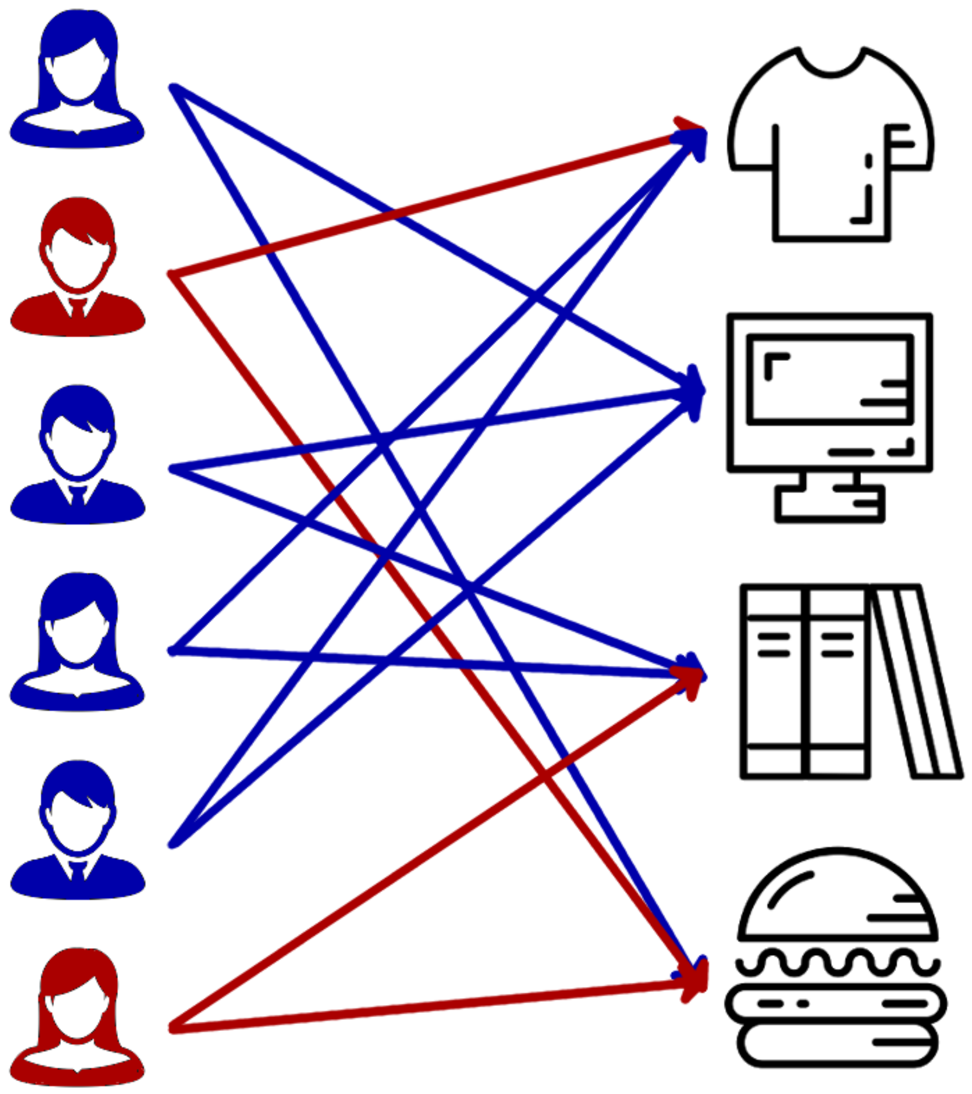
Learning from Untrusted Data

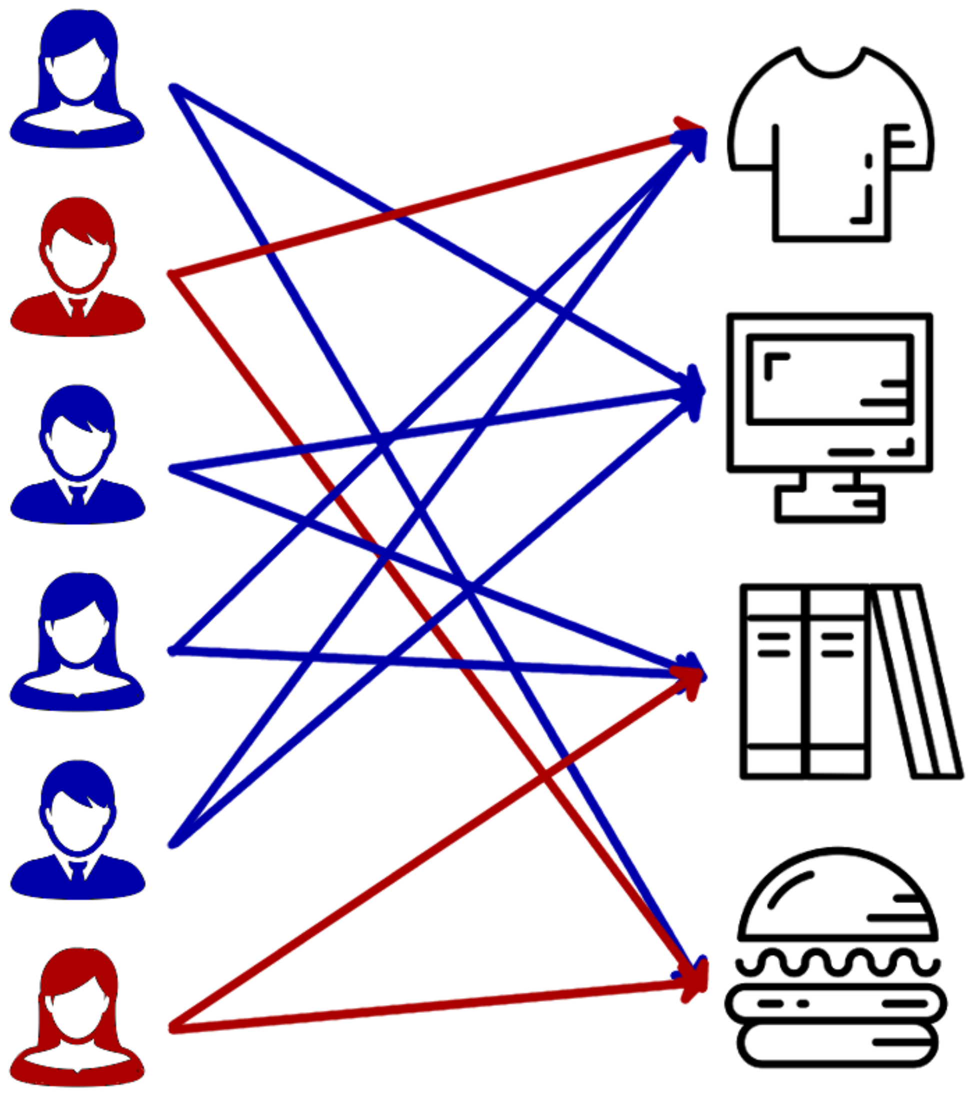
Moses Charikar, Jacob Steinhardt, Gregory Valiant

Symposium on the Theory of Computing
June 19, 2017

Motivation: **data poisoning** attacks:



Motivation: **data poisoning** attacks:



Question: what concepts can be learned in the presence of **arbitrarily corrupted** data?

