CS 59300 – Algorithms for Data Science

Classical and Quantum approaches

Lecture 9 (10/02)

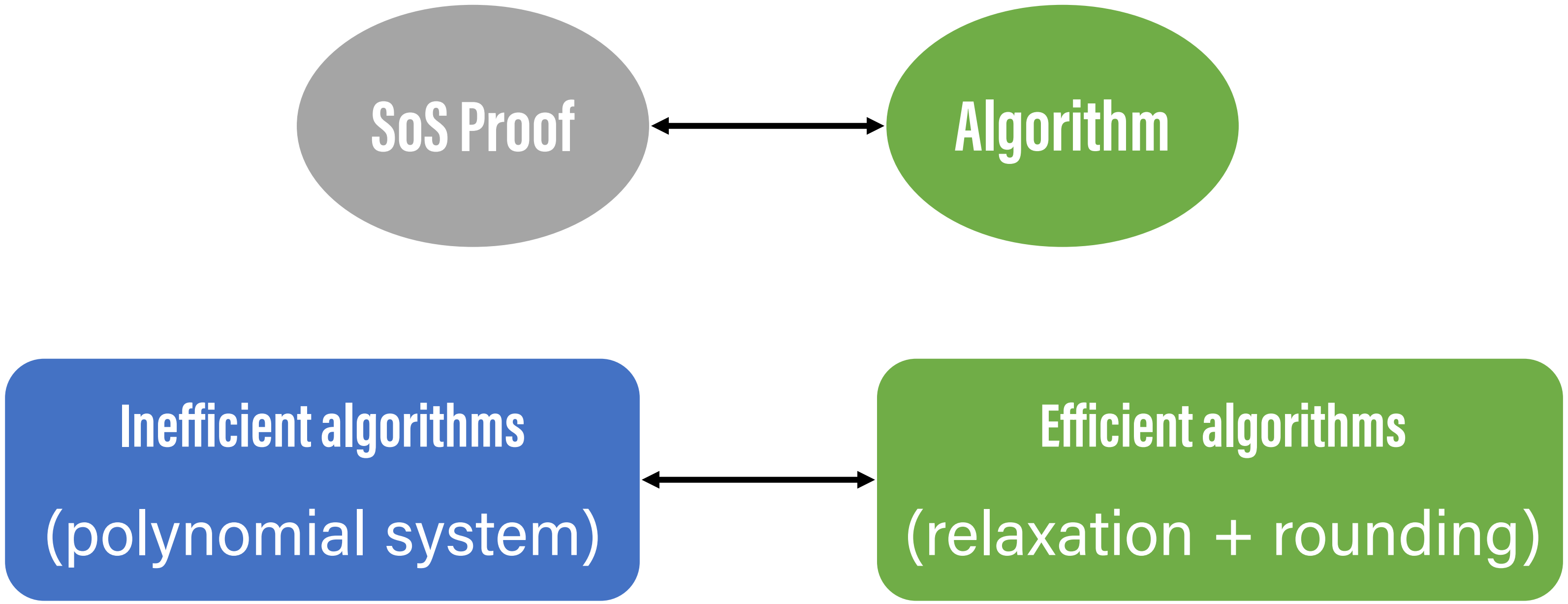
Sum-of-Squares (II)

https://ruizhezhang.com/course_fall_2025.html

The slides is based on Tselil Schramm's note

Sum-of-Squares (SoS)

Powerful generic framework for algorithm design/nonconvex optimization



Yields the most powerful approximation algorithms for many statistical/ML problems

