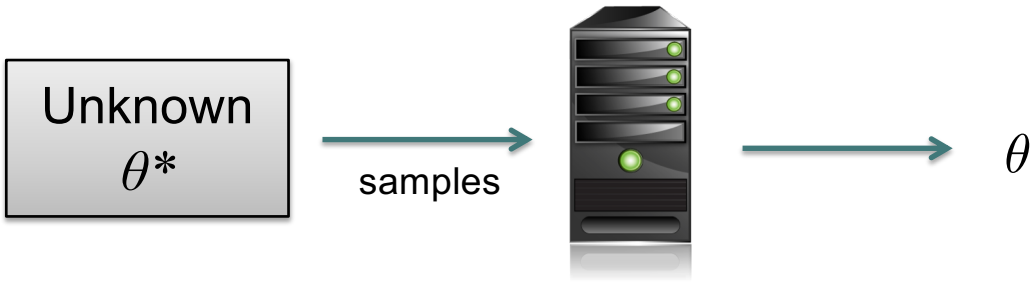
Part I: Introduction

- Motivation
- **Robust Statistics in Low and High Dimensions**

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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?

(Outlier-) 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.

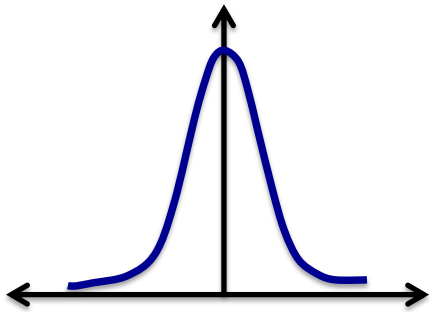
cf. Huber's contamination model [1964]

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:

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

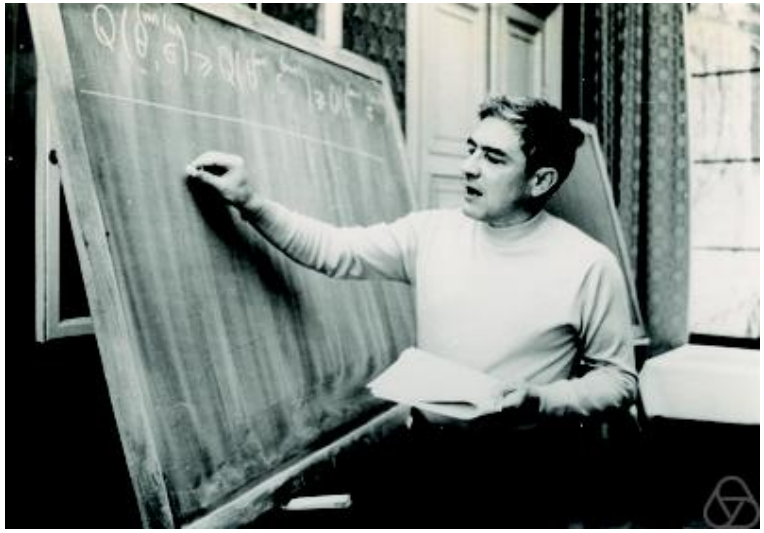
R. A. Fisher

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

J. W. Tukey

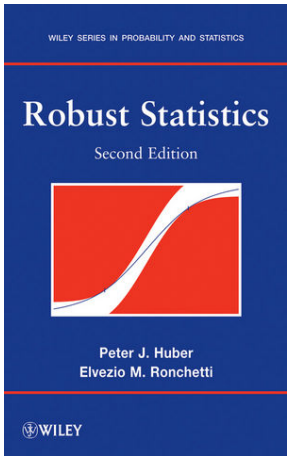
What about **errors** in the model itself? (1960)

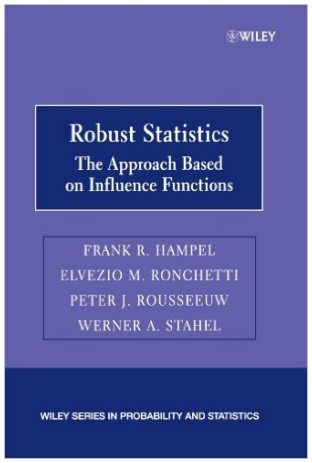
Peter J. Huber

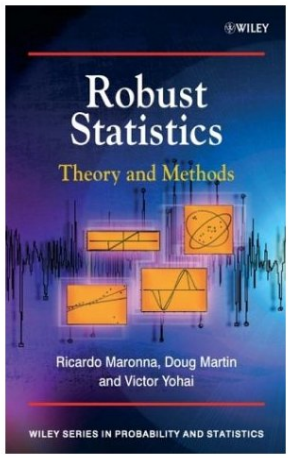


“Robust Estimation of a Location Parameter”
Annals of Mathematical Statistics, 1964.

