Algorithmic Robust Statistics

Ilias Diakonikolas (UW Madison)
Mathematical Methods of Statistics, CIRM
December 2023
Can we develop learning algorithms that are *robust* to a *constant* fraction of *corruptions* in the data?
MOTIVATION

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

• Outlier Detection/Removal

• Adversarial/Secure ML

So Many Misleading, “Fake” Reviews
THE STATISTICAL LEARNING PROBLEM

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

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

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

Main performance criteria:
- Sample size
- Running time
- Robustness
(Outlier-) Robustness

**Strong Contamination Model:**
Let $\mathcal{F}$ be a family of statistical models.
We say that a set of $N$ samples is $\epsilon$-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 $\epsilon$-fraction of them.

cf. Huber’s contamination model [1964]
EXAMPLE: PARAMETER ESTIMATION

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

e.g., a 1-D Gaussian

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

how do we accurately estimate its parameters?

**empirical mean:**

\[
\frac{1}{N} \sum_{i=1}^{N} X_i \to \mu
\]

**empirical variance:**

\[
\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \to \sigma^2
\]
John W. Tukey
Model Misspecification
(1960s)

Peter J. Huber
Robust Estimation of Location
(1964)
What estimators behave well in the presence of outliers?
Robust Estimation: One Dimension

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

- A single corrupted sample can arbitrarily corrupt the empirical mean and variance
- But the median and interquartile range work
Fact [Folklore]: Given a set $S$ of $N$ $\epsilon$-corrupted samples from a one-dimensional Gaussian

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

with high constant probability we have that:

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

where $\hat{\mu} = \text{median}(S)$.

What about robust estimation in high-dimensions?
**High-Dimensional Robust Mean Estimation**

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

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

**Remark**: Above convergence rate is optimal [Tukey’75, Donoho’82]
**Previous Approaches: Robust Mean Estimation**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Error Rate</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance-Based Pruning</td>
<td>$\Theta(\epsilon \sqrt{d})$</td>
<td>$O(dN)$</td>
</tr>
<tr>
<td>Coordinate-wise Median</td>
<td>$\Theta(\epsilon \sqrt{d})$</td>
<td>$O(dN)$</td>
</tr>
<tr>
<td>Geometric Median</td>
<td>$\Theta(\epsilon \sqrt{d})$</td>
<td>$\text{poly}(d, N)$</td>
</tr>
<tr>
<td>Tukey Median</td>
<td>$\Theta(\epsilon)$</td>
<td>NP-Hard</td>
</tr>
<tr>
<td>Tournament</td>
<td>$\Theta(\epsilon)$</td>
<td>$\mathcal{N}^{O(d)}$</td>
</tr>
</tbody>
</table>
DISTANCE-BASED PRUNING
DISTANCE-BASED PRUNING = NAÏVE OUTLIER REMOVAL
All known estimators either require exponential time to compute or can tolerate a negligible fraction of outliers.

Is robust estimation algorithmically possible in high-dimensions?
“The bad news is that with all currently known algorithms the effort of computing those estimates increases exponentially in \( d \). We might say they break down by failing to give a timely answer!

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.[…]”

Meta-Theorem [D-Kamath-Kane-Li-Moitra-Stewart’16]
Efficient robust estimators with *dimension-independent* error for robust mean and covariance estimation, if inlier distribution has bounded moments/nice concentration.

Related results by [Lai-Rao-Vempala’16]
**Robust Unsupervised Learning**

- Robustly Learning Graphical Models
- Computational/Statistical-Robustness Tradeoffs
- List-decodable Learning and Robustly Learning Mixture Models
Robust Supervised Learning

Robust Regression

Stochastic Convex Optimization
APPLICATIONS

Detecting Patterns in Biological Data

Provable Defenses against Data Poisoning

OOD Detection

[D-Kamath-Kane-Li-Moitra-Stewart, ICML’17]

[D-Kamath-Kane-Li-Moitra-Steinhardt, ICML’19]

[Du-Fang-D-Li, ’23]

Wild data scenario 1

Wild data scenario 2

(a) Data setup

(b) Filtered outliers (in green)