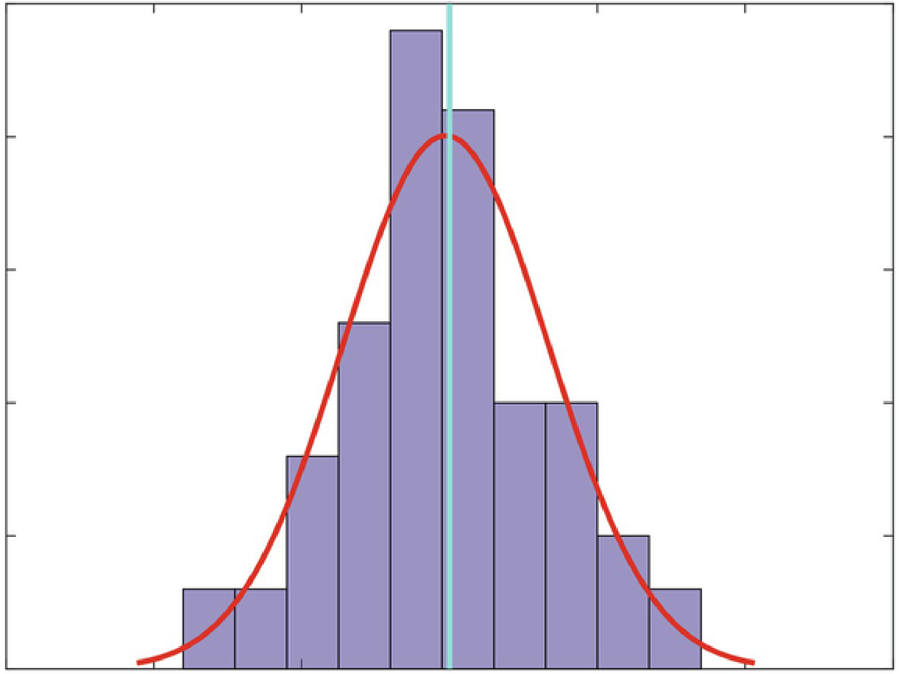
- Max-cut, tensor decomposition, dictionary learning, matrix/tensor completion, sparse PCA, Gaussian mixture models, planted clique, robust statistics, quantum separability, ...

Robust mean estimation

Mean estimation is a well-studied problem:

Given *i.i.d.* samples $v_1, v_2, \ldots, v_n \in \mathbb{R}^d$ from an unknown distribution $\mathcal{D}$, estimate $u \coloneqq \mathbb{E}_{\mathcal{D}}[x]$

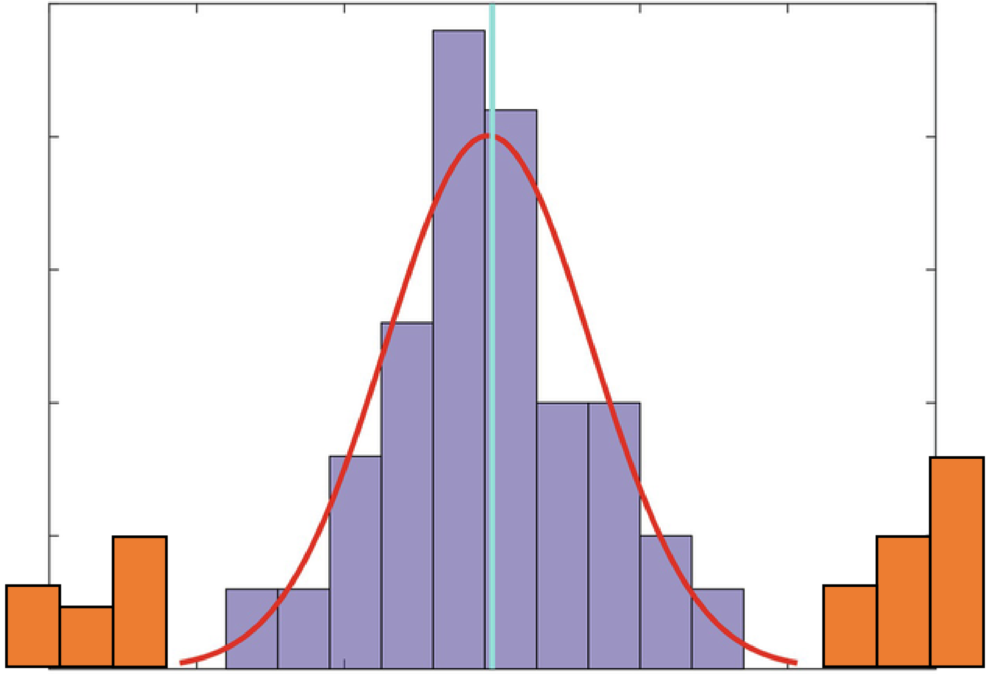
- For many scenarios, empirical mean (i.e. $\frac{1}{n}\sum_i v_i$) is enough



