Sum-of-Squares for your Average-Case Despairs

Pravesh Kothari

Princeton/IAS

based on joint works with Boaz Barak, Siu On Chan, Sam Hopkins, Jonathan Kelner, Adam Klivans, Ryan O'Donnell, Raghu Meka, Ankur Moitra, Ryuhei Mori, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, Jacob Steinhardt, David Steurer and David Witmer.

Modeling Woes

I WANT TO ANALYZE DATA!

YOU ARE IN LUCK!
I AM A COMPUTER SCIENTIST.

Modeling Woes

I WANT TO ANALYZE DATA!

YOU ARE IN LUCK!
I AM A COMPUTER SCIENTIST.

I SUSPECT THERE'S SOME LATENT STRUCTURE IN IT.

AH! I KNOW THIS PROBLEM!
IT'S CALLED *CLUSTERING*.

Modeling Woes

I WANT TO ANALYZE DATA!

YOU ARE IN LUCK!
I AM A COMPUTER SCIENTIST.

I SUSPECT THERE'S SOME LATENT STRUCTURE IN IT.

AH! I KNOW THIS PROBLEM!
IT'S CALLED CLUSTERING.

GREAT! HOW DO I DO IT?

WELL, IN GENERAL, YOU CAN'T.
IT'S NP-HARD!

Modeling Woes

I WANT TO ANALYZE DATA!

YOU ARE IN LUCK!
I AM A COMPUTER SCIENTIST.

I SUSPECT THERE'S SOME LATENT STRUCTURE IN IT.

AH! I KNOW THIS PROBLEM!
IT'S CALLED CLUSTERING.

GREAT! HOW DO I DO IT?

WELL, IN GENERAL, YOU CAN'T.
IT'S NP-HARD!

OH, WELL, AT LEAST YOU TRIED.

Simons Institue Public Lecture 2017

Does computational complexity restrict Artificial Intelligence (AI) and Machine Learning?

Sanjeev Arora

Princeton University

(on sabbatical at Simons Institute)

(Funding: NSF, Simons Foundation, ONR)

Simons Institute Public Lecture 2017

WORST CASE

Does computational complexity restrict Artificial Intelligence (AI) and Machine Learning? MOST LIKELY NOT!

Sanjeev Arora

Princeton University

(on sabbatical at Simons Institute)

Data usually doesn't conspire against us.
So worst-case instances may not be relevant.
Lower bounds may not be limiting.

(Funding: NSF, Simons Foundation, ONR)

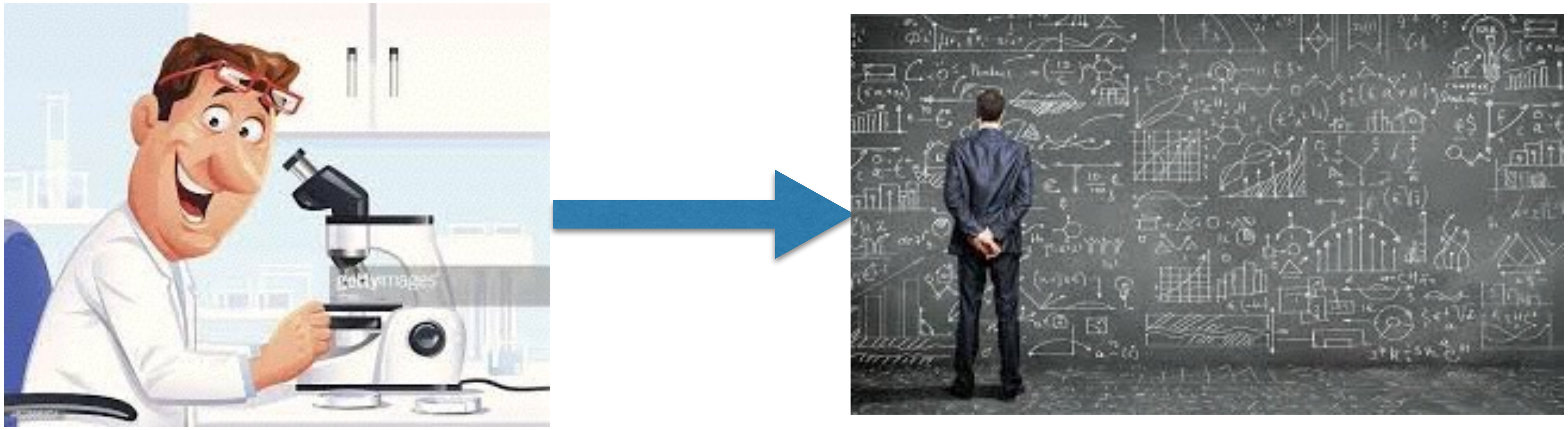
Average-Case Models

Choose a reasonable probabilistic *generation* model.

Solve typical instances according to this model.

Successful approach in Machine Learning

Average-Case Models



1. Choose a meaningful "latent variable" model = dist. family
2. Use data to learn the parameters of the model.

Latent Parameters → Model → Observed Data

Data Generation

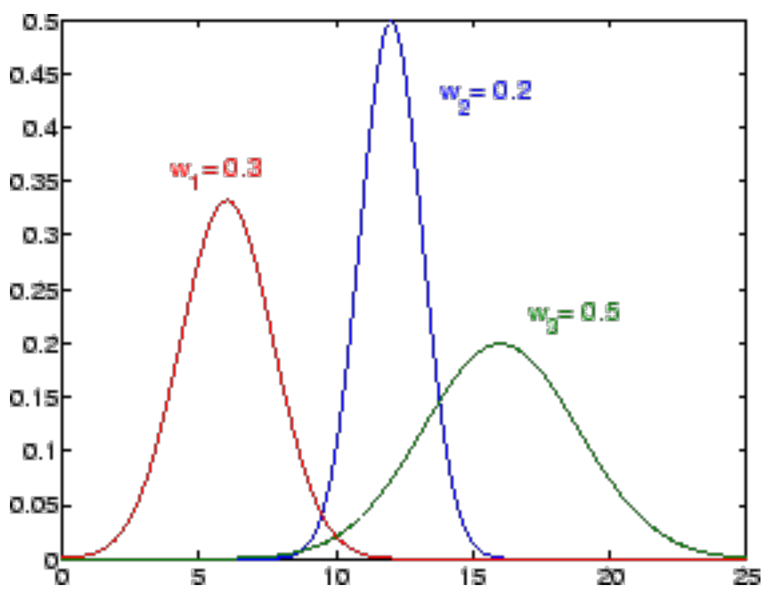
$\mu_1, \mu_2, \ldots, \mu_k$

$$\sum_i \frac{1}{k} \mathcal{N}(\mu_i, I)$$

"MIXTURE OF GAUSSIANS"

Learning

Average-Case Models



LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

HERE'S A GREAT ALGO!

Average-Case Models

First two ***moments*** { **Mean**, **Covariance** }

LET'S FIT "MIXTURE OF GAUSSIANS" MODEL TO YOUR DATA

HERE'S A GREAT ALGO!

Moments = summary of correlations of distributions.

Average-Case Models

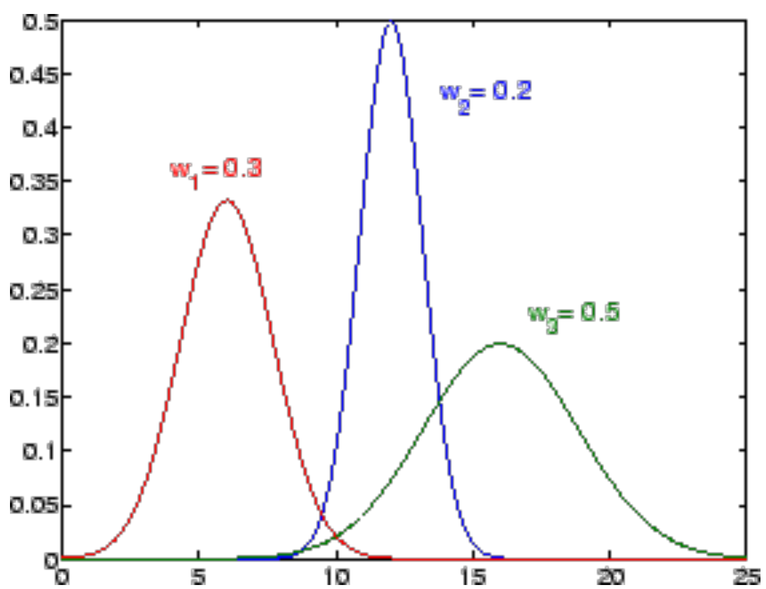
First two ***moments*** — **Mean**, **Covariance**

Learn with 3rd/4th moments!