(c) Uncertainty score
SUBSEQUENT WORKS

- **Sparse Models** [Balakrishan-Du-Li-Singh’17, D-Karmalkar-Kane-Price-Stewart’19, D-Kane-Lee-Pensia’22, …]
- **Graphical Models** [Cheng-D-Kane-Stewart’18, D-Kane-Stewart-Sun’21, D-Kane-Sun’22]
- **Robust Regression/Classification** [D-Kane-Stewart’18, Klivans-Kothari-Meka’18, D-Kong-Stewart’19 Bakshi-Prasad’21, …]
- **Robust Stochastic Optimization** [Prasad-Suggala-Balakrishnan-Ravikumar’19, D-Kamath-Kane-Li-Steinhard-Stewart’19, …]
- **Near-Linear Time Algorithms** [Chen-D-Ge’18, Cheng-D-Ge-Woodruff’19, Depersin-Lecue’19, Dong-Hopkins-Li’19, Li-Ye’20, Cherapanamjeri-Mohanty-Yau’20, D-Kane-Koongsgard-Li-Tian’21, …]
- **Computational-Statistical Tradeoffs** [D-Kane-Stewart’17, D-Kong-Stewart’19, Hopkins-Li’19, …]
- **Connections to Non-Convex Optimization** [Chen-D-Ge-Soltanolkotabi’20, Zhu-Jiao-Steinhardt’20, …]
- **List-Decodable Learning** [Charikar-Steinhardt-Valiant’17, D-Kane-Stewart’18, Meister-Valiant’18, Karmalkar-Klivans-Kothari’19, Raghavendra-Yau’19, D-Kane-Koongsgard’20, D-Kane-Koongsgard-Li-Tian’21, D-Kane-Karmalkar-Pensia-Pittas’22]
- **Applications in Data Analysis** [D-Kamath-Kane-Li-Moitra-Stewart’17, Tran-Li-Madry’18, D-Kamath-Kane-Li-Steinhard-Stewart’19, Hayase-Kong-Somani-Oh’21, Du-Fang-D-Li’23, …]
HIGH-DIMENSIONAL ROBUST MEAN ESTIMATION
**Robust Mean Estimation: Gaussian Case**

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

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

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

in the *additive contamination* model.

First-term of RHS Independent of $d$!

[D-Kamath-Kane-Li-Moitra-Stewart, SODA’18; D-Kane-Pensia-Pittas, NeurIPS’23]
**Robust Mean Estimation: Sub-Gaussian Case**

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

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

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

in the strong contamination model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS’16, ICML’17; D-Kane-Pensia-Pittas, ICML’22]
ROBUST MEAN ESTIMATION: BOUNDED COVARIANCE CASE

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

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

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

in the strong contamination model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, ICML’17; Steinhardt, Charikar, Valiant, ITCS’18]
Idea #1: If the empirical covariance is “close to what it should be”, then the empirical mean works.
CERTIFICATE FOR EMPRICAL MEAN

Detect when the empirical estimator may be compromised

\[ \hat{\mu} \triangleq \frac{1}{N} \sum_{i=1}^{N} X_i \]

- \(\bullet\) = uncorrupted
- \(\bullet\) = corrupted

There is no direction of large empirical variance
Lemma: Let \( X_1, X_2, \ldots, X_N \) be an \( \epsilon \)-corrupted set of samples from \( \mathcal{N}(\mu, I) \) and \( N = \Omega(d/\epsilon^2) \) for

\[
\hat{\mu} \triangleq \frac{1}{N} \sum_{i=1}^{N} X_i \quad \hat{\Sigma} \triangleq \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{\mu})(X_i - \hat{\mu})^T
\]

with high probability we have:

\[
\|\hat{\Sigma}\|_2 \leq 1 + \lambda \quad \longrightarrow \quad \|\hat{\mu} - \mu\|_2 \leq O(\epsilon \sqrt{\log(1/\epsilon)} + \sqrt{\epsilon \lambda})
\]

in strong contamination model.
Idea #2: Removing any $\epsilon$-fraction of inliers does not move the empirical mean and covariance by much.
Idea #3: Iteratively “remove outliers” to “fix” the empirical covariance.
Iterative Filtering

Iterative Two-Step Procedure:

Step #1: Test certificate of robustness of “standard” estimator
Step #2: If certificate is violated, detect and remove outliers
Iterate on “cleaner” dataset.

General recipe that works in general settings.

We’ll see how this works for robust mean estimation.
FILTERING SUBROUTINE

Either output empirical mean or remove many outliers.

Filtering Approach: Suppose that:

\[ \| \hat{\Sigma} \|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon)) \]

Let \( v^* \) be the direction of maximum variance.
FILTERING SUBROUTINE

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

\[ \| \hat{\Sigma} \|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon)) \]

Let \( v^* \) be the direction of maximum variance.

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

\[ \Pr_{X \sim U_S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] \geq 8 \cdot e^{-T^2/2}. \]

- 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:

\[ \| \hat{\Sigma} \|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon)) \]

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

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

Eventually the empirical mean works

Runtime: \( \tilde{O}(Nd^2) \)
**STABILITY CONDITION**

**Definition** Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A set $S \subset \mathbb{R}^d$ is $(\epsilon, \delta)$–stable with respect to $\mu$ if for all $v \in S^{d-1}$ and every $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon)|S|$, we have:

- $\left| \frac{1}{|S'|} \sum_{x \in S'} v \cdot (x - \mu) \right| \leq \delta \iff \|\mu_{S'} - \mu\|_2 \leq \delta$
- $\left| \frac{1}{|S'|} \sum_{x \in S'} (v \cdot (x - \mu))^2 - 1 \right| \leq \delta^2 / \epsilon \iff \|\Sigma_{S'} - I\|_2 \leq \delta^2 / \epsilon$