Related Work

- *60 years of work on robust statistics...*

PCA:

- XCM '10, CLMW '11, CSPW '11

Mean estimation:

- LRV '16, DKKLMS '16, DKKLMS '17, L '17, DBS '17, **S**CV '17

Regression:

- NTN '11, NT '13, CCM '13, BJK '15

Classification:

- FHKP '09, GR '09, KLS '09, ABL '14

Semi-random graphs:

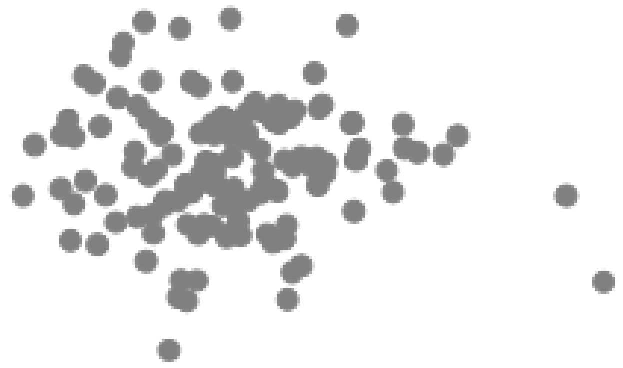
- FK '01, C '07, MMV '12, **S** '17

Other:

- HM '13, C '14, C '16, DKS '16, **S**CV '16

Problem Setting

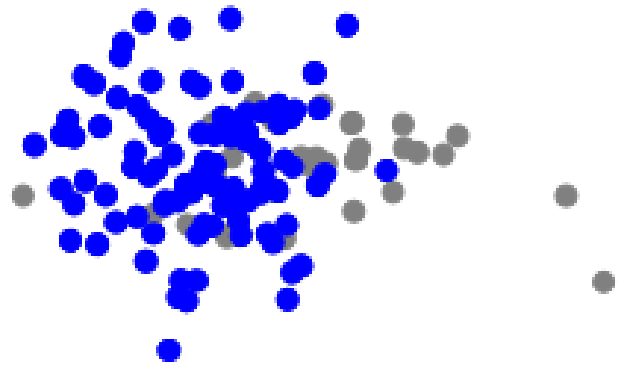
Observe n points $x_1, \ldots, x_n$



Problem Setting

Observe n points $x_1, \ldots, x_n$

Unknown subset of αn points drawn **i.i.d. from** p^*

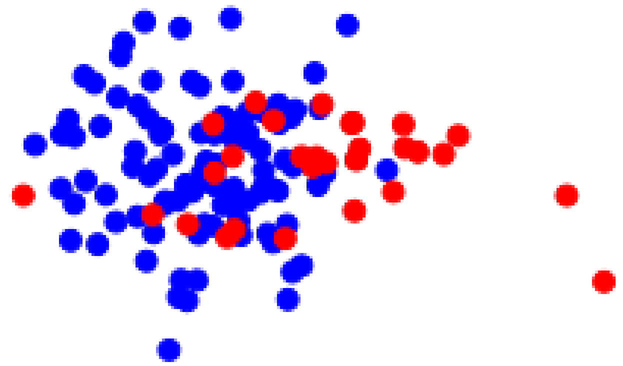


Problem Setting

Observe n points $x_1, \ldots, x_n$

Unknown subset of αn points drawn **i.i.d. from** p^*

Remaining $(1 - \alpha)n$ points are **arbitrary**

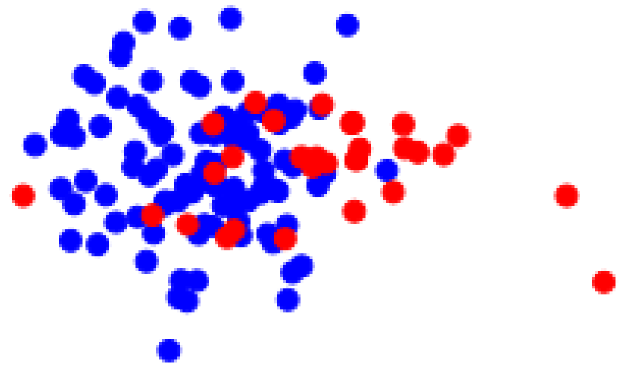


Problem Setting

Observe n points $x_1, \ldots, x_n$

Unknown subset of αn points drawn **i.i.d. from** p^*

Remaining $(1 - \alpha)n$ points are **arbitrary**



Goal: estimate parameter of interest $\theta(p^*)$

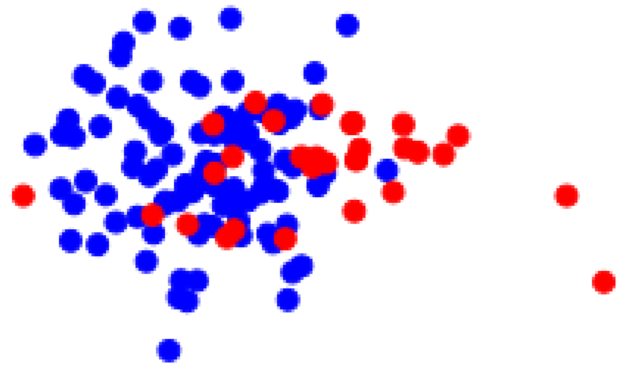
- assuming $p^* \in \mathcal{P}$ (e.g. bounded moments)
- $\theta(p^*)$ could be mean, best fit line, ranking, etc.