- Cluster via ***Mixture Models***
- Fit ***Topic Models***
- Do ***Independent Component Analysis***

Moments = summary of correlations of distributions.

Average-Case Models



LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

HERE'S A GREAT ALGO!

Learn by *decomposing* **3rd** moments $\mathbb{E}[X^{\otimes 3}]$!

"method of moments" for learning latent variable models

[Pearson'1894], [Kalai-Moitra-Valiant'10], [Belkin-Sinha'10],...

Average-Case Models

LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

UH, THERE'S AN ISSUE.

Average-Case Models

LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

UH, THERE'S AN ISSUE.

THE FIT DOESN'T "GENERALIZE"

Average-Case Models

LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

UH, THERE'S AN ISSUE.

THE FIT DOESN'T "GENERALIZE"

LACK ENOUGH DATA?
CLUSTERS NOT WELL-SEPARATED?
GAUSSIAN-ASSUMPTION FLAWED?
OUTLIERS?

Average-Case Models

LET'S FIT
"MIXTURE OF GAUSSIANS"
MODEL TO YOUR DATA

UH, THERE'S AN ISSUE.

THE FIT DOESN'T "GENERALIZE"

LACK ENOUGH DATA?

CLUSTERS NOT WELL-SEPARATED?

GAUSSIAN-ASSUMPTION FLAWED?

OUTLIERS?

IS THERE A DIFFERENT ALGO?

OR SHOULD I COLLECT MORE DATA?

UMM...

Average-Case Woes

- Typically, algorithmic techniques aren't easily adaptable.
- Limited applicability of *"web-of-reductions"* for lower bounds so hard to confirm fundamental impossibility

Hopes and Dreams

A meta-algorithm for average-case algorithm design?

1. Apply to *broad class of problems* in a *canonical* way
2. Capture *power of efficient algorithms* for this class.

Hopes and Dreams

A meta-algorithm for average-case algorithm design?

1. Apply to *broad class of problems* in a *canonical* way
2. Capture *power of efficient algorithms* for this class.
3. Admit *principled strategies* for *lower bounds*
4. *Easy-to-adapt analysis* for variants (e.g. outliers)

Hopes and Dreams

A meta-algorithm for average-case algorithm design?

1. Apply to *broad class of problems* in a *canonical* way
2. Capture *power of efficient algorithms* for this class.
3. Admit *principled strategies* for *lower bounds*
4. *Easy-to-adapt analysis* for variants (e.g. outliers)

Promising Candidate: **Sum-of-Squares Method**

[Shor ’87,Grigoriev-Vorobjov ’99,Nesterov ’99,Parillo ’00,Lasserre ’00]

This Talk

Sum-of-Squares Method for Average-Case Problems.

1. **Broadly-applicable Algorithmic Approach**

   Simple Generalization/ Identifiability proof → **Efficient Learning Algorithm**

   Example: Outlier-Robust *Method of Moments*

2. **Broadly-applicable Lower Bound Approach**

   "Pseudo-calibration"

   Applications: Tight *samples* vs *time* trade-offs for *Planted-Clique, Sparse PCA, Tensor PCA, Random-CSP*…

This Talk

Sum-of-Squares Method for Average-Case Problems.

1. **Broadly-applicable Algorithmic Approach**

Simple Generalization/ Identifiability proof → **Efficient Learning Algorithm!**

Example: Outlier-Robust *Method of Moments*

2. **Broadly-applicable Lower Bound Approach**

"Pseudo-calibration"

Applications: Tight *samples* vs *time* trade-offs for *Planted-Clique, Sparse PCA, Tensor PCA, Random-CSP*…

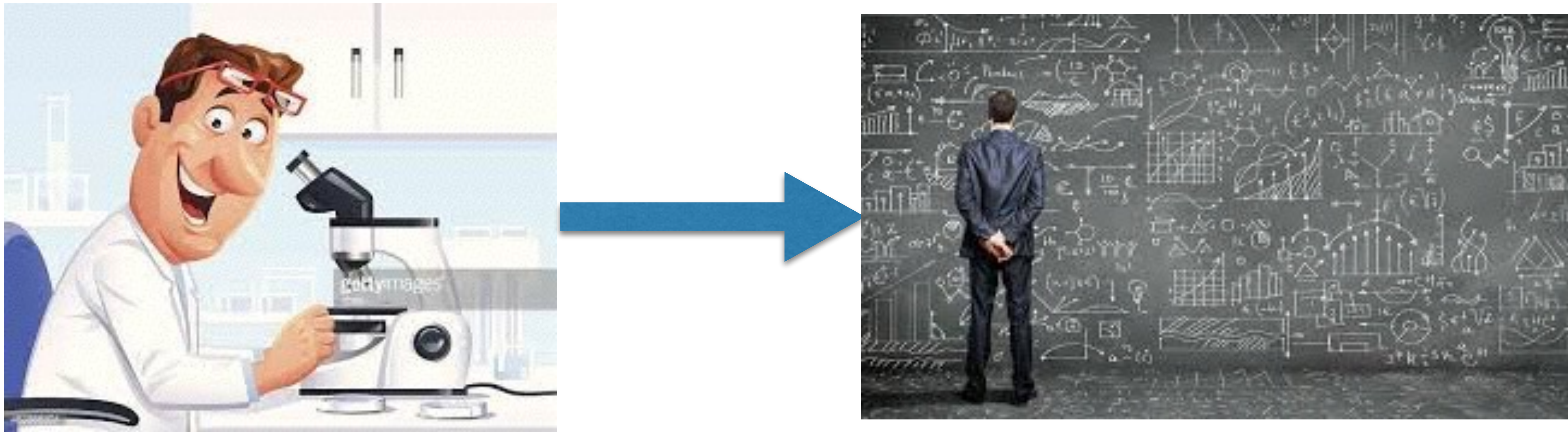
Moment Estimation

Given: i.i.d. samples from a distribution in some family **"Model"**

Goal: Accurately estimate **low-degree moments** of distribution

A basic primitive in unsupervised learning with many applications.

But data is not ideal…



Given: i.i.d. samples from a distribution in some family **"Model"**

Goal: Accurately estimate **low-degree moments** of distribution

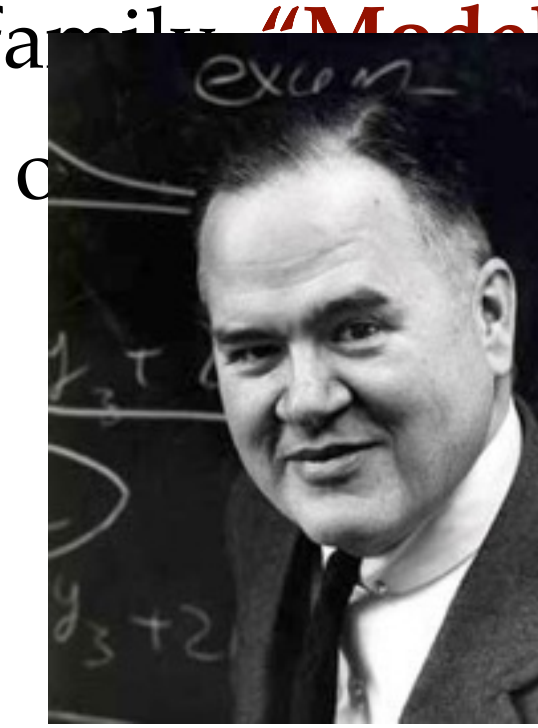
Issue

Can't assume data to be *perfectly* i.i.d.

