Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers

Jacob Steinhardt, Moses Charikar, Gregory Valiant

ITCS 2018
January 14, 2018
Motivation: Robust Learning

Question

What concepts can be learned robustly, even if some data is arbitrarily corrupted?
Example: Mean Estimation

Problem

Given data $x_1, \ldots, x_n \in \mathbb{R}^d$, of which $(1 - \epsilon)n$ come from $p^*$ (and remaining $\epsilon n$ are arbitrary outliers), estimate mean $\mu$ of $p^*$. 
Example: Mean Estimation

Problem

Given data $x_1, \ldots, x_n \in \mathbb{R}^d$, of which $(1 - \epsilon)n$ come from $p^*$ (and remaining $\epsilon n$ are arbitrary outliers), estimate mean $\mu$ of $p^*$. 
Example: Mean Estimation

Problem

Given data $x_1, \ldots, x_n \in \mathbb{R}^d$, of which $(1 - \epsilon)n$ come from $p^*$ (and remaining $\epsilon n$ are arbitrary outliers), estimate mean $\mu$ of $p^*$. 
Example: Mean Estimation

Problem

Given data $x_1, \ldots, x_n \in \mathbb{R}^d$, of which $(1 - \epsilon)n$ come from $p^*$ (and remaining $\epsilon n$ are arbitrary outliers), estimate mean $\mu$ of $p^*$. 

![Diagram showing data points with obvious outliers highlighted. Blue dots represent normal data points, red crosses represent outliers, and green circled points indicate estimated mean location.]
Example: Mean Estimation

Problem

Given data $x_1, \ldots, x_n \in \mathbb{R}^d$, of which $(1 - \epsilon)n$ come from $p^*$ (and remaining $\epsilon n$ are arbitrary outliers), estimate mean $\mu$ of $p^*$.

Issue: high dimensions
Mean Estimation: Gaussian Example

Suppose clean data is Gaussian:

\[ x_i \sim \mathcal{N}(\mu, I) \]

Gaussian mean \( \mu \)

variance 1 each coord.
Mean Estimation: Gaussian Example

Suppose clean data is Gaussian:

\[ x_i \sim \mathcal{N}(\mu, I) \]

Gaussian mean \( \mu \)

variance 1 each coord.

\[ \|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d} \]
Suppose clean data is Gaussian:

\[ x_i \sim \mathcal{N}(\mu, I) \]

Gaussian mean \( \mu \)

variance 1 each coord.

\[ \|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d} \]
Mean Estimation: Gaussian Example

Suppose clean data is Gaussian:

\[ x_i \sim \mathcal{N}(\mu, I) \]

Gaussian mean \( \mu \)

variance 1 each coord.

\[ \sqrt{d} \]

\[ \epsilon \sqrt{d} \]

\[ \|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d} \]

Cannot filter independently even if know true density!
History

Progress in high dimensions only recently:

- Tukey median [1975]: robust but NP-hard
- Donoho estimator [1982]: high error
- [DKKLMS16, LRV16]: first dimension-independent error bounds
History

Progress in high dimensions only recently:

- Tukey median [1975]: robust but NP-hard
- Donoho estimator [1982]: high error
- [DKKLMS16, LRV16]: first dimension-independent error bounds
- large body of work since then [CSV17, DKKLMS17, L17, DBS17]
- many other problems including PCA [XCM10], regression [NTN11], classification [FHKP09], etc.
This Talk

Question

What general and simple properties enable robust estimation?
What general and simple properties enable robust estimation?

New information-theoretic criterion: *resilience*. 
Resilience

Suppose \( \{x_i\}_{i \in S} \) is a set of points in \( \mathbb{R}^d \).

**Definition (Resilience)**

A set \( S \) is \((\sigma, \epsilon)\)-resilient in a norm \( \| \cdot \| \) around a point \( \mu \) if for all subsets \( T \subseteq S \) of size at least \((1 - \epsilon)|S|\),

\[
\left\| \frac{1}{|T|} \sum_{i \in T} (x_i - \mu) \right\| \leq \sigma.
\]

**Intuition:** all large subsets have similar mean.
Main Result

Let $S \subseteq \mathbb{R}^d$ be a set of $(1 - \epsilon)n$ “good” points.

Let $S_{\text{out}}$ be a set of $\epsilon n$ arbitrary outliers.

We observe $\tilde{S} = S \cup S_{\text{out}}$.

**Theorem**

If $S$ is $(\sigma, \frac{\epsilon}{1 - \epsilon})$-resilient around $\mu$, then it is possible to output $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq 2\sigma$.

In fact, outputting the center of any resilient subset of $\tilde{S}$ will work!
Pigeonhole Argument

Claim: If $S$ and $S'$ are $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient around $\mu$ and $\mu'$ and have size $(1-\epsilon)n$, then $\|\mu - \mu'\| \leq 2\sigma$. 
Pigeonhole Argument

Claim: If \( S \) and \( S' \) are \((\sigma, \frac{\epsilon}{1-\epsilon})\)-resilient around \( \mu \) and \( \mu' \) and have size \((1-\epsilon)n\), then \( \|\mu - \mu'\| \leq 2\sigma \).