Problem Setting

Observe n points $x_1, \dots, x_n$

Unknown subset of αn points drawn **i.i.d. from** p^*

Remaining $(1 - \alpha)n$ points are **arbitrary**



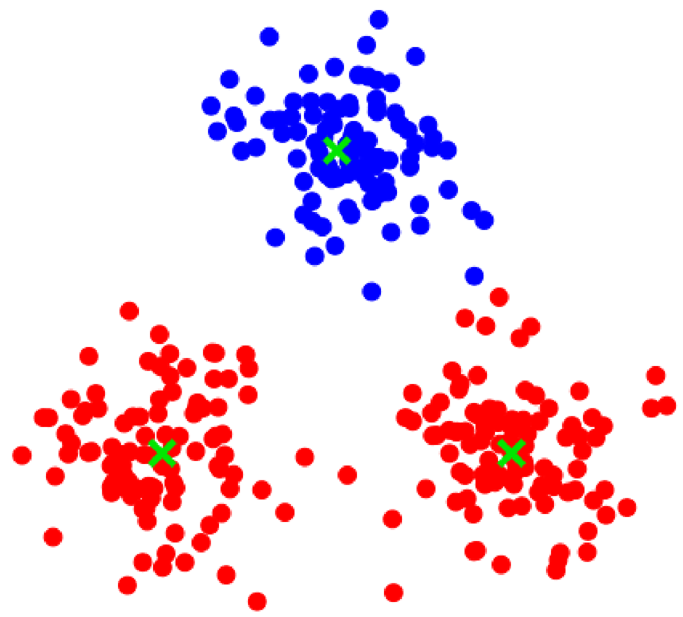
Goal: estimate parameter of interest $\theta(p^*)$

- assuming $p^* \in \mathcal{P}$ (e.g. bounded moments)
- $\theta(p^*)$ could be mean, best fit line, ranking, etc.

New regime: $\alpha \ll 1$

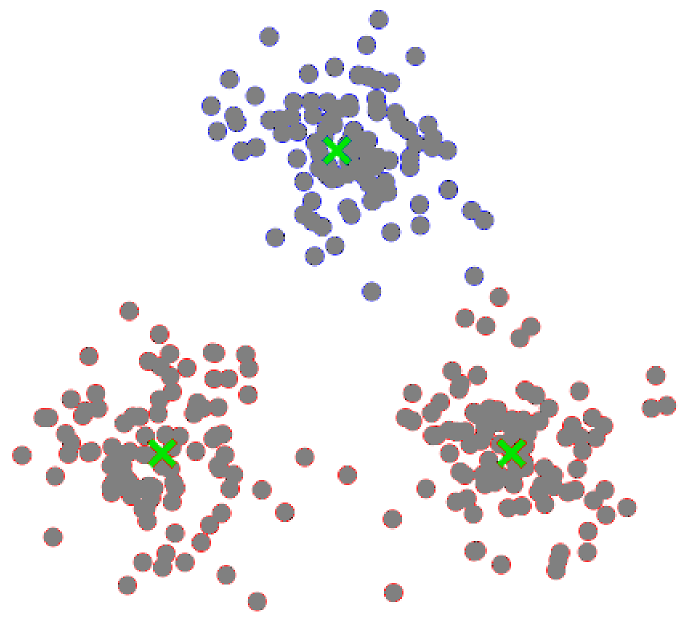
Why Is This Possible?

If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



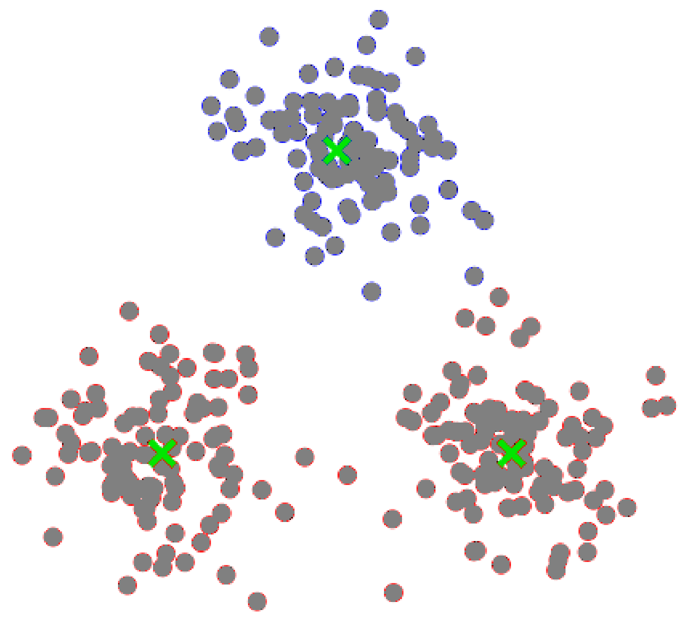
Why Is This Possible?

If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



Why Is This Possible?

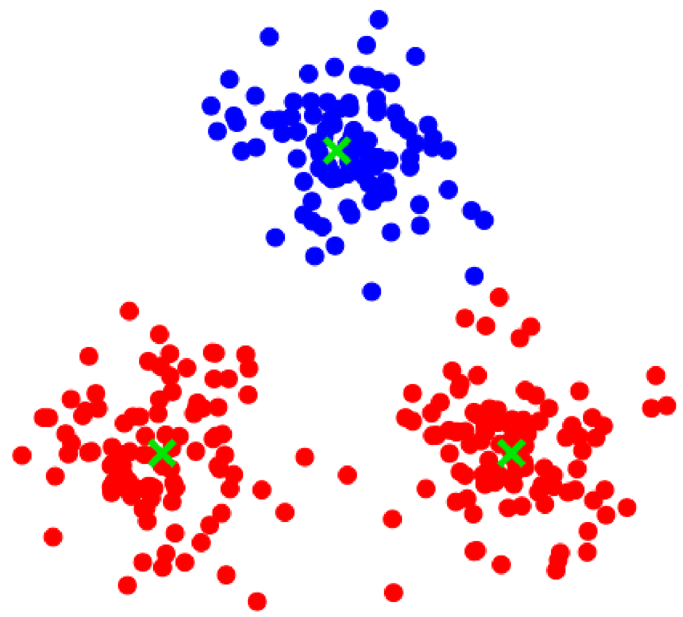
If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



But can narrow down to 3 possibilities!

Why Is This Possible?