But data is not ideal…

Given: i.i.d. samples from a distribution in some fa

Goal: Accurately estimate **low-degree moments** c



J. W. Tukey
1960s

Issue

Can't assume data to be *perfectly* i.i.d.

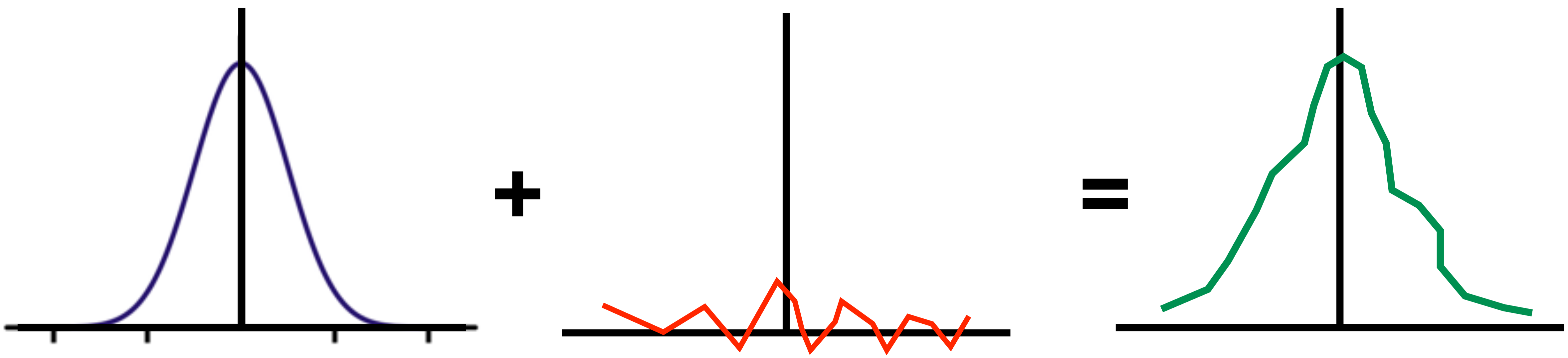
Are our learning algorithms *robust*?

Can they estimate moments from "noisy" data?

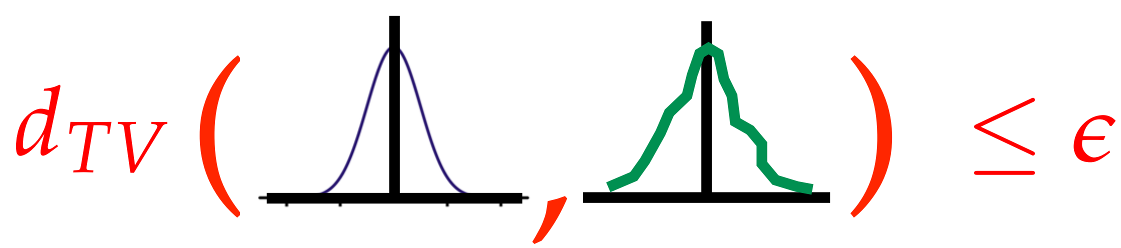
Robust Moment Estimation

Given: *ϵ-corruption* of a sample from an ideal model

Goal: accurately estimate moments of the model dist.



Ideal Model **Arbitrary Noise** **Observed Model**



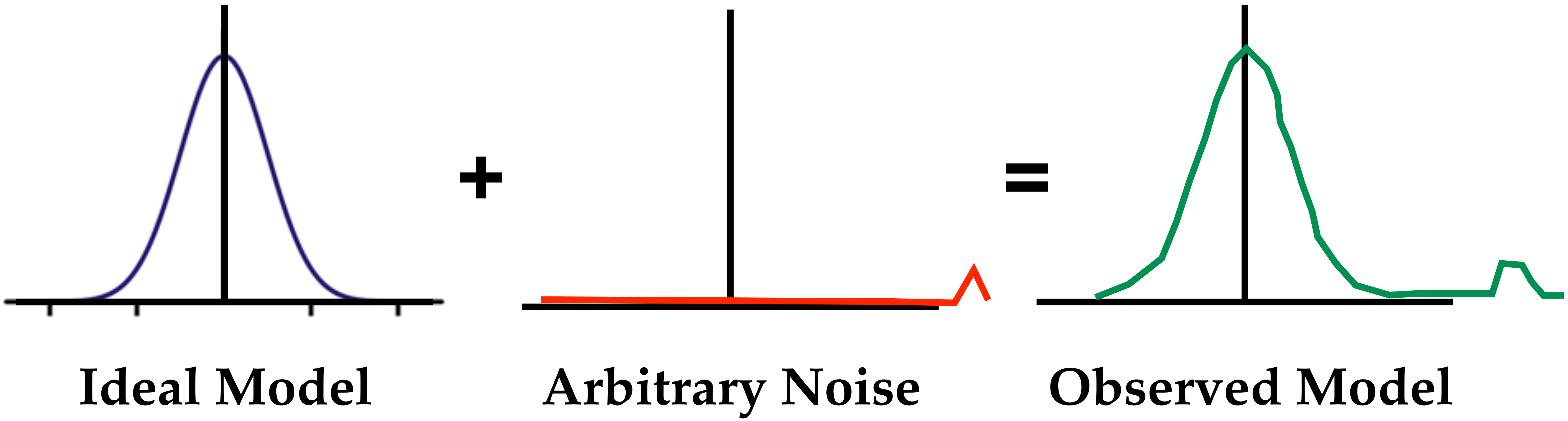
"Malicious" Noise

ϵ-fraction of the samples are adversarially corrupted

adversary can both *remove* points and *add* outliers

Robust Estimators

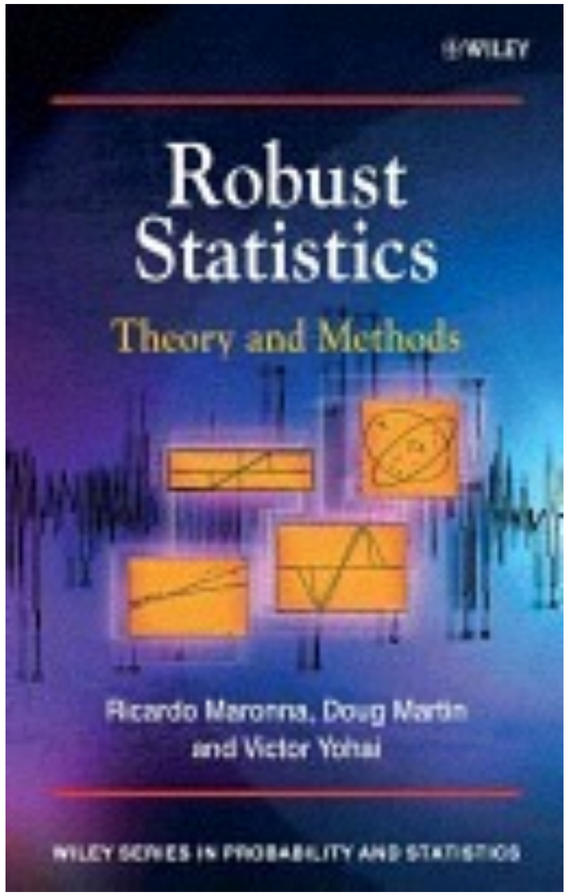
Do empirical moments work?