Robust mean estimation

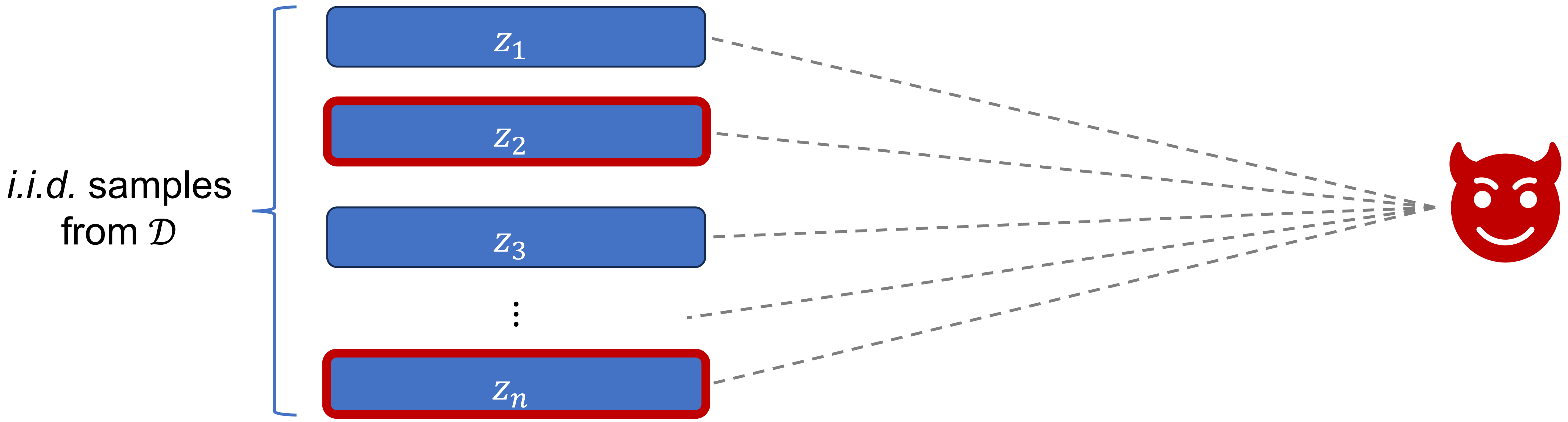
Robust mean estimation:

Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean u, covariance $\Sigma \preccurlyeq I$, and bounded 4th moments. Our goal is to estimate u from ϵ-corrupted samples: $v_1, \ldots, v_n \in \mathbb{R}^d$, a $(1-\epsilon)$-fraction sampled *i.i.d.* from $\mathcal{D}$ and the remaining ϵ-fraction are adversarially chosen.