- Intended for inlier distributions with $\Sigma \preceq I$
- Similar definition for distributions as opposed to datasets.
- A sufficiently large clean sample from a well-behaved distribution is stable with high probability.
# Efficient Robust Mean Estimation under Stability

**General Theorem** Let \( S \) be \((\epsilon, \delta)\)-stable with respect to a vector \( \mu \), and \( T \) an \( \epsilon \)-corruption of \( S \). There is an efficient algorithm that given \( \epsilon, \delta, T \) it computes an estimate \( \hat{\mu} \) such that

\[
\|\hat{\mu} - \mu\|_2 = O(\delta)
\]

**Fact** A set of \( N \) i.i.d. samples from a well-behaved distribution is \((\epsilon, \delta)\)-stable with high probability.

- For identity covariance sub-Gaussians, \( \delta \sim \epsilon \sqrt{\log(1/\epsilon)} \) and \( N \gg d/\delta^2 \)
- For identity covariance sub-exponentials, \( \delta \sim \epsilon \log(1/\epsilon) \) and \( N \gg d/\delta^2 \)
- For identity covariance with bounded \( k \)-th central moments \((k \geq 4)\), \( \delta \sim \epsilon^{1-1/k} \) and \( N \gg d(\log d)/\delta^2 \)
- For bounded covariance distributions, \( \delta \sim \sqrt{\epsilon} \) and \( N \gg d(\log d)/\delta^2 \)
  (after removing \( \epsilon \)-fraction of inliers)
**Certificate for Empirical Mean**

**Lemma** Let $S$ be $(\epsilon, \delta)$—stable with respect to $\mu$, and $T$ be an $\epsilon$—corruption of $S$. If $\|\Sigma_T\|_2 \leq 1 + \lambda$, for $\lambda \geq 0$, then

$$\|\mu_T - \mu\|_2 \leq O(\delta + \sqrt{\epsilon \lambda})$$

**Proof** Let $X, Y$ be uniform distribution over $S, T$ respectively. Can write $Y = (1 - \epsilon)X' + \epsilon E$, where $X'$ is $\epsilon$—subtraction of $X$.

$$\Sigma_Y = (1 - \epsilon)\Sigma_{X'} + \epsilon \Sigma_E + \epsilon(1 - \epsilon)(\mu_{X'} - \mu_E)(\mu_{X'} - \mu_E)^\top$$

Let $u$ be normalized version of $\mu_{X'} - \mu_E$.

$$1 + \lambda \geq u^\top \Sigma_Y u = (1 - \epsilon)u^\top \Sigma_{X'} u + \epsilon u^\top \Sigma_E u + \epsilon(1 - \epsilon)u^\top (\mu_{X'} - \mu_E)(\mu_{X'} - \mu_E)^\top u$$

$$\geq (1 - \epsilon)(1 - \delta^2/\epsilon) + \epsilon(1 - \epsilon)\|\mu_{X'} - \mu_E\|_2^2$$

$$\geq 1 - O(\delta^2/\epsilon) + (\epsilon/2)\|\mu_{X'} - \mu_E\|_2^2$$

Rearranging

$$\|\mu_{X'} - \mu_E\|_2 = O(\delta/\epsilon + \sqrt{\lambda/\epsilon})$$
**Certificate for Empirical Mean**

**Lemma** Let $S$ be $(\epsilon, \delta)$—stable with respect to $\mu$, and $T$ be an $\epsilon$-corruption of $S$. If $\|\Sigma_T\|_2 \leq 1 + \lambda$, for $\lambda \geq 0$, then

$$\|\mu_T - \mu\|_2 \leq O(\delta + \sqrt{\epsilon \lambda})$$

**Proof** Let $X, Y$ be uniform distribution over $S, T$ respectively. Can write $Y = (1 - \epsilon)X' + \epsilon E$, where $X'$ is $\epsilon$—subtraction of $X$.

$$\|\mu_{X'} - \mu_E\|_2 = O(\delta/\epsilon + \sqrt{\lambda/\epsilon})$$

For the means, have that $\mu_T = \mu_Y = (1 - \epsilon)\mu_{X'} + \epsilon \mu_E$.

$$\|\mu_T - \mu\|_2 = \|(1 - \epsilon)\mu_{X'} + \epsilon \mu_E - \mu\|_2 = \|\mu_{X'} - \mu + \epsilon(\mu_E - \mu_{X'})\|_2$$

$$\leq \|\mu_{X'} - \mu\|_2 + \epsilon \|\mu_{X'} - \mu_E\|_2$$

$$= O(\delta) + \epsilon \cdot O(\delta/\epsilon + \sqrt{\lambda/\epsilon})$$

$\blacksquare$
**Randomized Filtering: Idea**

**Main Idea:** Suppose we can find $f : T \to \mathbb{R}_{\geq 0}$ such that

$$\sum_{x \in T} f(x) \geq 2 \sum_{x \in S} f(x).$$

Then we can randomly filter by removing each point $x \in T$ with probability $\propto f(x)$.  