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?

High-Dimensional 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]

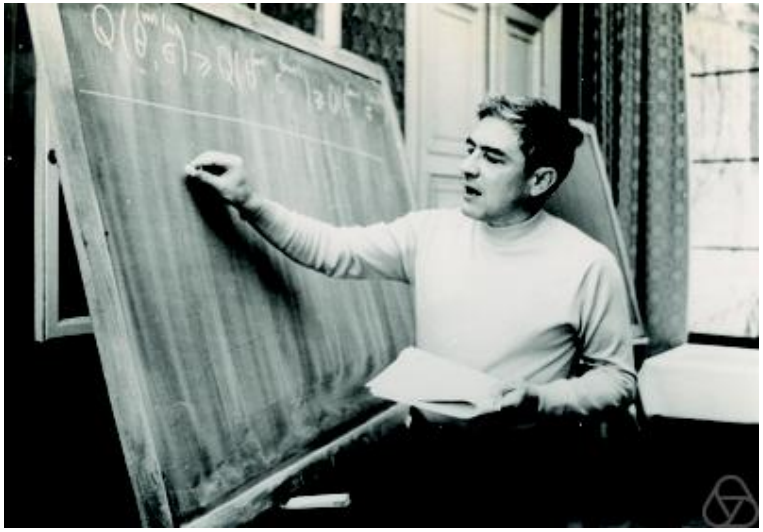
Previous Approaches: Robust Mean Estimation

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.*

Robust estimation in high-dimensions is algorithmically possible!

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

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

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

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

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

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

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

- Mean and Covariance Estimation
- Independent Component Analysis, SVD

Basic Result: Robust 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.

OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- **This Talk: Statements of Algorithmic Results**
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

Robust Mean Estimation: Gaussian 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: Let $\epsilon < 1/2$. If $N = \Omega(d/\epsilon^2)$ and $D = \mathcal{N}(\mu, I)$, then can efficiently recover $\widehat{\mu}$ with

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

in **additive contamination** model.

Error Guarantee Independent of d !

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

Robust Mean Estimation: *Sub*-Gaussian 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: Let $\epsilon < 1/2$. If $N = \Omega(d/\epsilon^2)$ and D is *sub-Gaussian* with identity covariance, then can efficiently recover $\widehat{\mu}$ with

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

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

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

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: Let $\epsilon < 1/2$. If $N = \Omega(d/\epsilon)$, and D has covariance $\Sigma \preceq \sigma^2 \cdot I$, then 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.

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

OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- **Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal**
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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

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

- Use $\mathrm{d_{TV}}\left(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)\right) \leq |\mu_1 - \mu_2|/2$
- 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 $\|x^* - \mu\|_2 \leq 4\delta$.