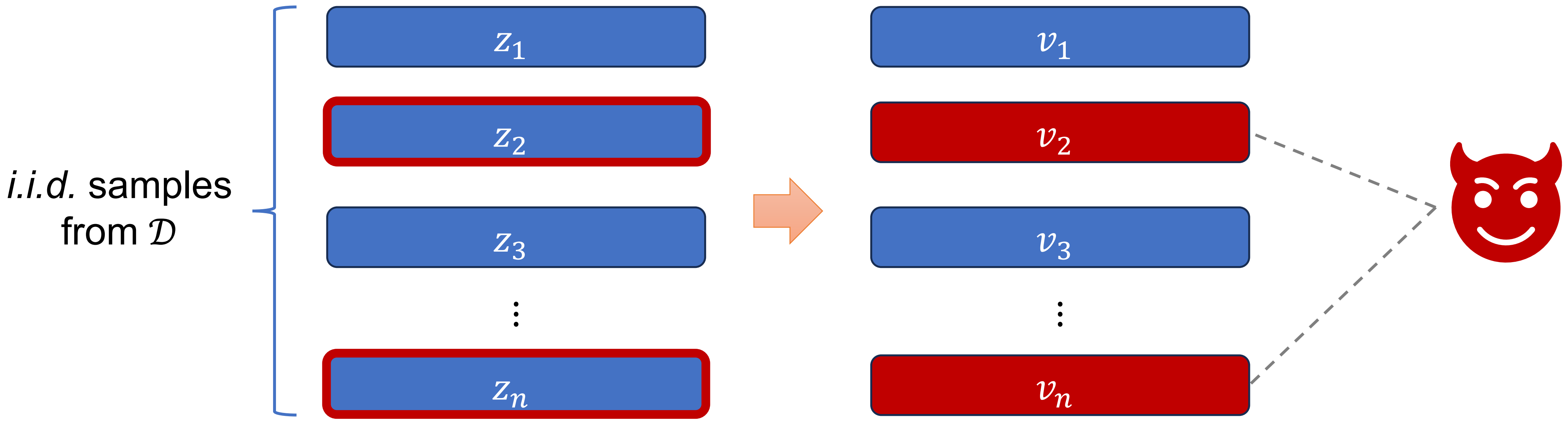
Robust mean estimation

Standard mean estimation setup, but where an ϵ-fraction of data is adversarially corrupted



Robust mean estimation

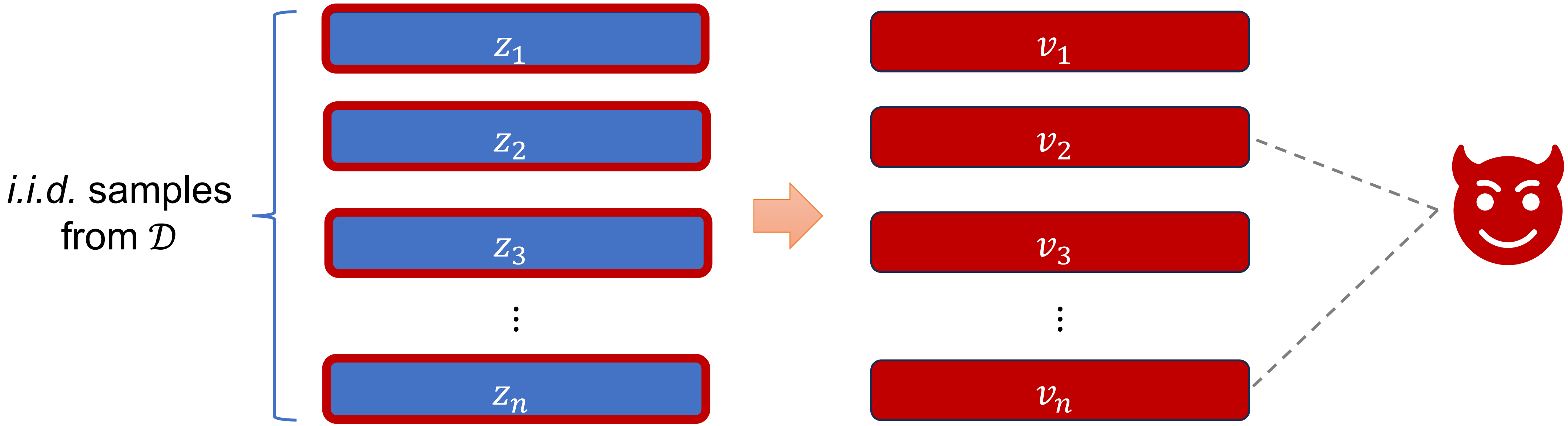
Standard mean estimation setup, but where an ϵ-fraction of data is adversarially corrupted



Robust mean estimation

Standard mean estimation setup, but where an ϵ-fraction of data is adversarially corrupted

- Not always solvable (e.g. $\epsilon = 1$)



Robust mean estimation

Standard mean estimation setup, but where an ϵ-fraction of data is adversarially corrupted

Under what conditions is estimating the mean u possible?

Identifiability

Robust mean estimation: Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean u, covariance $\Sigma \preccurlyeq I$, and bounded 4th moments. Our goal is to estimate u from ϵ-corrupted samples: $v_1, \dots, v_n \in \mathbb{R}^d$, a $(1-\epsilon)$-fraction sampled *i.i.d.* from $\mathcal{D}$ and the remaining ϵ-fraction are adversarially chosen.