Need this property to hold across iterations, assuming certificate not satisfied.

**Condition** Given any $T' \subseteq T$ such that $|T' \cap S| \geq (1 - 4\epsilon)|S|$, if $\|\Sigma_{T'}\|_2 \geq 1 + \lambda$ there is an explicit $f : T' \to \mathbb{R}_{\geq 0}$ such that

$$\sum_{x \in T'} f(x) \geq 2 \sum_{x \in T' \cap S} f(x).$$
**Randomized Filtering: Properties**

**Condition** Given any $T' \subseteq T$ such that $|T' \cap S| \geq (1 - 4\epsilon)|S|$ , if $\|\Sigma_{T'}\|_2 \geq 1 + \lambda$ there is an explicit $f : T' \rightarrow \mathbb{R}_{\geq 0}$ such that

$$\sum_{x \in T'} f(x) \geq 2 \sum_{x \in T' \cap S} f(x)$$

**Theorem** If condition holds, there is an efficient randomized algorithm that computes an estimate $\hat{\mu}$ such that with high probability

$$\|\hat{\mu} - \mu_X\|_2 = O(\delta + \sqrt{\epsilon \lambda})$$
1. Compute $\nu = \|\Sigma_T\|_2$
2. If $\nu \leq 1 + \lambda$, return $\mu_T$
3. Else
   - Compute the function $f$.
   - Remove each $x \in T$ with probability $f(x)/\max_{x \in T} f(x)$
   - Return to Step 1 with new set $T$. 

Randomized Filtering Pseudocode
**Randomized Filtering: Analysis**

At least one point is removed in each iteration, so algorithm runs in polynomial time.

**Claim** With probability at least 2/3, throughout the algorithm have that $|S \cap T_i| \geq (1 - 4\epsilon)|S|$.

**Proof** Consider $d(T_i) := |(S \cap T) \setminus T_i| + |T_i \setminus S|$.

$$d(T_i) - d(T_{i-1}) = (\#\text{Inliers removed in iteration } i) - (\#\text{Outliers removed in iteration } i)$$

$$\mathbb{E}[d(T_i) - d(T_{i-1})] = \sum_{x \in S \cap T_i} f(x) - \sum_{x \in T_i \setminus S} f(x) = 2 \sum_{x \in S \cap T_i} f(x) - \sum_{x \in T_i} f(x) \leq 0.$$  

Since $d(T_i) \geq 0$ and $\mathbb{E}[d(T_i)] \leq \mathbb{E}[d(T_0)] \leq \epsilon |S|$, by Ville’s inequality

$$\Pr[\max_i d(T_i) > 3\epsilon |S|] \leq 1/3.$$  

This implies that $|S \cap T_i| \geq (1 - 4\epsilon)|S|$ throughout. $\blacksquare$
**Proposition** Let $S$ be $(2\epsilon, \delta)$–stable and $T$ be an $\epsilon$–corruption of $S$. Suppose that $\|\Sigma_T\|_2 = 1 + \lambda > 1 + 8\delta^2 / \epsilon$. There exists an efficient algorithm that given $\epsilon, \delta, T$ it computes a function $f : T \rightarrow \mathbb{R}_{\geq 0}$ such that

$$\sum_{x \in T} f(x) \geq 2 \sum_{x \in T \cap S} f(x).$$

**Proof** Define the function $g(x) = (v \cdot (x - \mu_T))^2$, where $v$ is the top eigenvector. Let $L$ be the set of $\epsilon \cdot |T|$ points $x \in T$ for which $g(x)$ is largest. Then

$$f(x) = \begin{cases} 0 & x \notin L \\ g(x) & x \in L \end{cases}$$

![Diagram](image)

- $g(x^{(1)})$, $g(x^{(2)})$, $g(x^{(3)})$
- $L$
Universal Filtering: Analysis

- By definition
  \[ \sum_{x \in T} g(x) = |T| \text{Var}[v \cdot T] = |T|(1 + \lambda) \]
  and
  \[ \sum_{x \in S} g(x) = |S|\left(\text{Var}[v \cdot S] + (v \cdot (\mu_T - \mu_S))^2\right) \]

- By stability and our lemma \( \sum_{x \in S} g(x) \) is small so that
  \[ \sum_{x \in T \setminus S} g(x) \geq \sum_{x \in T} g(x) - \sum_{x \in S} g(x) \geq (2/3)|S|\lambda. \]

- By the definition of \( L \) and \( \lambda \)
  \[ \sum_{x \in T} f(x) = \sum_{x \in L} g(x) \geq \sum_{x \in T \setminus S} g(x) \]

- Similarly
  \[ \sum_{x \in S \cap T} f(x) = \sum_{x \in S \cap L} g(x) = \sum_{x \in S} g(x) - \sum_{x \in S \setminus L} g(x) \leq 2|S|\delta^2/\epsilon + |S|O(\delta^2 + \epsilon\lambda) \]
**Weighted Filtering**