If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



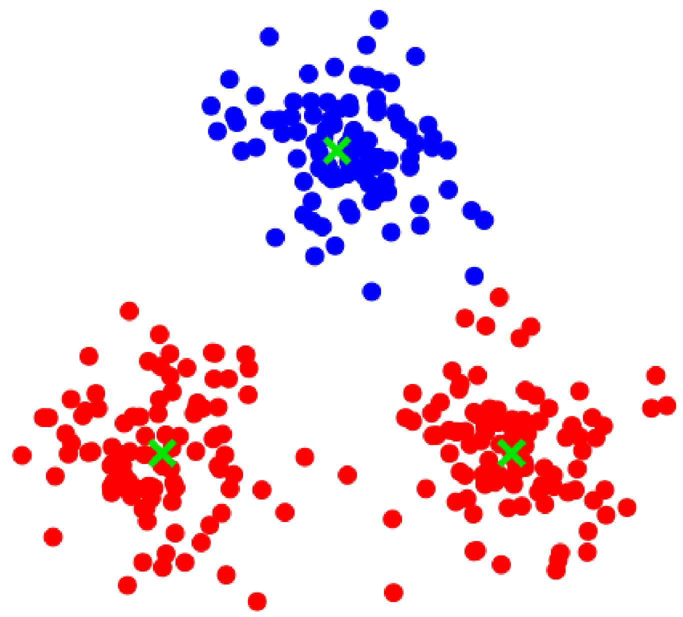
But can narrow down to 3 possibilities!

List-decodable learning [Balcan, Blum, Vempala '08]

- output $\mathcal{O}(1/\alpha)$ answers, one of which is approximately correct

Why Is This Possible?

If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



But can narrow down to 3 possibilities!

List-decodable learning [Balcan, Blum, Vempala '08]

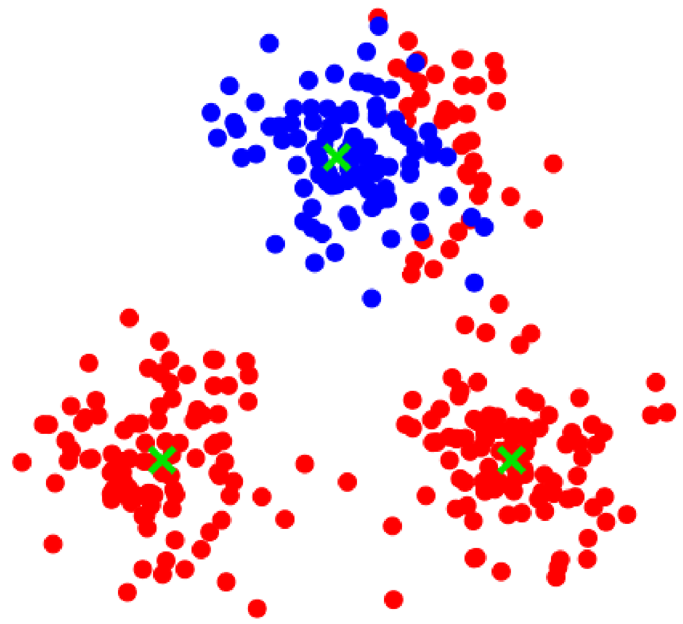
- output $\mathcal{O}(1/\alpha)$ answers, one of which is approximately correct

Semi-verified learning

- observe $\mathcal{O}(1)$ *verified* points from p^*

Why Is This Possible?

If e.g. $\alpha = \frac{1}{3}$, estimation seems impossible:



But can narrow down to 3 possibilities!

List-decodable learning [Balcan, Blum, Vempala '08]

- output $\mathcal{O}(1/\alpha)$ answers, one of which is approximately correct

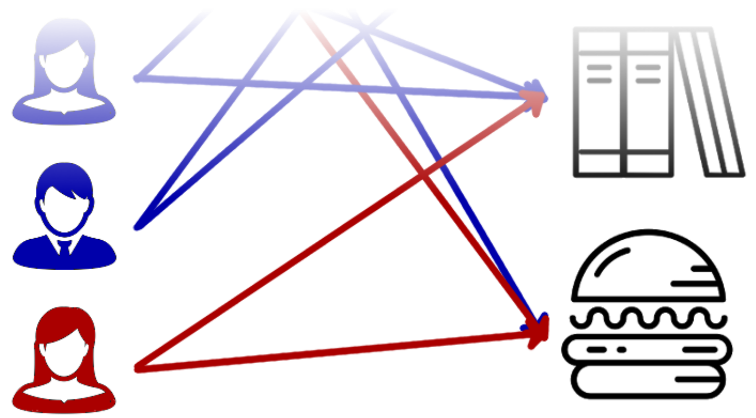
Semi-verified learning

- observe $\mathcal{O}(1)$ *verified* points from p^*

Why Care?

Practical problem: data poisoning attacks

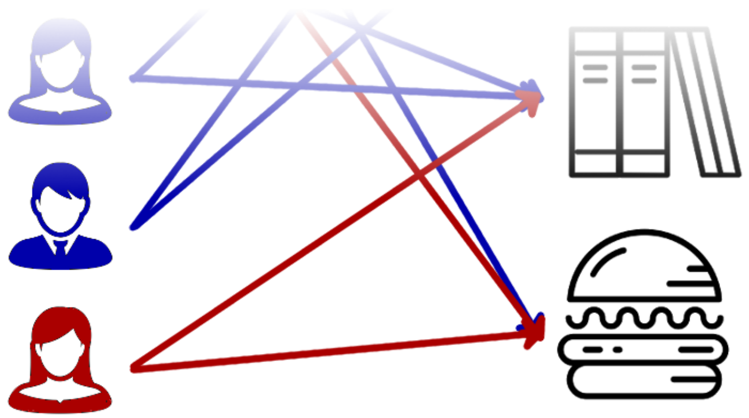
- How can we build learning algorithms that are **provably secure** to manipulation?



Why Care?

Practical problem: data poisoning attacks

- How can we build learning algorithms that are **provably secure** to manipulation?



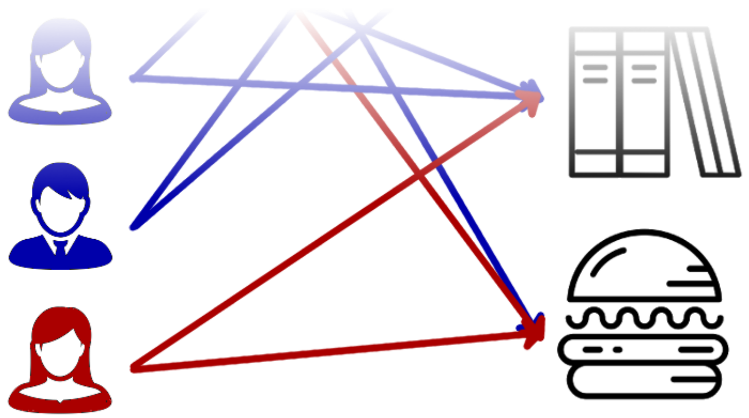
Fundamental problem in robust statistics

- What can be learned in presence of arbitrary outliers?

Why Care?

Practical problem: data poisoning attacks

- How can we build learning algorithms that are **provably secure** to manipulation?