Identifiability lemma. If $n = \mathrm{poly}(d)$ sufficiently large, and if $S \subset [n]$ of size $|S| = (1-\epsilon)n$ satisfies, for $u_S \coloneqq \frac{1}{|S|}\sum_{i \in S} v_i$,

$$\frac{1}{|S|}\sum_{i \in S}(v_i - u_S)(v_i - u_S)^\top \preccurlyeq 2I,$$

then $\|u_S - u\| \leq \mathcal{O}(\sqrt{\epsilon})$ with high probability over the samples

- This lemma gives an algorithm for robust mean estimation

Algorithm from identifiability lemma

Identifiability lemma. If $n = \mathrm{poly}(d)$ sufficiently large, and if $S \subset [n]$ of size $|S| = (1-\epsilon)n$ satisfies, for $u_S \coloneqq \frac{1}{|S|}\sum_{i \in S} v_i$,

$$\frac{1}{|S|}\sum_{i \in S}(v_i - u_S)(v_i - u_S)^\top \preccurlyeq 2I, \qquad (\bigstar)$$

then $\|u_S - u\| \le \mathcal{O}(\sqrt{\epsilon})$ with high probability over the samples

- The $(1-\epsilon)$-fraction of uncorrupted data form such set S
- For every possible S, check whether $(\bigstar)$ holds

Issue: running time is $\ge \binom{n}{(1-\epsilon)n} = 2^{\mathcal{O}(n)}$, inefficient!

Key idea:

- The proof is a low-degree SoS proof
- It will automatically give an efficient algorithm

Sum-of-squares proofs

Definition. For multivariate polynomials $p, q \in \mathbb{R}[x]$ in variables $x \in \mathbb{R}^N$, we say that $p \geq q$ is a degree-k sum-of-squares inequality if there exists polynomials $s_1, s_2, \ldots \in \mathbb{R}[x]$ of degree $\leq k/2$ such that

$$p - q = \sum_i s_i^2$$

We say that there is a degree-k SoS proof of $p \geq q$ modulo the axioms $\mathcal{A} = \{f_j = 0\}_{j\in[M]} \cup \{g_\ell \geq 0\}_{\ell\in[N]}$ if there exists polynomials $a_1, \ldots, a_M \in \mathbb{R}[x]$ and SoS polynomials $S, S_1, \ldots, S_N$ such that $\deg(a_j f_j) \leq k$, $\deg(S_\ell g_\ell) \leq k$, $\deg(S) \leq k$, and

$$p - q = S + \sum_{j=1}^{M} a_j f_j + \sum_{\ell=1}^{N} S_\ell g_\ell$$

We denote such an SoS proof as $\mathcal{A} \vdash_k p \geq q$

Expressive power of SoS proofs

Fact. If $p, q \in \mathbb{R}[x]$ are univariate and $p \geq q$, then there is always an SoS proof of degree at most $\max\{\deg(p), \deg(q)\}$

Proof sketch.

- Wlog, just consider $p \geq 0$, and induction on $l \coloneqq \deg(p)$
- For $l > 0$, let $p_{\min}$ be the global minimum value of p achieved at $x = a$
- Then, $p(x) - p_{\min} = (x - a)^t r(x)$ with some even number t
- Apply induction hypothesis to $r(x)$

Expressive power of SoS proofs

Fact. If $p, q \in \mathbb{R}[x]$ are univariate and $p \geq q$, then there is always an SoS proof of degree at most $\max\{\deg(p), \deg(q)\}$

This fact is not generally true for multivariate polynomials

- **Hilbert '1888:** non-constructible proof
- **Motzkin '1967:** explicit counterexample

$$p(x, y) = x^4y^2 + x^2y^4 - 3x^2y^2 + 1$$

Expressive power of SoS proofs

Fact. If $p, q \in \mathbb{R}[x]$ are univariate and $p \geq q$, then there is always an SoS proof of degree at most $\max\{\deg(p), \deg(q)\}$

- David Hilbert's 17th problem, resolved by Artin, Krivine, Stengle

Theorem (Positivestellensatz).