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

For \( w : T \rightarrow \mathbb{R}_+ \)

\[
\mu_w[T] := \frac{1}{\| w \|_1} \sum_{x \in T} w_x x \\
\Sigma_w[T] := \frac{1}{\| w \|_1} \sum_{x \in T} w_x (x - \mu_w)(x - \mu_w)^\top
\]

### Weighted Filtering Pseudocode

1. Set \( t = 1 \) and \( w_x^{(1)} = 1/|T| \) for \( x \in T \)
2. While \( \| \Sigma_{w(t)}[T] \|_2 > 1 + \lambda \)
   - Compute the function \( f \).
   - Set \( f_{\max} = \max \{ f(x) \mid x \in T \text{ and } w_x^{(t)} \neq 0 \} \)
   - Set \( w_x^{(t+1)} = w_x^{(t)} \left( 1 - f(x)/f_{\max} \right) \)
   - Set \( t \) to \( t + 1 \)
3. Return \( \mu_{w(t)} \)
**Lemma:** Let $T$ be an $\epsilon-$corruption of a $(3\epsilon, \delta)-$stable set. For any $w \in \Delta_{T,\epsilon}$, if

$$
\|\Sigma w [T]\|_2 \leq 1 + \lambda \quad \Rightarrow \quad \|\mu_w [T] - \mu\|_2 = O(\delta + \sqrt{\epsilon \lambda})
$$

**Non-Convex Optimization Formulation:**

$$
\min_{w \in \Delta_{T,\epsilon}} \|\Sigma w [T]\|_2
$$
NON-CONVEX OPTIMIZATION FORMULATION (II)

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

Let \( \Delta_{T,\epsilon} = \left\{ w \in \mathbb{R}_{\geq 0}^T \text{ with } \|w\|_1 = 1 \text{ and } w_x \leq \frac{1}{|T|(1-\epsilon)} \right\} \)

Non-Convex Optimization Formulation:
\[
\min_{w \in \Delta_{T,\epsilon}} \| \Sigma_w [T] \|_2
\]

Algorithmic Approaches:
• This is what filtering does!
• Ellipsoid Method [DKKLMS’16]
• Bi-level optimization [Cheng-D-Ge’18]
• Gradient Descent [Cheng-D-Ge-Soltanolkotabi’20]
**Concrete Open Problems**

- **Design near-linear time algorithms for robust statistics tasks**
  - Robust Mean Estimation [Cheng-D-Ge, SODA’19; Dong-Hopkins-Li, NeurIPS’19; Depersin-Lecue’19]
  - Robust Covariance Estimation [Cheng-D-Ge-Woodruff, COLT’19]
  - Clustering mixture models [D-Kane-Koongsgard-Li-Tian, STOC’22]
  - *Robust sparse estimation?*

- **Can we design robust estimators using first-order methods?**
  - Robust Mean Estimation [Cheng-D-Ge-Soltanolkotabi, ICML’20; Zhu et al. 2020]
  - *More general tasks?*

- **Obtain low-memory streaming robust learning algorithms**
  - [D-Kane-Pensia-Pittas, ICML’22] *Tradeoffs between memory and sample size?*

- **Robust Online Estimation?**
INFORMATION-COMPUTATION TRADEOFFS (IN ROBUST STATISTICS)
**Observed Statistical-Information Gaps**

**Problem 1:** Robust Mean Estimation for $\mathcal{N}(\mu, I)$ in strong contamination model

- Information-theoretic: $O(\epsilon)$
- Computational: $O(\epsilon \sqrt{\log(1/\epsilon)})$  [D-Kane-Kamath-Li-Moitra-Stewart’16]

**Problem 2:** Robust Sparse Mean Estimation for $\mathcal{N}(\mu, I)$ in Huber’s model

- Information-theoretic: $O(k \log(d)/\epsilon^2)$
- Computational: $O(k^2 \log(d)/\epsilon^2)$  [Li’17]

**Problem 3:** Robust covariance estimation for $\mathcal{N}(0, \Sigma)$ in spectral norm

- Information-theoretic: $O(d)$
- Computational: $\Omega(d^2)$  [D-Kane-Kamath-Li-Moitra-Stewart’16]

Are these observed information-computation gaps inherent?
**Statistical Query (SQ) Model [Kearns’93]**

**Statistical Query:** $Q_i : X \to [-1, 1]$

Complexity measures
- Number of queries: $q$
- Query tolerance: $\tau$

**Unrestricted algorithm**

$STAT_D(\tau)$ returns $a_i : |a_i - E_{x \sim D}[Q_i(x)]| \leq \tau$

Sample complexity
**Power of SQ Algorithms**

- **Restricted Model**: Can prove unconditional lower bounds.