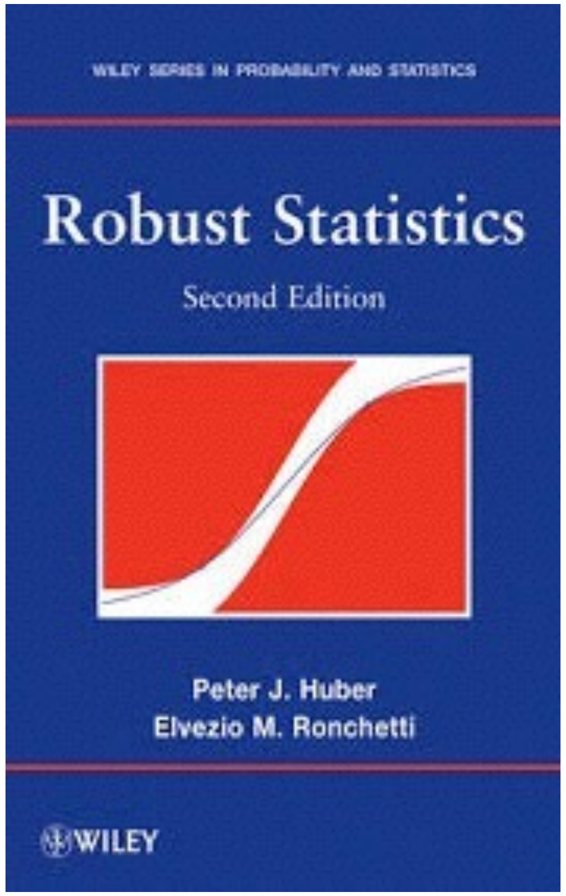


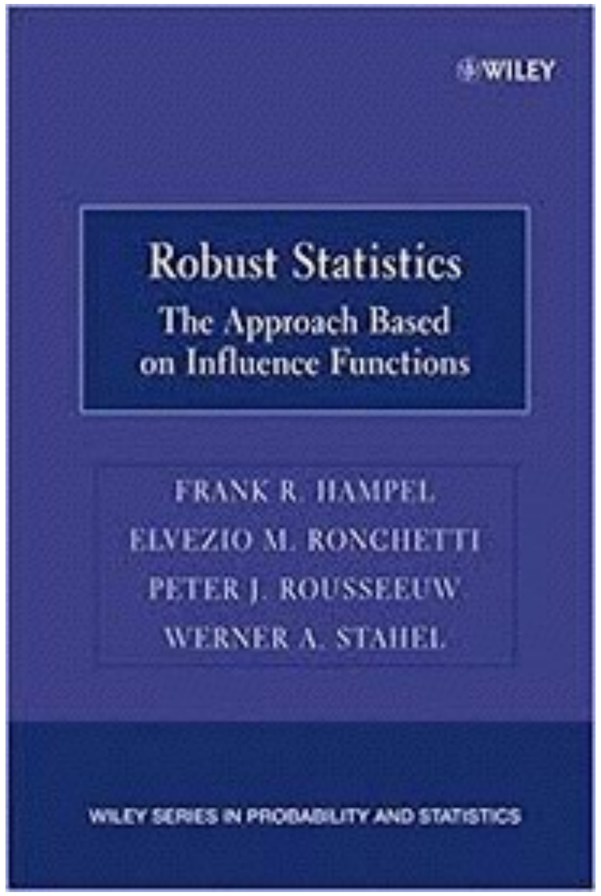
A **single** corrupted sample can arbitrarily change the empirical mean.

Method of Moments breaks under such sample corruption!

Robust Statistics







Estimators that work well in a neighborhood around the model.

Curse of Dimensionality

Typically need exponential time in dimension to compute.

Robust Moment Estimation in high dimensions?

Efficient Robust Estimation

[2016] first works on efficient robust mean/covariance estimation…

Theorem **[Lai-Rao-Vempala'16]**
[Charikar-Steinhardt-Valiant'17]
[Diakonikolas-Kamath-Kane-Lee-Moitra-Stewart'17]

Given: ϵ-*corrupted* sample from a dist. with covariance Σ.

Guarantee: $\|\hat{\mu} - \mu\| \leq O(\epsilon^{1/2}) \|\Sigma\|^{1/2}$

information theoretically optimal.

better results if unknown distribution is gaussian.

Efficient Robust Estimation

[2016] first works on efficient robust mean/covariance estimation…

Theorem **[Lai-Rao-Vempala'16]**
[Charikar-Steinhardt-Valiant'17]
[Diakonikolas-Kamath-Kane-Lee-Moitra-Stewart'17]

Given: ϵ-*corrupted* sample from a dist. with covariance Σ.

Guarantee: $\|\hat{\mu} - \mu\| \leq O(\epsilon^{1/2}) \|\Sigma\|^{1/2}$

[K-Steurer'17]

- weaker assumptions "bounded moments"
- information theoretically optimal accuracy
- trade-off niceness of model with error
- extends to higher moment estimation "injective norm guarantees"

Corollaries: outlier-robust algorithms for **ICA**, **Mixture Models**, …

Today

Simple proof to illustrate the SoS method for learning

Theorem [Lai-Rao-Vempala'16]
[Charikar-Steinhardt-Valiant'17]
[Diakonikolas-Kamath-Kane-Lee-Moitra-Stewart'17]

Given: ϵ-*corrupted* sample from a dist. with covariance Σ.

Guarantee: $\|\hat{\mu} - \mu\| \leq O(\epsilon^{1/2})\|\Sigma\|^{1/2}$

Proof from [K-Steurer'17]

- Same proof template yields all our results!
- **Algo:** return the output of a convex program
- No outlier-removal, no rounding…

Learn via SoS

Two steps to any unsupervised learning problem

Step 1 Identifiability

A finite sample **uniquely** determines the parameters of the models

Required for *any* algorithm to exist!

Yields an inefficient algorithm.

Step 2 Algorithm Design

Design an efficient algorithm for parameter recovery.

First step is usually easy.
Second step can be non-trivial.

Learn via SoS

Mechanically transform "simple" *identifiability* proofs to algorithms!

Step 1 Identifiability

A finite sample **uniquely** determines the parameters of the models

Required for *any* algorithm to exist!

Yields an inefficient algorithm.

Step 2 Algorithm Design

Design an efficient algorithm for parameter recovery.

First step is usually easy.
Second step can be non-trivial.

Learn via SoS

Mechanically transform "simple" *identifiability* proofs to algorithms!

Step 1 Identifiability

A finite sample **uniquely** determines the parameters of the models

Required for *any* algorithm to exist!

Yields an inefficient algorithm.

Many natural proof techniques are "simple"!

Step 2 Algorithm Design

Design an efficient algorithm for parameter recovery.

First step is usually easy.

Second step can be non-trivial.

Learn via SoS

Mechanically transform "simple" *identifiability* proofs to algorithms!

Step 1 Identifiability

A finite sample **uniquely** determines the parameters of the models

Required for *any* algorithm to exist!

Yields an inefficient algorithm.

Step 2 Algorithm Design

Design an efficient algorithm for parameter recovery.

First step is usually easy.
Second step can be non-trivial.

Specialization of "proofs to algorithms" paradigm to learning.