Any non-negative (modulo the axioms $\mathcal{A}$) polynomial can be written as a sum of squares of rational functions

SoS toolkit (2nd episode)

SoS Cauchy-Schwarz (plus): Let a, b be vector-valued polynomials of degree d_a and d_b. Then for any $\epsilon > 0$,

$$\vdash_{d_a+d_b} \langle a, b\rangle \leq \frac{\epsilon}{2}\|a\|^2 + \frac{1}{2\epsilon}\|b\|^2$$

and

$$\vdash_{2(d_a+d_b)} \langle a, b\rangle^2 \leq \|a\|^2\|b\|^2$$

SoS toolkit (2nd episode)

SoS operator norm: Let $y \in \mathbb{R}[x]^n$, $M \in \mathbb{R}[x]^{n\times n}$, and $B \in \mathbb{R}[x]^{n\times k}$. Then for any $\epsilon > 0$,

$$\{M = \lambda I - BB^\top\} \vdash_d y^\top M y \le \lambda \|y\|^2$$

for $d = \deg(y^\top M y + y^\top B B^\top y)$

Proof.

- The axioms imply that $y^\top M y = y^\top(\lambda I - BB^\top)y = \lambda\|y\|^2 - \|B^\top y\|^2$
- That is,

$$\lambda\|y\|^2 - y^\top M y = \underbrace{y^\top(\lambda I - BB^\top - M)y + \|B^\top y\|^2}_{\text{SoS}}$$

■

Recap: Pseudoexpectation

Definition. For a set of polynomial axioms $\mathcal{A} = \{f_i = 0\} \cup \{g_j \geq 0\}$, $\widetilde{\mathbb{E}}: \mathbb{R}[x] \to \mathbb{R}$ is a degree-k pseudoexpectation satisfying $\mathcal{A}$ if it is a linear operator with the following properties:

1. $\widetilde{\mathbb{E}}[1] = 1$
2. $\widetilde{\mathbb{E}}[h^2] \geq 0 \ \forall\, h \in \mathbb{R}[x], \deg(h) \leq k/2$
3. $\widetilde{\mathbb{E}}[af_i] = 0 \ \forall\, a \in \mathbb{R}[x], \deg(af_i) \leq k$, and $\widetilde{\mathbb{E}}\left[b^2 g_j\right] \geq 0 \ \forall\, b \in \mathbb{R}[x], \deg\left(b^2 g_j\right) \leq k$

SoS proof and pseudoexpectation duality

- If $\mathcal{A} \vdash p \geq q$, then $\widetilde{\mathbb{E}}[p] \geq \widetilde{\mathbb{E}}[q]$ for any $\widetilde{\mathbb{E}}$ satisfying $\mathcal{A}$
- If there exist an $\widetilde{\mathbb{E}}$ such that $\widetilde{\mathbb{E}}[p] \leq \widetilde{\mathbb{E}}[q]$, then $\mathcal{A} \nvdash p \geq q$

SoS proof of identifiability for robust mean estimation

Robust mean estimation: Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean u, covariance $\Sigma \preccurlyeq I$, and bounded 4th moments. Let $z_1, \dots, z_n$ be *i.i.d.* samples from $\mathcal{D}$. Our goal is to estimate u given ϵ-corrupted samples $v_1, \dots, v_n$ with the guarantee that for $(1-\epsilon)n$ indices $i \in [n]$, $v_i = z_i$.

Polynomial formulation:

- Variables: $Z_1, \dots, Z_n \in \mathbb{R}^n$, $W_1, \dots, W_n \in \mathbb{R}$, $B \in \mathbb{R}^{d \times d}$
- Polynomial system:

$$\mathcal{A} = \left\{ \begin{array}{ll} ① & W_i^2 = W_i \quad \forall\, i \in [n] \\ ② & \sum_{i=1}^{n} W_i = (1-\epsilon)n \\ ③ & W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n] \\ ④ & \overline{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top \end{array} \right.$$

$W_i \in \{0,1\}$ is an indicator of clean sample

(★) in Identifiability lemma

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

SoS algorithm:

1. Solve a degree-6 pseudoexpectation $\widetilde{\mathbb{E}}$ satisfying $\mathcal{A}$
2. Output $\widetilde{\mathbb{E}}[\overline{Z}]$

Proof-to-algorithm

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

- From the duality, the lemma gives that $\widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^4\right] \leq \mathcal{O}(\epsilon)\widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right]$
- We have

$$0 \leq \widetilde{\mathbb{E}}\left[\left(\left\|\overline{Z} - \overline{z}\right\|^2 - \widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right]\right)^2\right] = \widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^4\right] - \widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right]^2$$

$$\leq \underbrace{\widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right]}_{\geq 0}\left(\mathcal{O}(\epsilon) - \widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right]\right)$$

