Recent Advances in High-Dimensional Robust Statistics

Ilias Diakonikolas (UW Madison)

ICML 2020 Tutorial

July 2020

PART II: BASIC ALGORITHMIC TECHNIQUES

Outline

Part II

- Basics: Sample Complexity of Robust Estimation
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering

Outline

Part II

- **Basics: Sample Complexity of Robust Estimation**
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering

Basics of High-Dimensional Mean Estimation

Fact: Let $X_1, \ldots, X_N$ be IID samples from $\mathcal{N}(\mu, I)$. The empirical estimator $\widehat{\mu}$ satisfies $\|\widehat{\mu} - \mu\|_2 = O(\sqrt{d/N})$ with probability at least 9/10.
Moreover, this rate is optimal for any estimator.

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,

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

For lower bound, apply Assouad's lemma for $\mathcal{M} = \{\mu_j = \pm\sqrt{N}/d, j \in [d]\}$ ∎

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 $d_{\mathrm{TV}}(P_1, P_2) = \epsilon/(1-\epsilon)$. 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 $d_{\mathrm{TV}}\left(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)\right) \leq |\mu_1 - \mu_2|/2$
- Same argument gives:
  - For sub-gaussian distributions: $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$
  - For bounded variance distributions: $\Omega(\sqrt{\epsilon})$

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: $\|x\|_2 = \max_{v:\|v\|_2=1} |v \cdot x|$

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

Proof:

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

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 *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

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

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 median satisfies $\delta = O(\epsilon)$ with probability at least $1 - \tau$ using $O((1/\epsilon^2)\log(1/\tau))$ samples.

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

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

■

OUTLINE

Part II

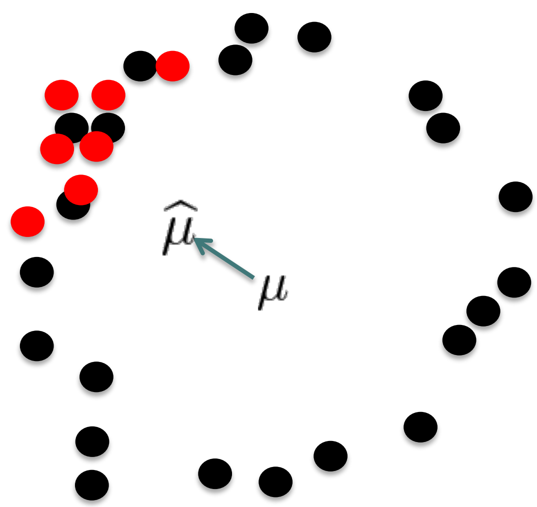
- Basics: Sample Complexity of Robust Estimation
- **Certificate of Robustness**
- Recursive Dimension Halving
- Iterative Filtering

Certificate Of Robustness For Empirical Estimator

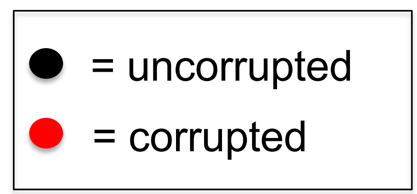
Idea #1 [DKKLMS'16, LRV'16]: If the empirical covariance is "close to what it should be", then the empirical mean works.

Certificate For Empirical Mean

Detect when the empirical estimator *may* be compromised



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



There is *no* direction of large empirical variance

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

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

with high probability we have:

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

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

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

with high probability we have:

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

in **strong** contamination model.

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

Remarks on Key Lemma

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

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

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

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

If D satisfies $\Sigma \preceq I$:

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

OUTLINE

Part II

- Basics: Sample Complexity of Robust Estimation
- Certificate of Robustness
- **Recursive Dimension Halving**
- Iterative Filtering

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

Recursive Dimension-Halving [LRV'16]

LRV Procedure:

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

Combine Results.

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

Finding A Good Subspace (I)

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

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

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

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

Question: How do we find such a subspace?

Finding A Good Subspace (II)

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

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

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

Recursive Dimension-Halving Algorithm [LRV'16]

Algorithm works as follows:

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

Error Analysis:

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

OUTLINE

Part II

- Basics: Sample Complexity of Robust Estimation
- Certificate of Robustness
- Recursive Dimension Halving
- **Iterative Filtering**

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

ITERATIVE FILTERING [DKKLMS'16]

Iterative Two-Step Procedure:

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

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

Iterate on "cleaner" dataset.

General recipe that works in general settings.

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

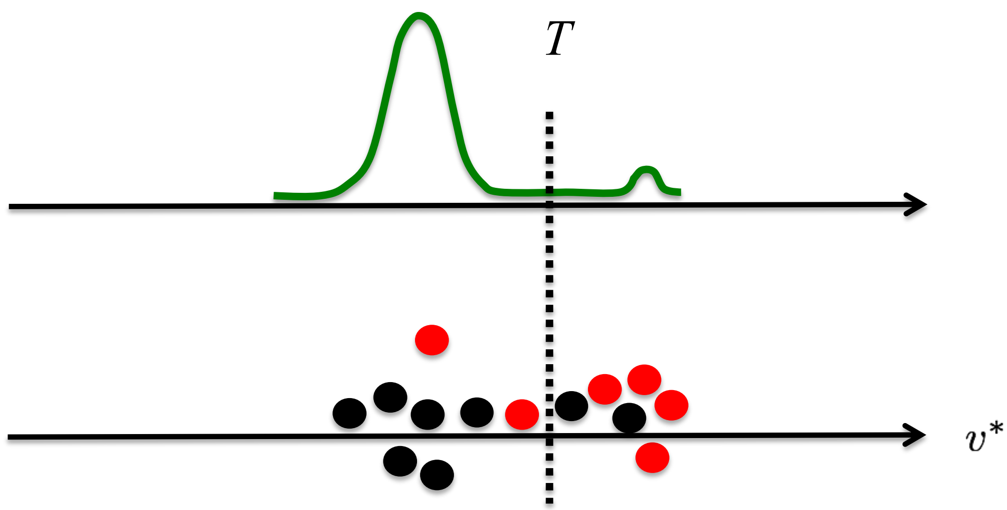
Filtering Subroutine

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

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

Let v^* be the direction of maximum variance.



FILTERING SUBROUTINE

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

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

Let v^* be the direction of maximum variance.

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

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

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

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

Eventually the empirical mean works

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

FILTERING PSEUDO-CODE

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

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

OR robust estimate of the unknown mean μ

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

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

5. **Return**

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

Remarks on Filtering Method(s)

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

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

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

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

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

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

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.