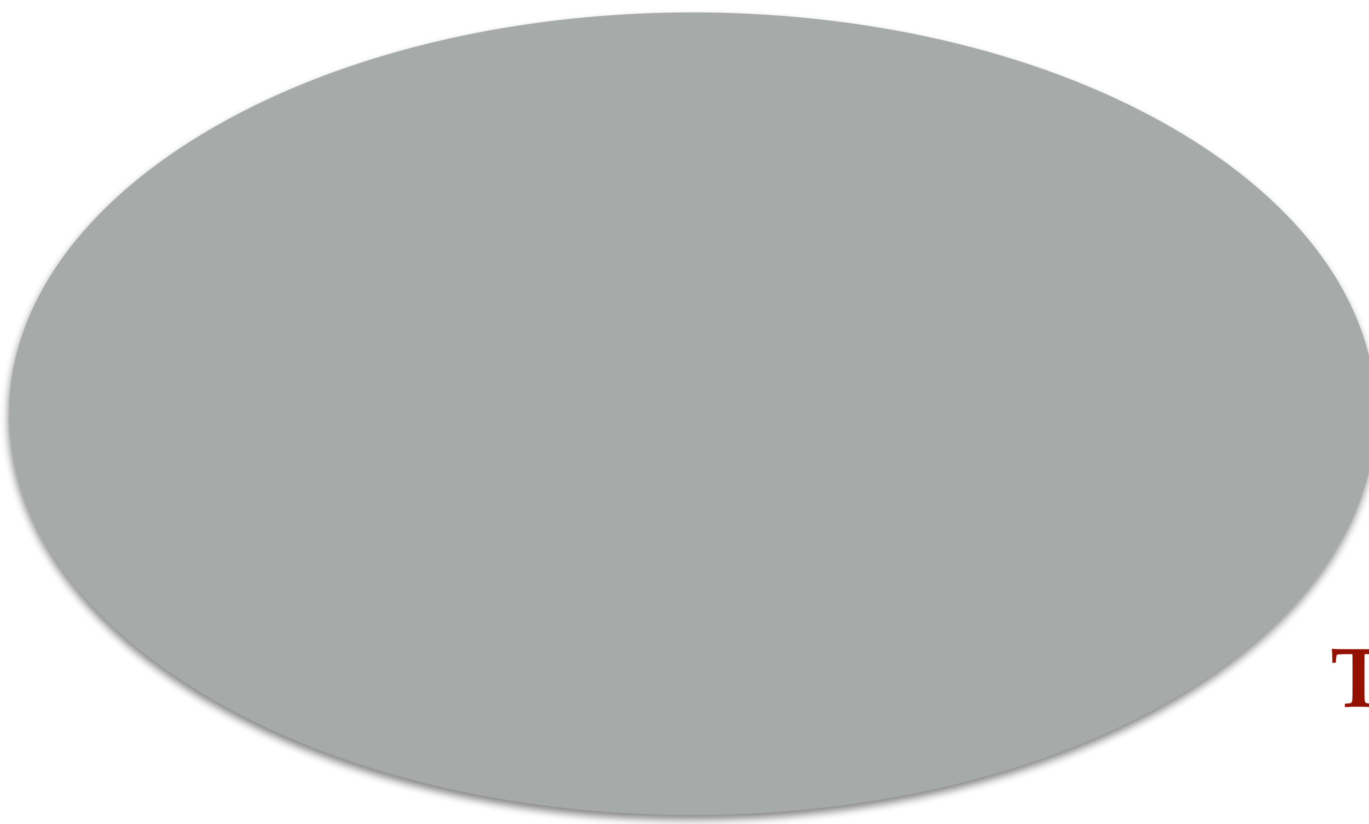
Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

*up to a small error

Identifiability for Mean Estimation

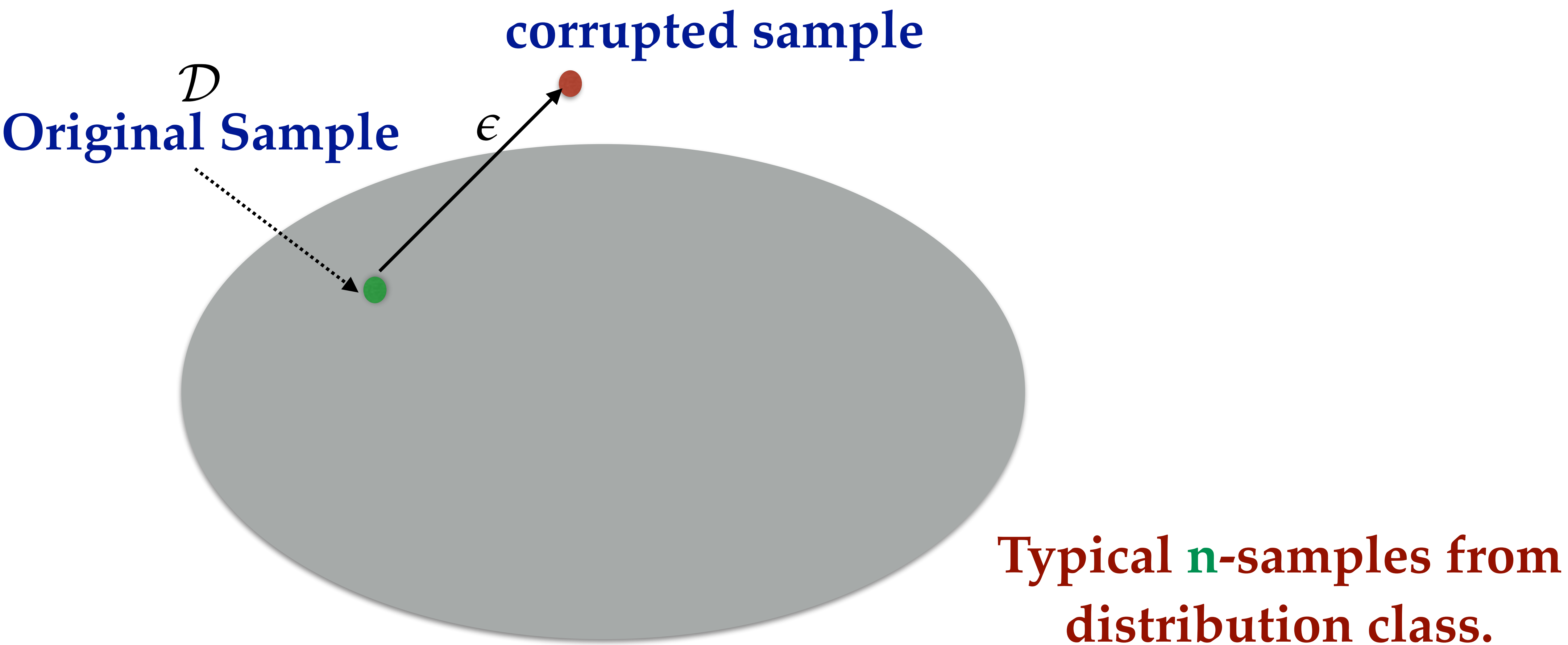
Why does a corrupted sample uniquely* determine the mean?



Typical n-samples from distribution class.