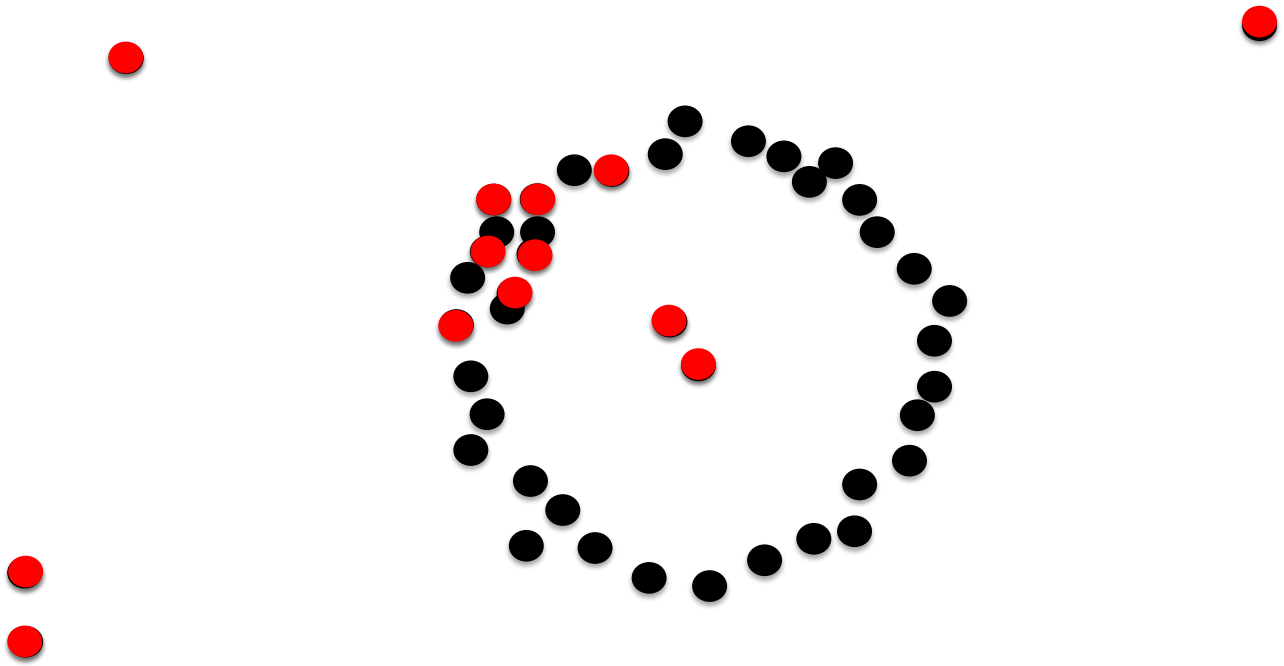
Sample Complexity: Note that the empirical median satisfies $\delta = O(\epsilon)$ with probability at least $1 - \tau$ after $O((1/\epsilon^2)\log(1/\tau))$ samples.

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 $N = O(d/\epsilon^2)$.

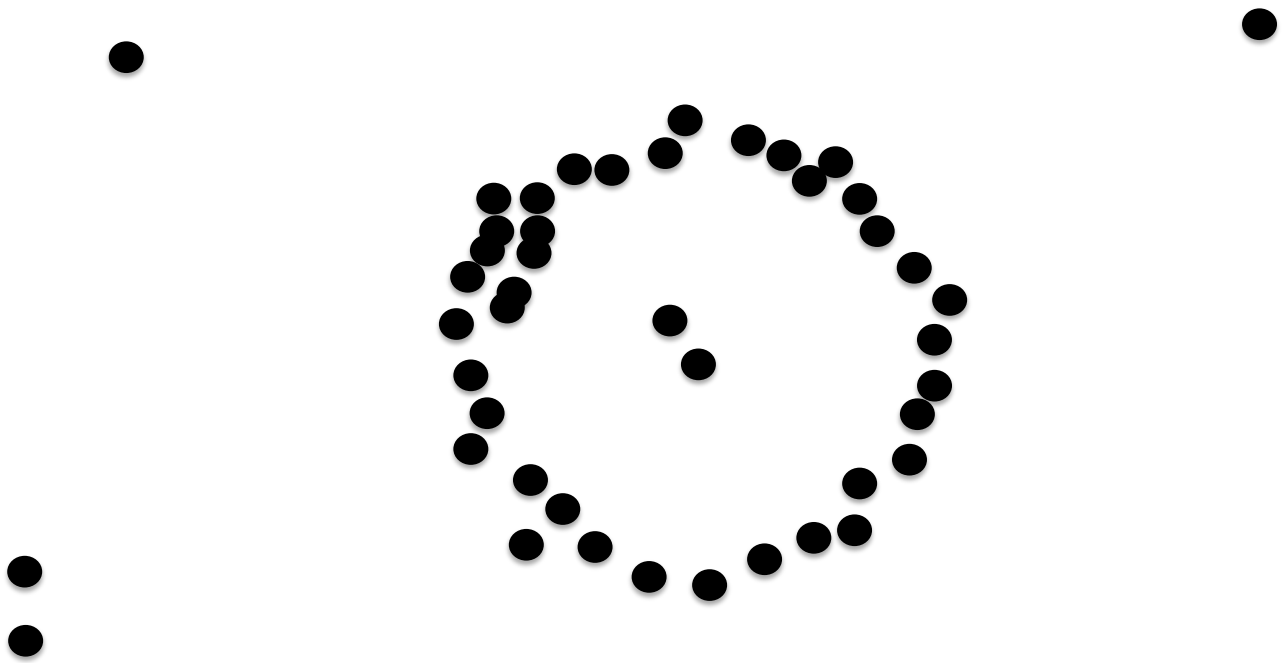
Runtime: $\mathrm{poly}(N, 2^d)$.