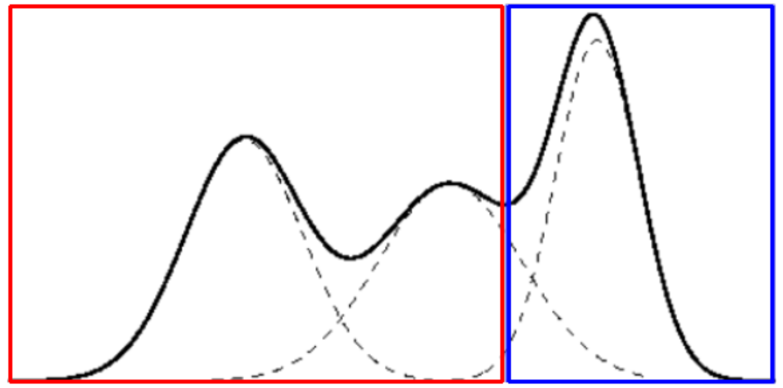


Fundamental problem in robust statistics

- What can be learned in presence of arbitrary outliers?

Agnostic learning of mixtures

- When is it possible to learn about one mixture component, with **no assumptions** about the other components?



Main Theorem

Observed functions: $f_1, \ldots, f_n$

Want to minimize unknown target function: $\bar{f}$

Main Theorem

Observed functions: $f_1, \dots, f_n$

Want to minimize unknown target function: $\bar{f}$

Key quantity: **spectral norm bound** on a subset I:

$$\frac{1}{\sqrt{|I|}} \max_{w \in \mathbb{R}^d} \|[\nabla f_i(w) - \nabla \bar{f}(w)]_{i \in I}\|_{\mathrm{op}} \le S.$$

Main Theorem

Observed functions: $f_1, \ldots, f_n$

Want to minimize unknown target function: $\bar{f}$

Key quantity: **spectral norm bound** on a subset I:

$$\frac{1}{\sqrt{|I|}} \max_{w \in \mathbb{R}^d} \|[\nabla f_i(w) - \nabla \bar{f}(w)]_{i \in I}\|_{\mathrm{op}} \leq S.$$

Meta-Theorem

Given a spectral norm bound on an unknown subset of αn functions, learning is possible:

- in the semi-verified model (for convex f_i)
- in the list-decodable model (for strongly convex f_i)

Main Theorem

Observed functions: $f_1, \dots, f_n$

Want to minimize unknown target function: $\bar{f}$

Key quantity: **spectral norm bound** on a subset I:

$$\frac{1}{\sqrt{|I|}} \max_{w \in \mathbb{R}^d} \|[\nabla f_i(w) - \nabla \bar{f}(w)]_{i \in I}\|_{\mathrm{op}} \leq S.$$

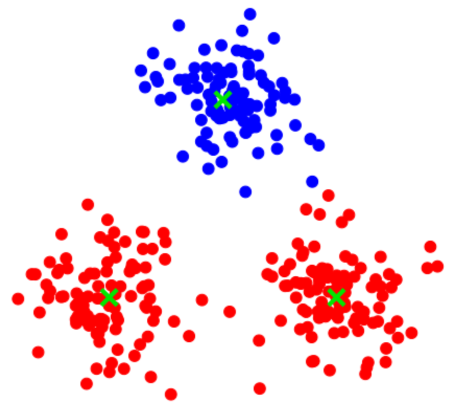
Meta-Theorem

Given a spectral norm bound on an unknown subset of αn functions, learning is possible:

- in the semi-verified model (for convex f_i)
- in the list-decodable model (for strongly convex f_i)

All results direct corollaries of meta-theorem!

Corollary: Mean Estimation



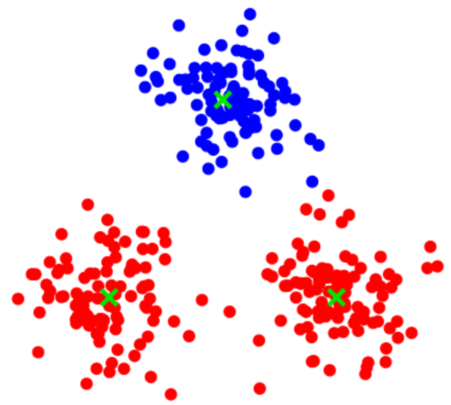
Setting: distribution p^* on $\mathbb{R}^d$ with mean μ and bounded 1st moments:

$$\mathbb{E}_{p^*}[|\langle x - \mu, v \rangle|] \leq \sigma \|v\|_2 \text{ for all } v \in \mathbb{R}^d.$$

Corollary: Mean Estimation

Setting: distribution p^* on $\mathbb{R}^d$ with mean μ and bounded 1st moments:

$$\mathbb{E}_{p^*}[|\langle x - \mu, v \rangle|] \le \sigma \|v\|_2 \text{ for all } v \in \mathbb{R}^d.$$

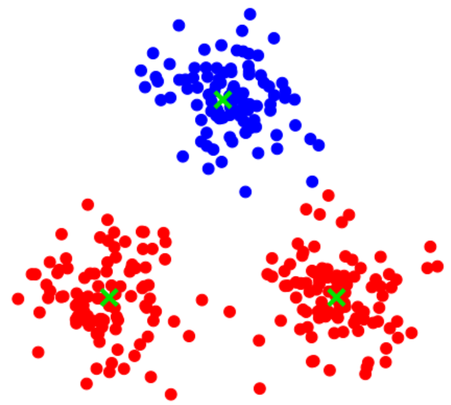


Observe αn samples from p^* and $(1 - \alpha)n$ arbitrary points, and want to estimate μ.

Corollary: Mean Estimation

Setting: distribution p^* on $\mathbb{R}^d$ with mean μ and bounded 1st moments:

$$\mathbb{E}_{p^*}[|\langle x - \mu, v\rangle|] \le \sigma \|v\|_2 \text{ for all } v \in \mathbb{R}^d.$$



Observe αn samples from p^* and $(1-\alpha)n$ arbitrary points, and want to estimate μ.

Theorem (Mean Estimation)

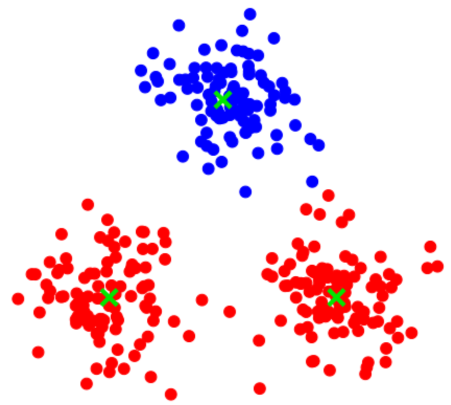
If $\alpha n \ge d$, it is possible to output estimates $\hat{\mu}_1, \dots, \hat{\mu}_m$ of the mean μ such that

- $m \le 2/\alpha$, and
- $\min_{j=1}^m \|\hat{\mu}_j - \mu\|_2 = \tilde{\mathcal{O}}(\sigma/\sqrt{\alpha})$ w.h.p.