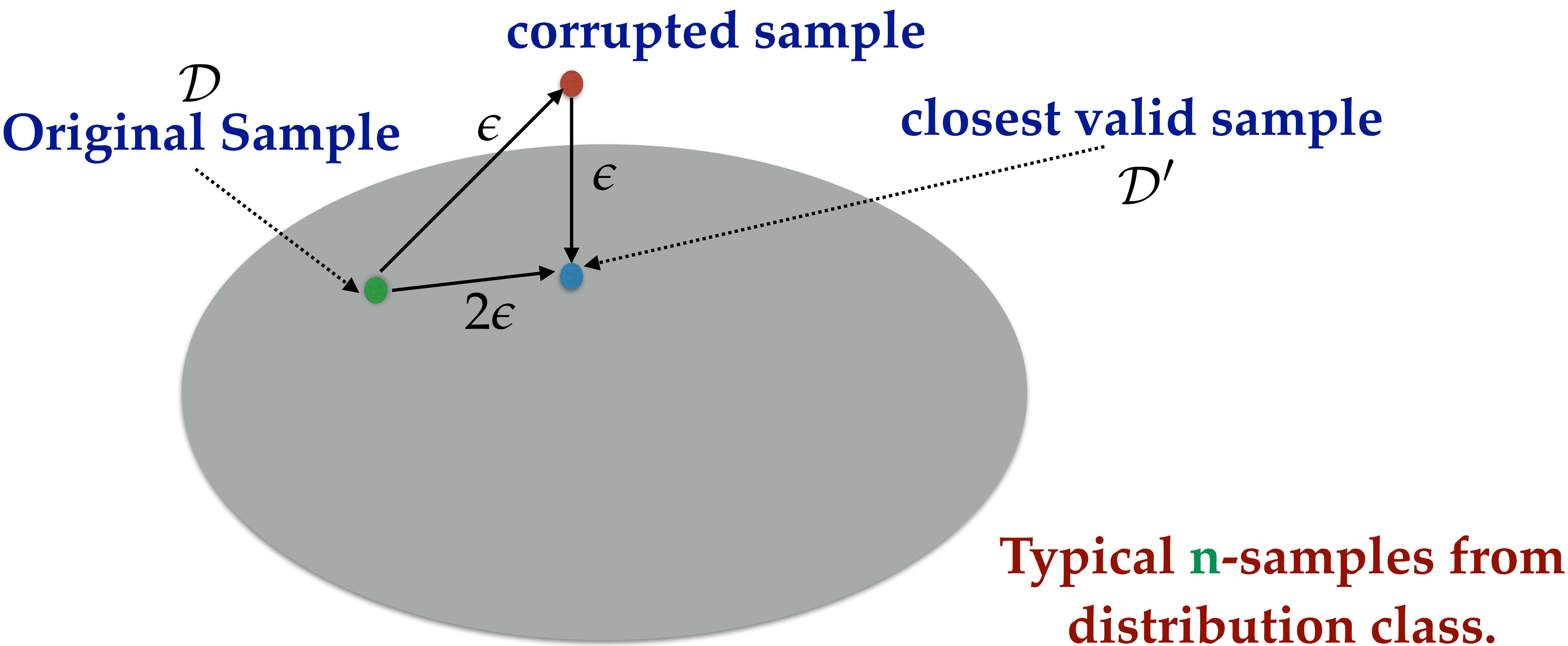
Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?



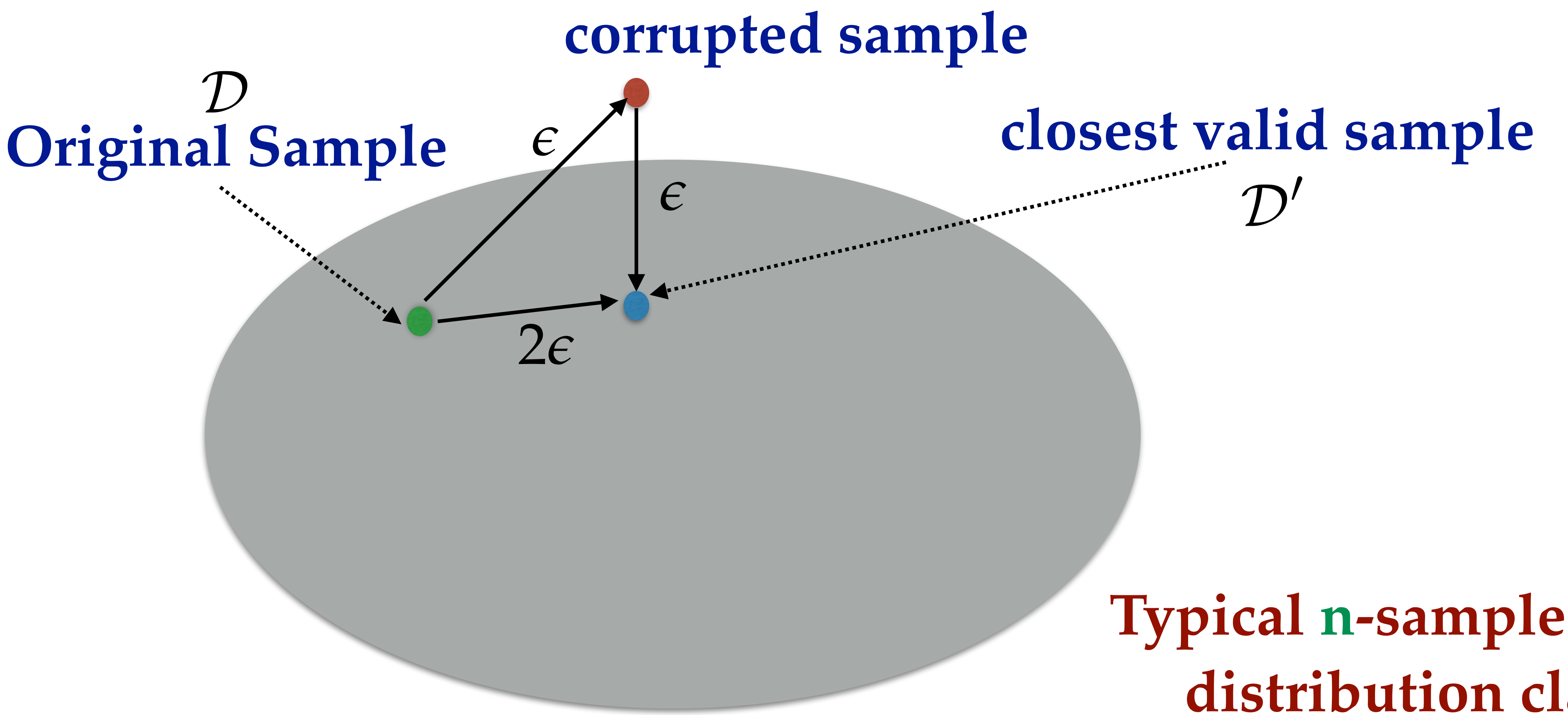
Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?



Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?



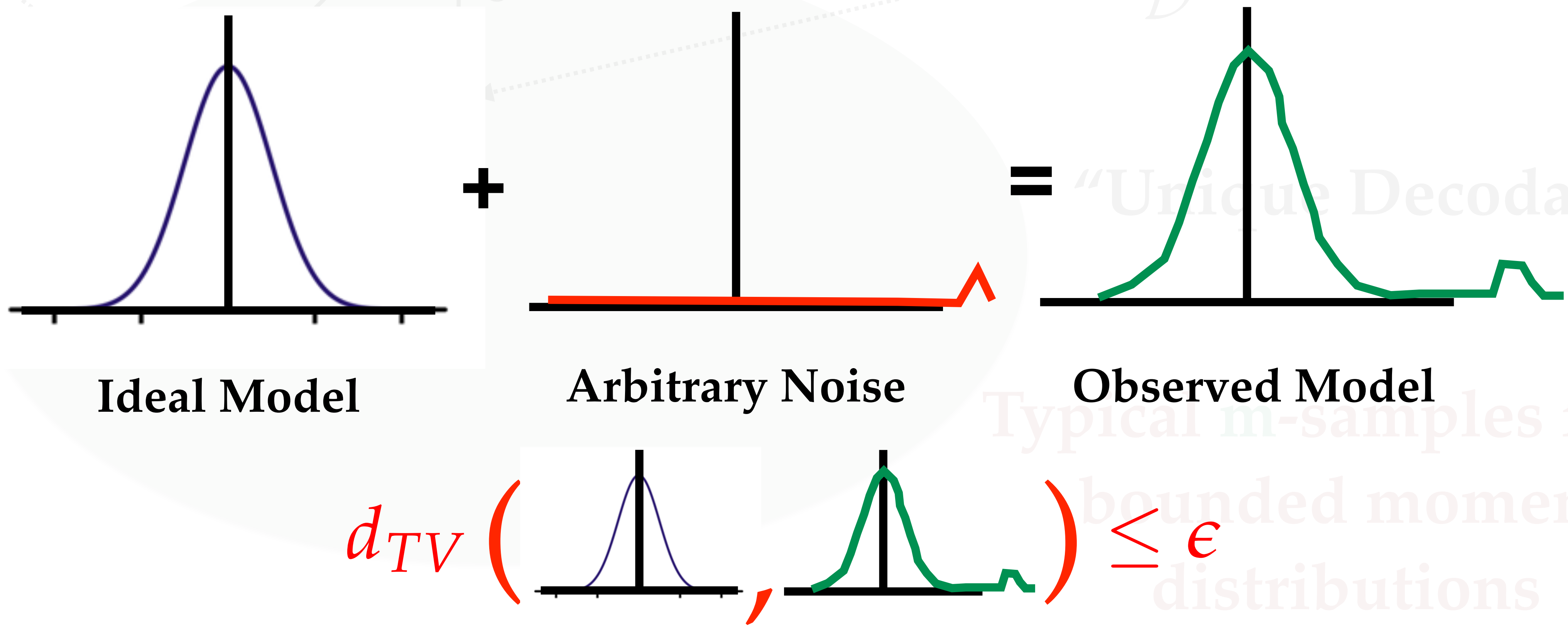
Typical n-samples from distribution class.