■

OUTLIER DETECTION ?



Naïve Outlier Removal (Naïve Pruning)



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

OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- **Overview of Algorithmic Approaches**
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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

Part II: High-Dimensional Robust Mean Estimation

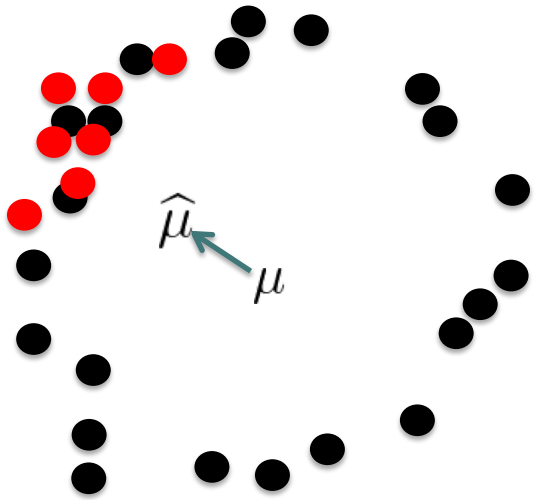
- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- **Certificate of Robustness**
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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



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

● = uncorrupted
● = corrupted

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

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

with high probability we have:

- **[LRV'16]:**

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

  in ***additive*** contamination model

- **[DKKLMS'16]:**

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

  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]:**

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

  in ***additive*** contamination model

- **[DKKLMS'16]:**

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

  in ***strong*** contamination model

KEY LEMMA: ADDITIVE CORRUPTIONS

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]:**

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

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.
Can assume wlog that $\mu = \mathbf{0}$.

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

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

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

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

Proof: We can write

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

■

Proof of Key Lemma: Additive Corruptions (II)

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

Let $\widehat{\Sigma} = (1/N) \sum_{i \in [N]} X_i X_i^T - \widehat{\mu}\widehat{\mu}^T$

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

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

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

Proof: Note that

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

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

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 $\|\widehat{\mu}_B\|_2^2 \leq O(1 + \delta/\epsilon)$

Using Claim 1, we have shown:

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

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

■

Remarks on Key Lemma (Additive Corruptions)

- Same argument holds in finite sample setting.

The following concentration inequalities suffice:

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

$$\|\mu - \widehat{\mu}_G\|_2 \ll \epsilon \qquad \text{and} \qquad \|\widehat{\Sigma}_G - I\|_2 \ll \epsilon$$

- Same proof holds for any “reasonable” isotropic distribution.

KEY LEMMA: STRONG CORRUPTIONS

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:

- **[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.
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 points, and $L \subset G$ is the removed set of samples.

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

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

Proof: We can write

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

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

■

Proof of Key Lemma: Strong Corruptions (II)

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

Let $\widehat{\Sigma} = (1/N)\sum_{i\in[N]} X_i X_i^T - \widehat{\mu}\widehat{\mu}^T$, $\widehat{\Sigma}_G = (1/|G|)\sum_{i\in I_G} X_i X_i^T - \widehat{\mu}_G\widehat{\mu}_G^T$, similarly $\widehat{\Sigma}_B$

$\widehat{M_L} = (1/|L|)\sum_{i\in I_L} X_i X_i^T$.

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

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

Will 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 Claim 3, 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 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 $\widehat{M}_L := (1/|L|)\sum_{i\in I_L} X_i X_i^T = \mathbf{E}_{X\sim_U L}[XX^T]$. 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, $|L|\cdot \mathbf{Pr}_{X\sim_U L}[X\in\mathcal{E}] \le |S|\cdot \mathbf{Pr}_{X\sim_U G}[X\in\mathcal{E}]$.