Proof:
Pigeonhole Argument

Claim: If $S$ and $S'$ are $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient around $\mu$ and $\mu'$ and have size $(1-\epsilon)n$, then $\|\mu - \mu'\| \leq 2\sigma$.

Proof:
Pigeonhole Argument

Claim: If $S$ and $S'$ are $(\sigma, \frac{\epsilon}{1-\epsilon})$-resilient around $\mu$ and $\mu'$ and have size $(1-\epsilon)n$, then $\|\mu - \mu'\| \leq 2\sigma$.

Proof:

- Let $\mu_{S \cap S'}$ be the mean of $S \cap S'$.
- By Pigeonhole, $|S \cap S'| \geq \frac{\epsilon}{1-\epsilon} |S'|$. 
**Pigeonhole Argument**

**Claim:** If \( S \) and \( S' \) are \((\sigma, \frac{\epsilon}{1-\epsilon})\)-resilient around \( \mu \) and \( \mu' \) and have size \((1 - \epsilon)n\), then \( \| \mu - \mu' \| \leq 2\sigma \).

**Proof:**

- Let \( \mu_{S \cap S'} \) be the mean of \( S \cap S' \).
- By Pigeonhole, \( |S \cap S'| \geq \frac{\epsilon}{1-\epsilon} |S'| \).
- Then \( \| \mu' - \mu_{S \cap S'} \| \leq \sigma \) by resilience.
- Similarly, \( \| \mu - \mu_{S \cap S'} \| \leq \sigma \).
- Result follows by triangle inequality.
Implication: Mean Estimation

Lemma

If a dataset has bounded covariance, it is \((\epsilon, O(\sqrt{\epsilon}))\)-resilient (in the \(\ell_2\)-norm).
Lemma

If a dataset has bounded covariance, it is \((\epsilon, O(\sqrt{\epsilon}))\)-resilient (in the \(\ell_2\)-norm).

Proof: If \(\epsilon n\) points \(\gg 1/\sqrt{\epsilon}\) from mean, would make variance \(\gg 1\). Therefore, deleting \(\epsilon n\) points changes mean by at most \(\approx \epsilon \cdot 1/\sqrt{\epsilon} = \sqrt{\epsilon}\).
Implication: Mean Estimation

**Lemma**

If a dataset has bounded covariance, it is $(\epsilon, O(\sqrt{\epsilon}))$-resilient (in the $\ell_2$-norm).

**Proof:** If $\epsilon n$ points $\gg 1/\sqrt{\epsilon}$ from mean, would make variance $\gg 1$.

Therefore, deleting $\epsilon n$ points changes mean by at most $\approx \epsilon \cdot 1/\sqrt{\epsilon} = \sqrt{\epsilon}$.

**Corollary**

If the **clean data** has bounded covariance, its mean can be estimated to $\ell_2$-error $O(\sqrt{\epsilon})$ in the presence of $\epsilon n$ outliers.
Implication: Mean Estimation

Lemma

If a dataset has bounded covariance, it is \((\epsilon, O(\sqrt{\epsilon}))\)-resilient (in the \(\ell_2\)-norm).

Proof: If \(\epsilon n\) points \(\gg 1/\sqrt{\epsilon}\) from mean, would make variance \(\gg 1\).

Therefore, deleting \(\epsilon n\) points changes mean by at most \(\approx \epsilon \cdot 1/\sqrt{\epsilon} = \sqrt{\epsilon}\).

Corollary

If the clean data has bounded \(k\)th moments, its mean can be estimated to \(\ell_2\)-error \(O(\epsilon^{1-1/k})\) in the presence of \(\epsilon n\) outliers.
Implication: Learning Discrete Distributions

Suppose we observe samples from a distribution $\pi$ on $\{1, \ldots, m\}$.

Samples come in $r$-tuples, which are either all good or all outliers.
Implication: Learning Discrete Distributions

Suppose we observe samples from a distribution \( \pi \) on \( \{1, \ldots, m\} \).

Samples come in \( r \)-tuples, which are either all good or all outliers.

**Corollary**

The distribution \( \pi \) can be estimated (in TV distance) to error \( O(\epsilon \sqrt{\log(1/\epsilon)/r}) \) in the presence of \( \epsilon n \) outliers.
Implication: Learning Discrete Distributions

Suppose we observe samples from a distribution $\pi$ on $\{1, \ldots, m\}$.

Samples come in $r$-tuples, which are either all good or all outliers.

**Corollary**

The distribution $\pi$ can be estimated (in TV distance) to error $O(\epsilon \sqrt{\log(1/\epsilon)/r})$ in the presence of $\epsilon n$ outliers.