"Unique Decodability"

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

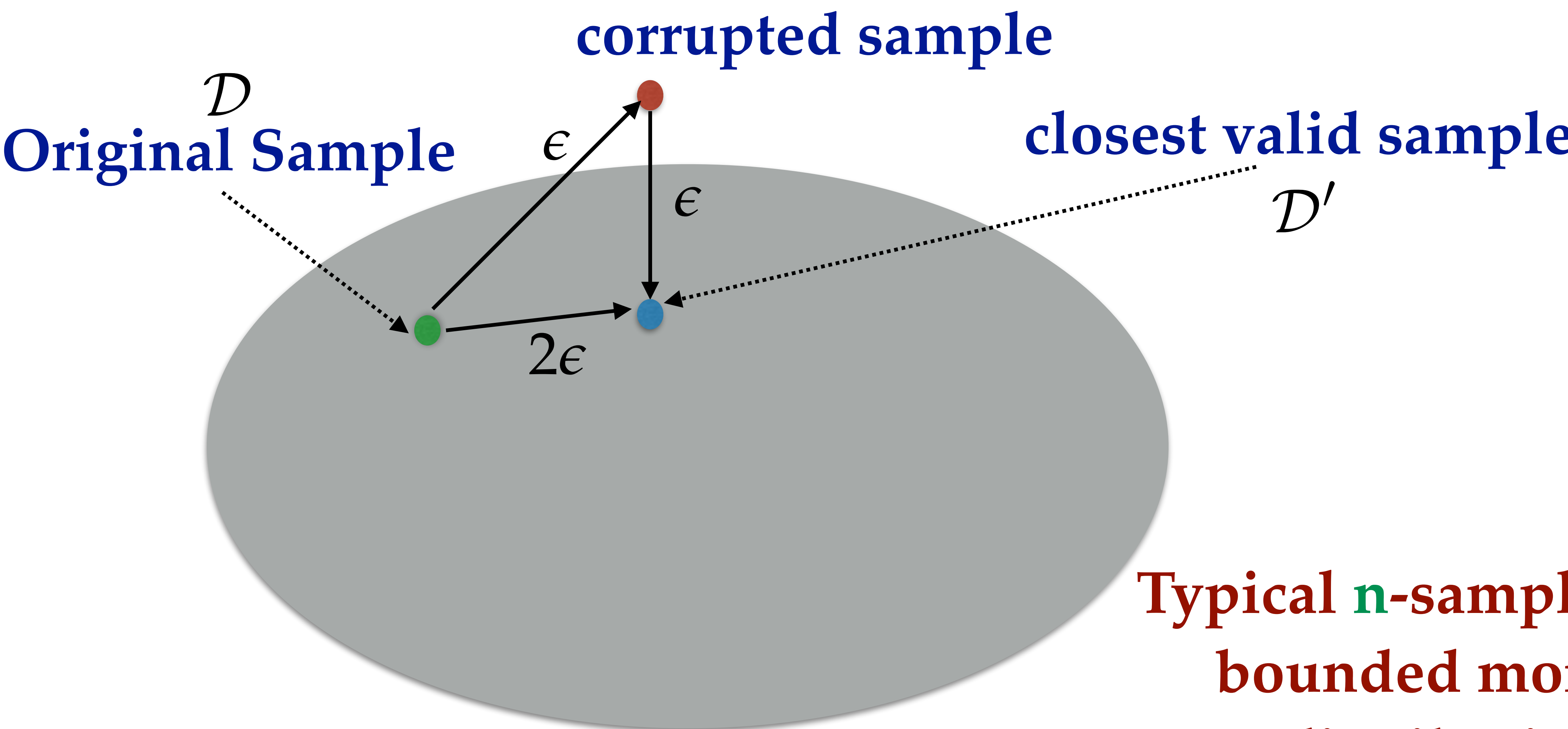
When is robust estimation possible?



Nearby dist. in the family must have close parameters!

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?



Typical n-samples from bounded moment distributions

Why do nearby samples have close parameters?

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

Lemma (Robust Identifiability of Mean)

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $X' = \{x'_1, x'_2, \ldots, x'_n\}$ be such that:

$\Pr_{i \in [n]} \{x_i \neq x'_i\} = \epsilon < 0.9$. Then,

$$\|\mu(X) - \mu(X')\| < O(\epsilon^{1/2})(\sigma_X + \sigma_{X'})$$

$$\sigma_X^2 = \|\Sigma(X)\|$$
$$\sigma_{X'}^2 = \|\Sigma(X')\|$$

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

Lemma (Robust Identifiability of Mean)

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $X' = \{x'_1, x'_2, \ldots, x'_n\}$ be such that:

$\Pr_{i \in [n]} \{x_i \neq x'_i\} = \epsilon < 0.9$. Then,

$$\|\mu(X) - \mu(X')\| < O(\epsilon^{1/2})(\sigma_X + \sigma_{X'})$$

$$\sigma_X^2 = \|\Sigma(X)\|$$
$$\sigma_{X'}^2 = \|\Sigma(X')\|$$

Inefficient Algorithm Using Identifiability

1. Find an ϵ-close sample that has the smallest covariance

2. Return its mean.

In 1-D, corresponds to modifying the largest/smallest points.

~ median

Simple proof of Lemma → convex relaxation of this algo works !

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

Lemma (Robust Identifiability of Mean)

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $X' = \{x'_1, x'_2, \ldots, x'_n\}$ be such that:

$\Pr_{i \in [n]} \{x_i \neq x'_i\} = \epsilon < 0.9$. Then,

$$\|\mu(X) - \mu(X')\| < O(\epsilon^{1/2})(\sigma_X + \sigma_{X'})$$

$$\sigma_X^2 = \|\Sigma(X)\|$$
$$\sigma_{X'}^2 = \|\Sigma(X')\|$$

Proof By Cauchy-Schwarz

$$\frac{1}{n}\sum_i \langle u, x_i - x'_i \rangle = \frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x'_i\}) \cdot \langle u, x_i - x'_i \rangle$$

$$\leq \left(\frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x'_i\})\right)^{1/2} \cdot \left(\frac{1}{n}\sum_i \langle u, x_i - x'_i \rangle\right)^{1/2}$$

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

Lemma (Robust Identifiability of Mean)

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $X' = \{x'_1, x'_2, \ldots, x'_n\}$ be such that:

$\Pr_{i \in [n]} \{x_i \neq x'_i\} = \epsilon < 0.9$. Then,

$$\|\mu(X) - \mu(X')\| < O(\epsilon^{1/2})(\sigma_X + \sigma_{X'})$$

$$\sigma_X^2 = \|\Sigma(X)\| \qquad \sigma_{X'}^2 = \|\Sigma(X')\|$$

Proof By Cauchy-Schwarz

$$\frac{1}{n}\sum_i \langle u, x_i - x'_i \rangle = \frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x'_i\}) \cdot \langle u, x_i - x'_i \rangle$$

$$\leq \left(\frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x'_i\})\right)^{1/2} \cdot \left(\frac{1}{n}\sum_i \langle u, x_i - x'_i \rangle\right)^{1/2}$$

$$\leq \epsilon^{1/2} \cdot \left(\mathbb{E}_i \langle u, x_i - \mu(X) \rangle\right) + \left(\langle u, x'_i - \mu(X') \rangle\right) + \langle u, \mu(X) - \mu(X') \rangle\right)^{1/2}$$

Identifiability for Mean Estimation

Why does a corrupted sample uniquely* determine the mean?