- **Powerful Model**: Wide range of algorithmic techniques in ML are implementable using SQs:
  - PAC Learning: AC$^0$, decision trees, linear separators, boosting

- **Exceptions**: Gaussian elimination, lattice basis-reduction [D-Kane’22, Zadik-Song-Wein-Bruna’22]

- **SQ Model $\approx$ Low-degree Polynomial Tests** [Brennan-Bresler-Hopkins-Li-Schramm’21]
**INTERPRETATION OF SQ LOWER BOUNDS**

Suppose we have proved:

Any SQ algorithm for problem $P$
- either requires queries of **tolerance** at most $\tau$
- or makes at least $q$ **queries**.

Then we can interpret:

Any SQ algorithm* for problem $P$
- either requires at least $1/\tau^2$ **samples**
- or has **runtime** at least $q$. 
**SQ Lower Bound for Robust Mean Estimation**

**Theorem:** Any SQ algorithm that learns an $\epsilon$-corrupted Gaussian $\mathcal{N}(\mu, I)$ in the strong contamination model within error $o(\epsilon \sqrt{\log(1/\epsilon)})$ requires either:

- SQ queries of accuracy $d^{-\omega(1)}$
- or
- at least $d^{\omega(1)}$ many SQ queries.

**Take-away:** Any asymptotic improvement in error guarantee over filtering algorithm requires super-polynomial time.
**SQ LOWER BOUND FOR ROBUST SPARSE MEAN ESTIMATION**

**Theorem:** Any SQ algorithm that learns an $\epsilon$-corrupted Gaussian $\mathcal{N}(\mu, I)$ where is $k$-sparse within constant error requires either:
- $\Omega(k^2)$ samples
- or
- at least $d^{k^{\Omega(1)}}$ many SQ queries.

Minimax sample complexity is $\Theta(k \log(d/k)/\epsilon^2)$

**Take-away:** Any asymptotic improvement in error guarantee over known efficient algorithms [Li’17, DKKPS’19, …] requires super-polynomial time.
**SQ LOWER BOUND FOR LEARNING GMMs**

**Theorem:** Any SQ algorithm that learns GMMs on $\mathbb{R}^d$ to constant total variation error requires either:

- $d^{\Omega(k)}$ samples
- or
- at least $2^{d^{\Omega(1)}}$ many SQ queries.

even if the components are pairwise separated in total variation distance.

Minimax sample complexity is $\text{poly}(d, k)$

**Take-away:** Computational complexity of learning separated GMMs is inherently exponential in number of components.
NON-GAUSSIAN COMPONENT ANALYSIS (NGCA)

Given samples from a distribution on $\mathbb{R}^d$, find a hidden “non-Gaussian” direction.

• Introduced in [Blanchard-Kawanabe-Sugiyama-Spokoiny-Muller’06].

• Studied extensively from algorithmic standpoint.
  [Kawanabe-Theis’06; Kawanabe-Sugiyama-Blanchard-Muller’07; Diederichs-Juditsky-Spokoiny-Schutte’10; Diederichs-Juditsky-Nemirovski-Spokoiny’13; Bean’14; Sasaki-Niu-Sugiyama’16; Virta-Nordhausen-Oja’16; Vempala-Xiao’11; Tan-Vershynin’18; Goyal-Shetty’19]
**Non-Gaussian Component Analysis (NGCA): Definition**

**Definition:** Let \( \mathbf{v} \) be a unit vector in \( \mathbb{R}^d \) and \( A : \mathbb{R} \rightarrow \mathbb{R}_+ \) be a pdf. We define \( P_v^A \) to be the distribution with \( \mathbf{v} \)-projection equal to \( A \) and \( \mathbf{v}^\perp \)-projection an independent standard Gaussian.

**NGCA Problem:** Given \( A \) that matches the first \( m \) moments with \( \mathcal{N}(0, 1) \):
Using i.i.d. samples from \( P_v^A \) where \( \mathbf{v} \) is unknown, find the hidden direction \( \mathbf{v} \).
NGCA captures interesting instances of several (robust) learning tasks
• Learning Gaussian Mixtures [D-Kane-Stewart’17, D-Kane-Pittas-Zarifis’23]
• Robust mean and covariance estimation [D-Kane-Stewart’17]
• Robust sparse mean estimation, sparse PCA [D-Kane-Stewart’17, D-Stewart’18]
• Robust linear regression [D-Kong-Stewart’19]
• List-decodable learning [D-Kane-Stewart’18, D-Kane-Pensia-Pittas-Stewart’21]
• Adversarially robust PAC learning [Bubeck-Price-Razenshteyn’18]
• Agnostic PAC Learning [Goel-Gollakota-Klivans’20, D-Kane-Zarifis’20, D-Kane-Pittas-Zarifis’21]
• Learning LTFs with (Semi)-random Noise [D-Kane’20, Nasser-Tiegel’22, D-J.D.-Kane-Wang-Zarifis’23]
• Learning (Very Simple) NNs and Generative Models [Goel-Gollakota-Jin-Karmalkar-Klivans’20, D-Kane-Kontonis-Zarifis’20 Chen-Li-Li’22]
• Learning Mixtures of LTFs [D-Kane-Sun’23]
• …
INFORMAL LOWER BOUND RESULT

Fact: Non-Gaussian Component Analysis
• Can be solved with \( \text{poly}(d, m) \) samples.
• All known efficient algorithms require at least \( d^{\Omega(m)} \) samples (and time).