For any unit vector v:

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

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

■

Remarks on Key Lemma (Strong Corruptions)

- For finite sample version, also need that for every direction v and $T > 0$

$$\mathbf{Pr}_{X \sim_U G}[|v \cdot X| > T] \approx \mathbf{Pr}_{X \sim \mathcal{D}}[|v \cdot X| > T] \; .$$

- Proof holds *as is* for any isotropic distribution with sub-Gaussian tails.
- Essentially same argument goes through if covariance is *approximately* known.

- Argument extends for (approximately known) covariance and weaker concentration.

If $\mathcal{D}$ 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 *only* assumption is that $\mathcal{D}$ has $\Sigma \preceq I$:

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

OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- **Recursive Dimension Halving**
- Iterative Filtering
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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 Idea of Good Subspace Lemma

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

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

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- **Iterative Filtering**
- Soft Outlier Removal
- Application: Robust Stochastic Optimization

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.

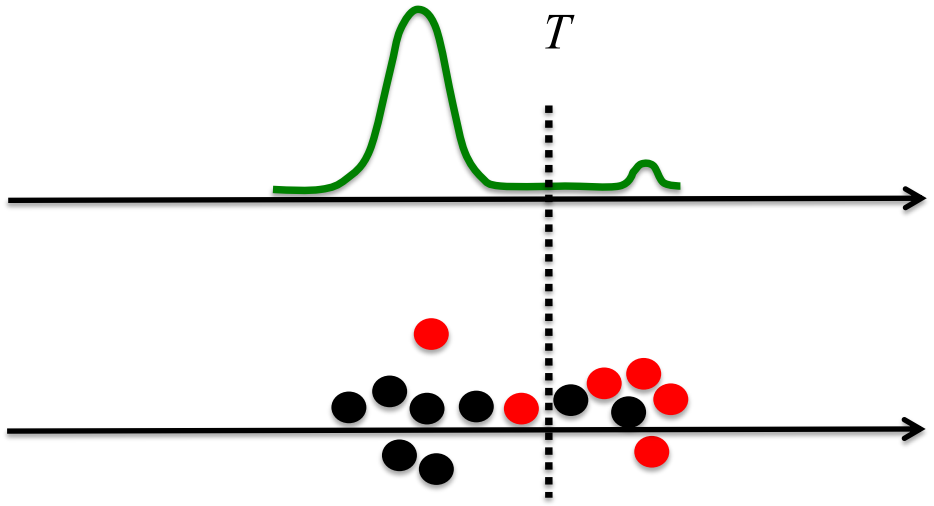
FILTERING SUBROUTINE

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

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

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

After a number of iterations, we stop removing 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

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

5. **Return**

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

Sketch of Correctness

Claim 2: Can always find a threshold satisfying the Condition of Step 4.

Proof Sketch:
By contradiction. Suppose that for all $T > 0$ we have

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

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

Main Idea: Exploit concentration.

Summary: Robust Mean Estimation via Filtering

Certificate for Robustness:

"Spectral norm of empirical covariance is *close* to 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.

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, need *randomized* filtering.

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

- Analogue of Claim 1: Remove more corrupted than good points *in expectation*.

Optimal 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 $f : \mathbb{R}^d \to [0, 1]$

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

OUTLINE

Part I: Introduction

- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- **Soft Outlier Removal**
- Application: Robust Stochastic Optimization

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

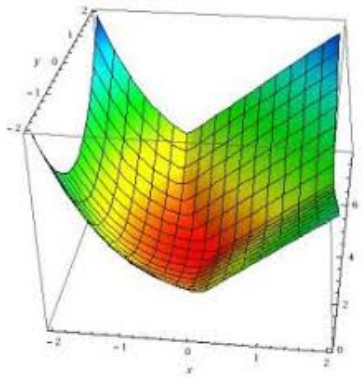
- Motivation
- Robust Statistics in Low and High Dimensions

Part II: High-Dimensional Robust Mean Estimation

- This Talk: Statements of Algorithmic Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- **Application: Robust Stochastic Optimization**

APPLICATION: ROBUST *SUPERVISED* LEARNING

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, …].

CONCLUSIONS

- First Computationally Efficient Robust Estimators in High-Dimensions
- Robust Mean Estimation and Applications
- General Methodology for Various Robust Estimation Tasks

Thank you!
Questions?