Corollary: Mean Estimation

Setting: distribution p^* on $\mathbb{R}^d$ with mean μ and bounded 1st moments:

$$\mathbb{E}_{p^*}[|\langle x - \mu, v\rangle|] \leq \sigma \|v\|_2 \text{ for all } v \in \mathbb{R}^d.$$



Observe αn samples from p^* and $(1-\alpha)n$ arbitrary points, and want to estimate μ.

Theorem (Mean Estimation)

If $\alpha n \geq d$, it is possible to output estimates $\hat{\mu}_1, \ldots, \hat{\mu}_m$ of the mean μ such that

- $m \leq 2/\alpha$, and
- $\min_{j=1}^{m} \|\hat{\mu}_j - \mu\|_2 = \tilde{\mathcal{O}}(\sigma/\sqrt{\alpha})$ w.h.p.

Alternately, it is possible to output an estimate $\hat{\mu}$ given a single verified point from p^*.

Comparisons

Mean estimation:

	Bound	Regime	Assumption	Samples
LRV '16	$\sigma\sqrt{1-\alpha}$	$\alpha > 1 - c$	4th moments	d
DKKLMS '16	$\sigma(1-\alpha)$	$\alpha > 1 - c$	sub-Gaussian	d^3
CSV '17	$\sigma/\sqrt{\alpha}$	$\boxed{\alpha > 0}$	1st moments	d

Comparisons

Mean estimation:

	Bound	Regime	Assumption	Samples
LRV '16	$\sigma\sqrt{1-\alpha}$	$\alpha > 1-c$	4th moments	d
DKKLMS '16	$\sigma(1-\alpha)$	$\alpha > 1-c$	sub-Gaussian	d^3
CSV '17	$\sigma/\sqrt{\alpha}$	$\boxed{\alpha > 0}$	1st moments	d

Estimating mixtures:

	Separation	Robust?
AM '05	$\sigma(k + 1/\sqrt{\alpha})$	no
KK '10	σk	no
AS '12	$\sigma\sqrt{k}$	no
CSV '17	$\sigma/\sqrt{\alpha}$	**yes**

Other Results

Stochastic Block Model: (sparse regime: cf. GV '14, LLV '15, RT '15, RV '16)

	Average Degree	Robust?
GV '14	$1/\alpha^4$	no
AS '15	$1/\alpha^2$	no
CSV '17	$1/\alpha^3$	yes

Other Results

Stochastic Block Model: (sparse regime: cf. GV '14, LLV '15, RT '15, RV '16)

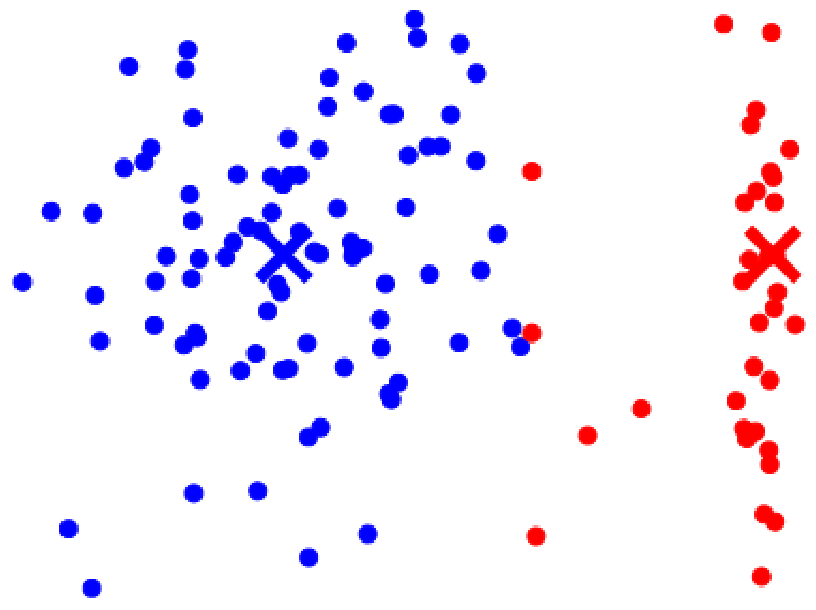
	Average Degree	Robust?
GV '14	$1/\alpha^4$	no
AS '15	$1/\alpha^2$	no
CSV '17	$1/\alpha^3$	yes

Others:

- discrete product distributions
- exponential families
- ranking

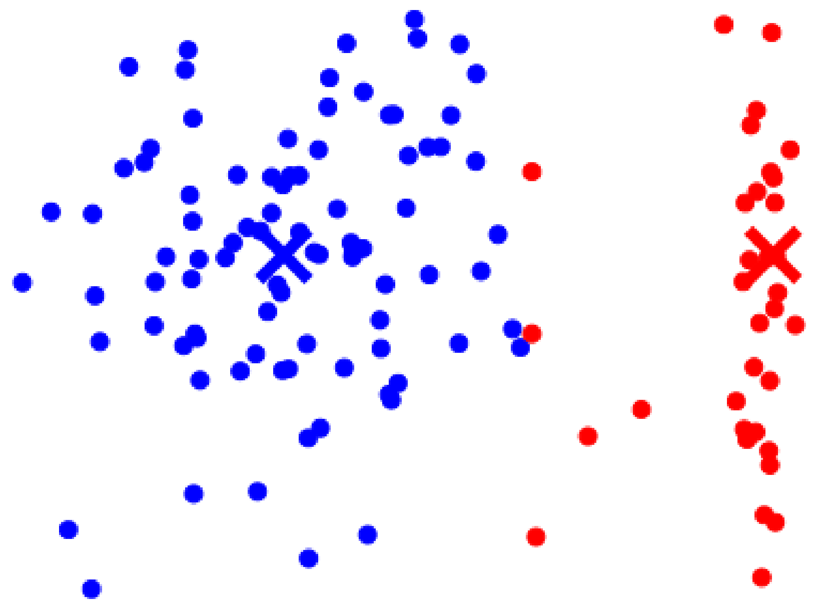
Proof Overview (Mean Estimation)

Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



Proof Overview (Mean Estimation)