Informal Theorem: For any “nice” univariate distribution \( \mathcal{A} \) matching its first \( m \) moments with the standard Gaussian, any* algorithm that solves NGCA
• either draws at least \( d^{\Omega(m)} \) samples
• or has runtime \( 2d^{\Omega(1)} \).

*holds for any Statistical Query (SQ) algorithm

[D-Kane-Stewart, FOCS’17; D-Kane-Ren-Sun, NeurIPS’23]
**General Methodology For SQ Lower Bounds**

**Hypothesis Testing Problem**: Given access to a distribution $D$ on $\mathbb{R}^d$ with promise that
- either $D = D_0$
- or $D$ is selected randomly from $\mathcal{D} = \{D_u\}_{u \in S}$ according to prior $\mu$
the goal is to distinguish between the two cases.

**Pairwise correlation**: $\chi_{D_0}(p, q) = \mathbb{E}_{x \sim D_0}[(p/D_0)(x)(q/D_0)(x)] - 1$

**Theorem [FGRVX'17]**: Suppose there exists a “large” set of distributions in $\mathcal{D}$ with “small” pairwise correlation with respect to $D_0$. Then any SQ algorithm for hypothesis testing task:
- either requires at least one “high-accuracy” query
- or requires a ”large” number of queries.
**STATISTICAL QUERY HARDNESS OF NGCA**

**Testing Version of NGCA:** Given access to a distribution $D$ on $\mathbb{R}^d$ with the promise that
- either $D = \mathcal{N}(0, I)$
- or $D = P^A_v$, where $v$ is a uniformly random unit vector
the goal is to distinguish between the two cases.

**Main Theorem [D-Kane-Stewart’17]**
Suppose that $A$ matches its first $m$ moments with $\mathcal{N}(0, 1)$ and $\chi^2(A, \mathcal{N}(0, 1)) < \infty$.
Any SQ algorithm for the testing version of NGCA:
- either requires a query of tolerance at most $d^{-\Omega(m)} \chi^2(A, \mathcal{N}(0, 1))^{1/2}$
- or requires at least $2^{d^{\Omega(1)}}$ many queries.
**Intuition: Why is NGCA “Hard”?**

**Claim 1:** Low-degree moments do not help.

- Degree at most $m$ moment tensor of $P_v^A$ identical to that of $\mathcal{N}(0, I_d)$

**Claim 2:** Random projections do not help.

Distinguishing requires exponentially many random projections.
**Key Lemma:** Let $Q$ be the distribution of $v' \cdot X$, where $X \sim P^A_v$. Then, we have that:

$$\chi^2(Q, \mathcal{N}(0, 1)) \leq (v \cdot v')^{2(m+1)} \chi^2(A, \mathcal{N}(0, 1))$$

**Diagram:**

- $X \sim P^A_v$
- $v' \cdot X \sim Q$
- $v_{\perp} \cdot X \sim \mathcal{N}(0, 1)$
- $v' \cdot X \sim A$

**Ornstein-Uhlenbeck operator**

\[ Q = U_{\theta}(A) \]
SQ LOWER BOUND: PROOF OVERVIEW

Want exponentially many $P^A_v$'s that are nearly uncorrelated.

- Pick set $\mathcal{V}$ of near-orthogonal unit vectors. Can get $|\mathcal{V}| = 2^{\Omega(1)}$

- Have

$$\chi_{\mathcal{N}(0, I_d)}(P^A_v, P^A_{v'}) = \chi_{\mathcal{N}(0, 1)}(A, U\theta A) \leq |\cos^{m+1}(\theta)|^2 \chi^2(A, \mathcal{N}(0, 1))$$
**Recipe For SQ Hardness Results**

**Main Theorem** [D-Kane-Stewart’17]
Suppose that $A$ matches its first $m$ moments with $\mathcal{N}(0, 1)$ and $\chi^2(A, \mathcal{N}(0, 1)) < \infty$.

Any SQ algorithm for the testing version of NGCA:
• either requires a query of tolerance at most $d^{-\Omega(m)} \chi^2(A, \mathcal{N}(0, 1))^{1/2}$
• or requires at least $2^{d^{\Omega(1)}}$ many queries.