- follows from resilience in $\ell_1$-norm
- see also [Qiao & Valiant, 2018] later in this session!
A Majority of Outliers

Can also handle the case where clean set has size only $\alpha n$ ($\alpha < \frac{1}{2}$):
A Majority of Outliers

Can also handle the case where clean set has size only $\alpha n$ ($\alpha < \frac{1}{2}$):

- cover $\tilde{S}$ by resilient sets

$\tilde{S}$

$S$

• cover $\tilde{S}$ by resilient sets
Can also handle the case where clean set has size only $\alpha n$ ($\alpha < \frac{1}{2}$):

- cover $\tilde{S}$ by resilient sets
- at least one set $S'$ must have high overlap with $S$...
A Majority of Outliers

Can also handle the case where clean set has size only $\alpha n$ ($\alpha < \frac{1}{2}$):

- cover $\tilde{S}$ by resilient sets
- at least one set $S'$ must have high overlap with $S$...
- ...and hence $\|\mu' - \mu\| \leq 2\sigma$ as before.
A Majority of Outliers

Can also handle the case where clean set has size only $\alpha n$ ($\alpha < \frac{1}{2}$):

- cover $\tilde{S}$ by resilient sets
- at least one set $S'$ must have high overlap with $S$...
- ...and hence $\|\mu' - \mu\| \leq 2\sigma$ as before.
- Recovery in list-decodable model [BBV08].
Implication: Stochastic Block Models

Set of $\alpha n$ good and $(1 - \alpha)n$ bad vertices.
Set of $\alpha n$ good and $(1 - \alpha)n$ bad vertices.

- good $\leftrightarrow$ good: dense (avg. deg. = $a$)
- good $\leftrightarrow$ bad: sparse (avg. deg. = $b$)
Implication: Stochastic Block Models

Set of $\alpha n$ good and $(1 - \alpha)n$ bad vertices.

- good $\leftrightarrow$ good: dense (avg. deg. = $a$)
- good $\leftrightarrow$ bad: sparse (avg. deg. = $b$)
- bad $\leftrightarrow$ bad: arbitrary

Set of $\alpha n$ good and $(1 - \alpha)n$ bad vertices.

- good $\leftrightarrow$ good: dense (avg. deg. = $a$)
- good $\leftrightarrow$ bad: sparse (avg. deg. = $b$)
- bad $\leftrightarrow$ bad: arbitrary
Implication: Stochastic Block Models

Set of $\alpha n$ good and $(1 - \alpha)n$ bad vertices.

- good $\leftrightarrow$ good: dense (avg. deg. = $a$)
- good $\leftrightarrow$ bad: sparse (avg. deg. = $b$)
- bad $\leftrightarrow$ bad: arbitrary

**Question:** when can good set be recovered (in terms of $\alpha, a, b$)?
Using resilience in “truncated $\ell_1$-norm”, can show:

**Corollary**

The set of good vertices can be approximately recovered whenever

\[
\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^2}.
\]
Implication: Stochastic Block Models

Using resilience in “truncated $\ell_1$-norm”, can show:

**Corollary**

The set of good vertices can be approximately recovered whenever

$$\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^2}.$$  

Matches **Kesten-Stigum threshold** up to log factors!
Implication: Stochastic Block Models

Using resilience in “truncated \(\ell_1\)-norm”, can show:

**Corollary**

The set of good vertices can be approximately recovered whenever
\[
\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^2}.
\]

Matches **Kesten-Stigum threshold** up to log factors!

For planted clique \((a = n, b = n/2)\), recover cliques of size \(\Omega(\sqrt{n \log n})\).
- this is tight [S’17]
Algorithmic Results

Can (sometimes) turn info-theoretic into algorithmic results.
Algorithmic Results

Can (sometimes) turn info-theoretic into algorithmic results.

Most existing algorithmic results rely on bounded covariance.
Algorithmic Results

Can (sometimes) turn info-theoretic into algorithmic results.

Most existing algorithmic results rely on bounded covariance.

We show:

- for strongly convex norms, resilient sets can be “pruned” to have bounded covariance
- if injective norm is approximable, bounded covariance $\rightarrow$ efficient algorithm with $\sqrt{\varepsilon}$ error
- both true for $\ell_p$-norms! ($p \in [2, \infty]$)
Algorithmic Results

Can (sometimes) turn info-theoretic into algorithmic results.

Most existing algorithmic results rely on bounded covariance.

We show:

- for strongly convex norms, resilient sets can be “pruned” to have bounded covariance
- if injective norm is approximable, bounded covariance $\rightarrow$ efficient algorithm with $\sqrt{\epsilon}$ error
- both true for $\ell_p$-norms! ($p \in [2, \infty]$)

See [Li, 2017] and [Du, Balakrishnan, & Singh, 2017] for a non-$\ell_p$-norm.
Other Results

Finite-sample bounds

Extension to SVD
Summary

Information-theoretic criterion yielding (tight?) robust recovery bounds.
  • based on simple pigeonhole arguments
Summary

Information-theoretic criterion yielding (tight?) robust recovery bounds.
  • based on simple pigeonhole arguments

Benefit: from \textit{statistical} problem to \textit{algorithmic} problem.
Summary

Information-theoretic criterion yielding (tight?) robust recovery bounds.
  • based on simple pigeonhole arguments

Benefit: from statistical problem to algorithmic problem.

Open questions:
  • resilience for other problems (e.g. regression)
  • efficient algos under other assumptions
  • matching lower bounds?