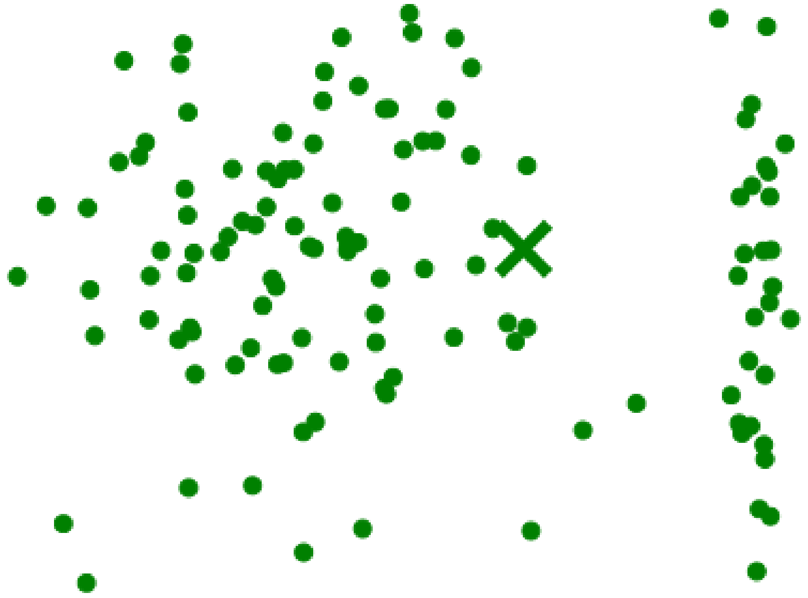
Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



Key tension: balance **adversarial** and **statistical** error

Proof Overview (Mean Estimation)

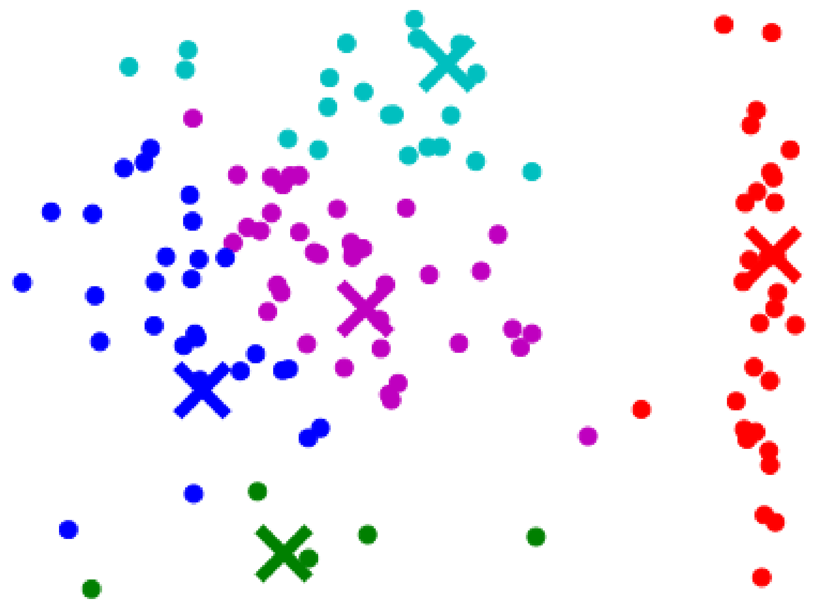
Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



Key tension: balance **adversarial** and **statistical** error

Proof Overview (Mean Estimation)

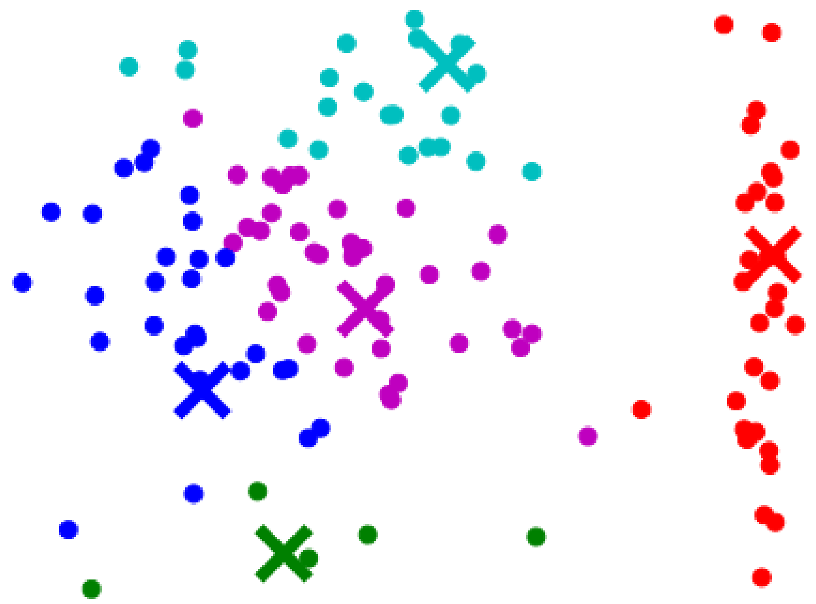
Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



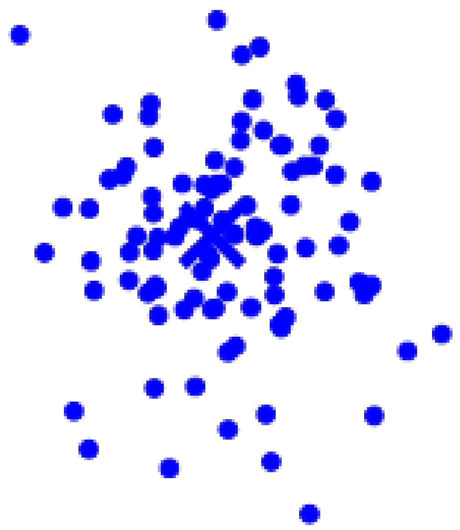
Key tension: balance **adversarial** and **statistical** error

Proof Overview (Mean Estimation)

Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



Key tension: balance **adversarial** and **statistical** error

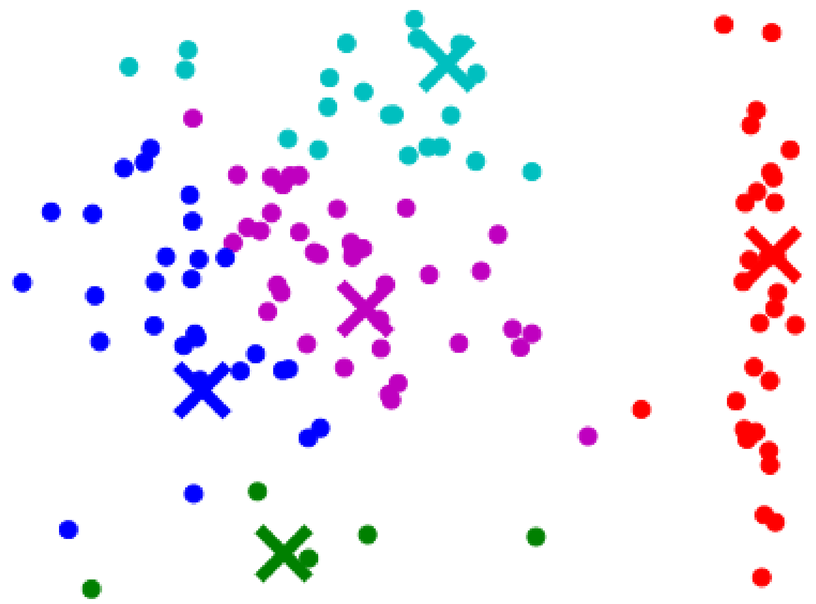


High-level strategy: solve convex optimization problem

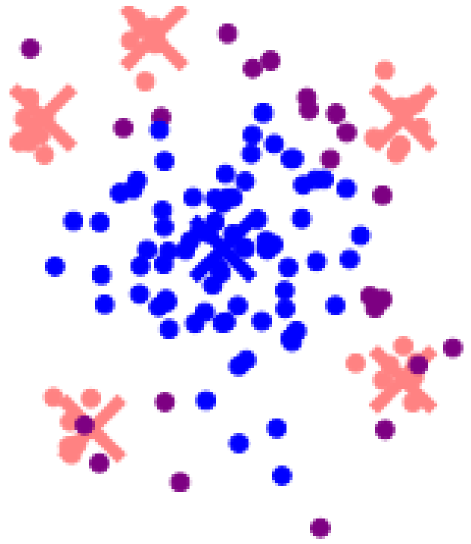
- if cost is low, estimation succeeds (spectral norm bound)
- if cost is high, identify and remove **outliers**

Proof Overview (Mean Estimation)

Recall goal: given n points $x_1, \ldots, x_n$, αn drawn from p^*, estimate mean μ of p^*



Key tension: balance **adversarial** and **statistical** error



High-level strategy: solve convex optimization problem

- if cost is low, estimation succeeds (spectral norm bound)
- if cost is high, identify and remove **outliers**

Algorithm

First pass: $\mathrm{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Algorithm

First pass: $\text{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Second pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2$

Algorithm

First pass: $\text{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Second pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2$

Final pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2 + \lambda F(\mu_1, \ldots, \mu_n)$

Algorithm

First pass: $\text{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Second pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2$

Final pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2 + \lambda F(\mu_1, \ldots, \mu_n)$

Choices for F:

- nuclear norm: error σ/α
- maximum nuclear norm over subsets: error $\sigma/\sqrt{\alpha}$ (intractable)
- minimum trace ellipsoid: error $\sigma/\sqrt{\alpha}$ (tractable)

Algorithm

First pass: $\text{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Second pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2$

Final pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2 + \lambda F(\mu_1, \ldots, \mu_n)$

Choices for F:

- nuclear norm: error σ/α
- maximum nuclear norm over subsets: error $\sigma/\sqrt{\alpha}$ (intractable)
- minimum trace ellipsoid: error $\sigma/\sqrt{\alpha}$ (tractable)

Clean-up: remove outliers, cluster the μ_i, output the cluster means

- padded decompositions [FRT '03]

Algorithm

First pass: $\text{minimize}_{\mu} \sum_{i=1}^{n} \|x_i - \mu\|_2^2$

Second pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \|x_i - \mu_i\|_2^2$

Final pass: $\text{minimize}_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} f_i(\mu_i) + \lambda F(\mu_1, \ldots, \mu_n)$

Choices for F:

- nuclear norm: error σ/α
- maximum nuclear norm over subsets: error $\sigma/\sqrt{\alpha}$ (intractable)
- minimum trace ellipsoid: error $\sigma/\sqrt{\alpha}$ (tractable)

Clean-up: remove outliers, cluster the μ_i, output the cluster means

- padded decompositions [FRT '03]

Summary

Method for robustness to **large fraction of adversarial data**

Summary

Method for robustness to **large fraction of adversarial data**

Can handle **arbitrary convex loss functions**

- based on **spectral norm bound** on gradients

Summary

Method for robustness to **large fraction of adversarial data**

Can handle **arbitrary convex loss functions**

- based on **spectral norm bound** on gradients

Strong bounds in many concrete settings

- mixtures, stochastic block model

Summary

Method for robustness to **large fraction of adversarial data**

Can handle **arbitrary convex loss functions**

- based on **spectral norm bound** on gradients

Strong bounds in many concrete settings

- mixtures, stochastic block model

Open questions:

- Can larger amounts of **verified data** yield stronger bounds?
- Can we exploit strong convexity / gradient bounds in **other norms**?
- Can we obtain guarantees in the **online setting**?

Main Theorem

Meta-Theorem

Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be a collection of κ-strongly convex functions, and let $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ an unknown target function minimized at w^*.

Suppose there is an (unknown) subset $I \subseteq [n]$ of size αn such that

$$\frac{1}{\sqrt{|I|}} \max_{w \in \mathbb{R}^d} \|[\nabla f_i(w) - \nabla \bar{f}(w)]_{i \in I}\|_{\mathrm{op}} \leq S.$$

Then, there is an algorithm outputting $m = \frac{2}{\alpha}$ candidates $\hat{w}_1, \ldots, \hat{w}_m$ such that

$$\min_{j=1}^{m} \|\hat{w}_j - w^*\|_2 = \tilde{\mathcal{O}}(S/(\kappa\sqrt{\alpha})).$$

Main Theorem

Meta-Theorem

Let $f_1, \dots, f_n : \mathbb{R}^d \to \mathbb{R}$ be a collection of κ-strongly convex functions, and let $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ an unknown target function minimized at w^*.

Suppose there is an (unknown) subset $I \subseteq [n]$ of size αn such that

$$\frac{1}{\sqrt{|I|}} \max_{w \in \mathbb{R}^d} \|[\nabla f_i(w) - \nabla \bar{f}(w)]_{i \in I}\|_{\mathrm{op}} \leq S.$$

Then, there is an algorithm outputting $m = \frac{2}{\alpha}$ candidates $\hat{w}_1, \dots, \hat{w}_m$ such that

$$\min_{j=1}^{m} \|\hat{w}_j - w^*\|_2 = \tilde{O}(S/(\kappa\sqrt{\alpha})).$$

- Can remove strong convexity (semi-verified model)