**Recipe.** Encode $\Pi$ as a NGCA instance:

• Construct moment-matching distribution $A$ such that $P^A_v$ is a valid instance of $\Pi$.
• Match as many low-degree moments as possible.
Lemma: There exists a univariate distribution $A$ such that:

- $A$ agrees with $\mathcal{N}(0, 1)$ on the first $m$ moments
- $A$ satisfies $d_{TV}(A, N(\delta, 1)) \leq O(\delta m^2/\sqrt{\log(1/\delta)})$

Proof Idea:

- Take $C = \Theta(\sqrt{\log(1/\delta)})$
- Define

\[
A(x) = \begin{cases} 
  G(x - \delta), & x \notin [-C, C] \\
  G(x - \delta) + p(x), & x \in [-C, C]
\end{cases}
\]

where $p$ is degree-$m$ moment-matching polynomial.
**Moment-Matching for Learning GMMs**

**Lemma:** There exists a univariate $k$-GMM $A$ with nearly non-overlapping components such that: $A$ agrees with $\mathcal{N}(0, 1)$ on the first $2k-1$ moments.

**Proof Idea:**
- Construct discrete distribution $B$ with support $k$ matching its first $2k-1$ moments with $\mathcal{N}(0, 1)$.
- Rescale $B$ and add a “skinny” Gaussian to get $A$. 
SQ HARD INSTANCES FOR GMMS: PARALLEL PANCAKES
SQ HARDNESS FOR WIDE RANGE OF PROBLEMS

NGCA captures SQ hard instances of several well-studied learning tasks

- Learning GMMs [D-Kane-Stewart’17, D-Kane-Pittas-Zarifis’23]
- Robust mean and covariance estimation [D-Kane-Stewart’17]
- Robust sparse mean estimation, sparse PCA [D-Kane-Stewart’17, D-Stewart’18]
- Robust linear regression [D-Kong-Stewart’19]
- List-decodable learning [D-Kane-Stewart’18, D-Kane-Pensia-Pittas-Stewart’21]
- Adversarially robust PAC learning [Bubeck-Price-Razenshteyn’18]
- Agnostic PAC Learning [Goel-Gollakota-Klivans’20, D-Kane-Zarifis’20, D-Kane-Pittas-Zarifis’21]
- Learning LTFs with (Semi)-random Noise [D-Kane’20, Nasser-Tiegal’22, D-J.D.-Kane-Wang-Zarifis’23]
- Learning (Very Simple) NNs and Generative Models [Goel-Gollakota-Jin-Karmalkar-Klivans’20, D-Kane-Kontonis-Zarifis’20 Chen-Li-Li’22]
- Learning Mixtures of LTFs [D-Kane-Sun’23]
- ...
OPEN PROBLEMS

NGCA leads to wide range of hardness results in SQ model

Open Problem 1: Alternative evidence of hardness?

Already known for special cases (reductions):
- Robust sparse mean estimation [Brennan-Bresler’20]
- Learning GMMs [Bruna-Regev-Song-Tang’21]
- Learning with Semi-random Noise [D-Kane-Panurangsi-Ren’22, D-Kane-Ren’23]

Open Problem 2: How general is this phenomenon?

SQ hard instances are computationally hard

Open Problem 3: Prove SoS lower bounds for NGCA.
LEARNING WITH A MAJORITY OF OUTLIERS

• So far focused on setting where \( \epsilon < 1/2 \).

• What can we learn from a dataset in which the \textit{majority} of points are corrupted?

Problem: Given a set of points \( x_1, \ldots, x_N \in \mathbb{R}^d \) and \( 0 < \alpha \leq 1/2 \) such that:
• An unknown subset of \( \alpha N \) points are drawn from an unknown \( D \in \mathcal{F} \), and
• The remaining \( (1 - \alpha)N \) points are arbitrary,
approximate the mean \( \mu \) of \( D \).

Which is the “real” \( D \)?
LIST-DECODABLE LEARNING

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

List-Decodable Mean Estimation:
Given a set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ and $0 < \alpha \leq 1/2$ such that:
• An unknown subset of $\alpha N$ points are drawn from an unknown $D \in \mathcal{F}$, and
• The remaining $(1 - \alpha) N$ points are arbitrary,
output a small list of $s$ hypotheses vectors such that one is close to the mean $\mu$ of $D$.

• Model defined in [Balcan-Blum-Vempala’08]
• First studied for mean estimation [Charikar-Steinhardt-Valiant’17]
• Application: Learning Mixture Models
LIST-DECODABLE MEAN ESTIMATION

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

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

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

- Initial algorithm [CSV’17] based on ellipsoid method.
- Generalization of filtering (“multi-filtering”) works for list-decodable setting [DKS’18].
- Near-linear time algorithm [D-Kane-Koongsgard-Li-Tian’22].
FUTURE DIRECTIONS: ALGORITHMS

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

General Algorithmic Theory of Robustness

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

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

Broader Challenges:

• Relation to Related Notions of Algorithmic Stability (Differential Privacy, Adaptive Data Analysis)
• Resource tradeoffs (e.g., memory, communication)
• Further Applications (ML Security, Computer Vision, …)
• Connections to Adversarial Examples/Distribution Shift
• Other notions of robustness?
  (heavy-tailed, semi-random, oblivious noise, missing data, …)

Thank you!

Questions?