Lemma **(Robust Identifiability of Mean)**

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $X' = \{x_1', x_2', \ldots, x_n'\}$ be such that:

$\Pr_{i \in [n]} \{x_i \neq x_i'\} = \epsilon < 0.9$. Then,

$$\|\mu(X) - \mu(X')\| < O(\epsilon^{1/2})(\sigma_X + \sigma_{X'})$$

$$\sigma_X^2 = \|\Sigma(X)\|$$
$$\sigma_{X'}^2 = \|\Sigma(X')\|$$

Proof By Cauchy-Schwarz

$$\frac{1}{n}\sum_i \langle u, x_i - x_i'\rangle = \frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x_i'\}) \cdot \langle u, x_i - x_i'\rangle$$

$$\leq \left(\frac{1}{n}\sum_i \mathbb{1}(\{x_i \neq x_i'\})\right)^{1/2} \cdot \left(\frac{1}{n}\sum_i \langle u, x_i - x_i'\rangle\right)^{1/2}$$

$$\leq O(\epsilon^{1/2})(\sigma_X + \sigma_{X'} + |\langle u, \mu(X) - \mu(X')\rangle|^{1/2})$$

Rearrange to get the lemma!

Efficient Algorithm?

"Lemma"

There's a magic box to mechanically convert such proofs into efficient algorithms based on semi-definite programming.

SDP relaxation for the following quadratic program works!

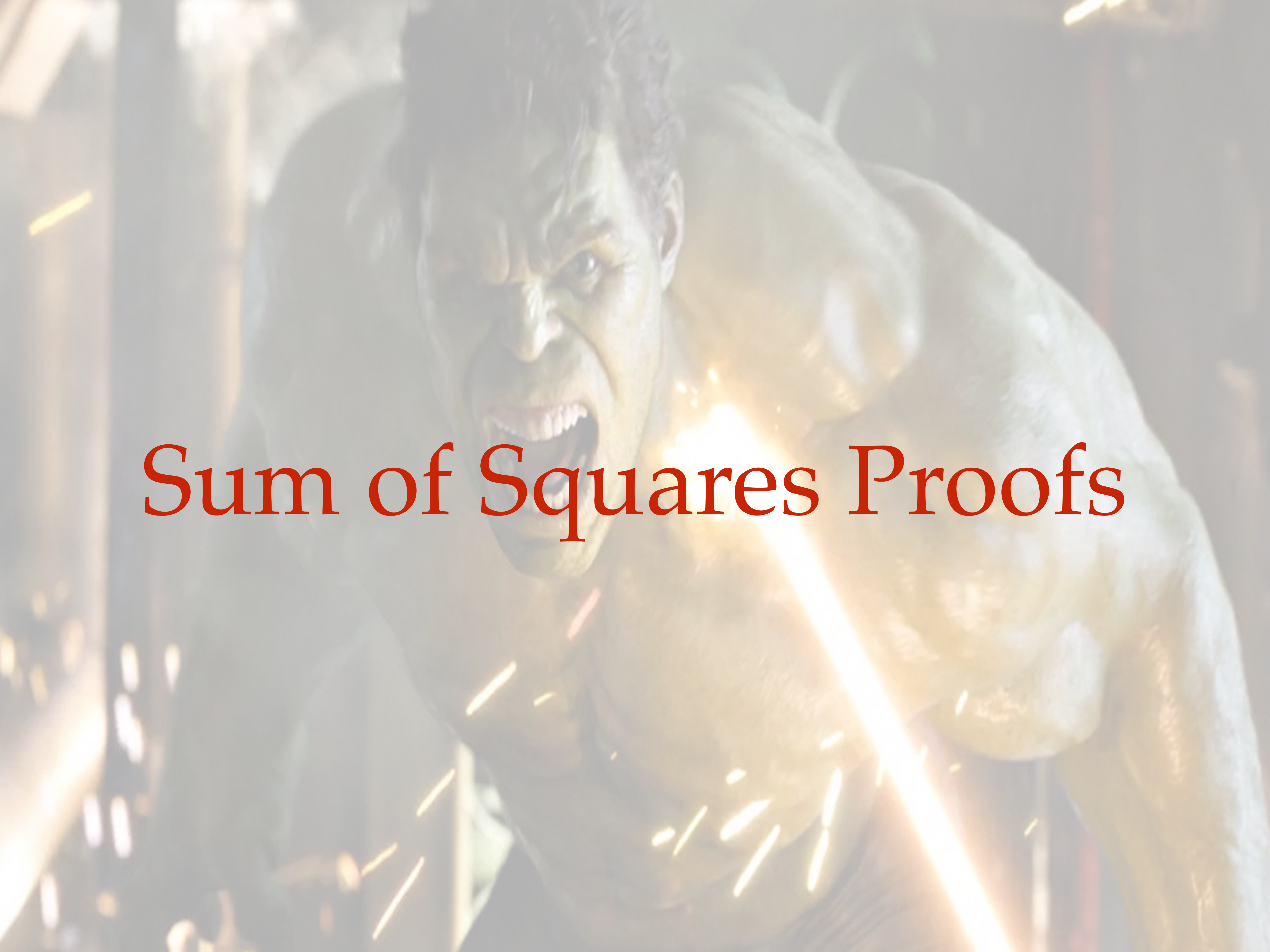
Input $\{y_1, y_2, \ldots, y_n\}$ ϵ-***corrupted*** sample.

Variables/Constraints

$X' = \{x'_1, x'_2, \ldots, x'_n\}$ a guess for original sample. A coupling w.

$w_i^2 = w_i \quad w_i(y_i - x'_i) = 0 \;\; \forall i \qquad \sum_i w_i = (1-\epsilon)n$

Minimize $\|\Sigma(X')\|$



Sum of Squares Proofs

Identifiability to Algorithms

Automatically translate "simple" *identifiability* proofs into algorithms!

What does simple mean?

captured in the sum of squares proof system

- A "proof system" that reasons about polynomial inequalities
- Degree t proofs can be found in time $d^{O(t)}$
- Many natural inequalities have low-degree SoS proofs

Holder's, Cauchy-Schwarz, Triangle Inequality, Brascamp-Lieb inequalities…

growing general toolkit of **ready-to-use** SoS facts*!

***See notes at sumofsquares.org**

SoS in Average Case

"Simple proofs of identifiability = algorithm"

TENSOR DECOMPOSITION, DICTIONARY LEARNING [Barak-Kelner-Steurer'14, Ge-Ma'15, Ma-Shi-Steurer'16]

TENSOR COMPLETION [Barak-Moitra'15, Potechin-Steurer'16]

BEYOND SPECTRAL CLUSTERING [**K-**Steinhardt'17]

ROBUST REGRESSION [Klivans-**K-**Meka'17]

BREAK CRYPTO ASSUMPTIONS [Barak-Brakerski-Komargodski-**K'**17]

Tight lower bounds via Pseudocalibration

RANDOM CSPS [Barak-Chan-**K'15, K-**Mori-O'Donnell-Witmer'17]

PLANTED CLIQUE [Barak-Hopkins-Kelner-**K-**Moitra-Potechin'16]

SPARSE/ TENSOR PCA [Hopkins-**K**-Potechin-Raghavendra-Schramm-Steurer17]

More SoS

Many great open directions!

SoS for Worst-Case Problems?

Quantum information, UGC/Small-Set Expansion,…

SoS for crypto assumptions?

SoS for computing equilibria?

Thank you for your attention!