Proof-to-algorithm

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

- We have

$$\begin{aligned}
\mathcal{O}(\epsilon) &\geq \widetilde{\mathbb{E}}\left[\left\|\overline{Z} - \overline{z}\right\|^2\right] = \widetilde{\mathbb{E}}\left[\left\|\overline{Z}\right\|^2\right] - 2\langle \widetilde{\mathbb{E}}[\overline{Z}], \overline{z}\rangle + \|\overline{z}\|^2 \\
&= \widetilde{\mathbb{E}}[\|\overline{Z}\|]^2 - 2\langle \widetilde{\mathbb{E}}[\overline{Z}], \overline{z}\rangle + \|\overline{z}\|^2 + \left(\widetilde{\mathbb{E}}\left[\left\|\overline{Z}\right\|^2\right] - \widetilde{\mathbb{E}}[\|\overline{Z}\|]^2\right) \\
&= \left\|\overline{z} - \widetilde{\mathbb{E}}[\overline{Z}]\right\|^{\mathbf{2}} + \left(\widetilde{\mathbb{E}}\left[\left\|\overline{Z}\right\|^2\right] - \widetilde{\mathbb{E}}[\|\overline{Z}\|]^2\right) \\
&\geq \left\|\overline{z} - \widetilde{\mathbb{E}}[\overline{Z}]\right\|^{\mathbf{2}}
\end{aligned}$$

≥ 0 by SoS Cauchy-Schwarz

Proof-to-algorithm

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

- $\left\|\overline{z} - \widetilde{\mathbb{E}}\left[\overline{Z}\right]\right\| \leq \mathcal{O}(\sqrt{\epsilon})$
- For sufficiently large n, $\|\overline{z} - u\| \leq \sqrt{\epsilon}$
- By triangle inequality, $\left\|u - \widetilde{\mathbb{E}}\left[\overline{Z}\right]\right\| \leq \mathcal{O}(\sqrt{\epsilon})$
- Thus, the SoS algorithm works

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Z_i - \overline{Z}\right)^{\top} = 2I - BB^{\top}$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- We can expand $\left\|\overline{Z} - \overline{z}\right\|^4$ as follows:

$$\left\langle \overline{z} - \overline{Z}, \overline{z} - \overline{Z}\right\rangle^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - W_i \mathbf{1}_{z_i = v_i}\right)\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle + \frac{1}{n}\sum_{i=1}^{n} W_i \mathbf{1}_{z_i = v_i} \cancel{\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle}\right)^2$$

$Z_i = v_i = z_i$

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- We can expand $\left\|\overline{Z} - \overline{z}\right\|^4$ as follows:

$$\left\langle \overline{z} - \overline{Z}, \overline{z} - \overline{Z}\right\rangle^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(1 - W_i \mathbf{1}_{z_i = v_i}\right)}_{\deg = 1}\underbrace{\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle}_{\deg = 2}\right)^2$$

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^{\top} = 2I - BB^{\top}$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- By SoS Cauchy-Schwarz,

$$\mathcal{A} \vdash_6 \left\langle \overline{z} - \overline{Z}, \overline{z} - \overline{Z} \right\rangle^2 \le \boxed{\left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - W_i \mathbf{1}_{z_i = v_i}\right)^2\right)} \left(\frac{1}{n}\sum_{i=1}^{n}\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle^2\right)$$

SoS proof of identifiability for robust mean estimation

(1) $W_i^2 = W_i \quad \forall\, i \in [n]$ (3) $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$

(2) $\sum_{i=1}^{n} W_i = (1-\epsilon)n$ (4) $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,,\quad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- (1) $\vdash_2 \left(1 - W_i \mathbf{1}_{z_i = v_i}\right)^2 = 1 - 2W_i \mathbf{1}_{z_i=v_i} + W_i^2 \mathbf{1}_{z_i=v_i} = 1 - W_i \mathbf{1}_{z_i=v_i}$
- (2) $\vdash_1 \frac{1}{n}\sum_i\left(1 - W_i \mathbf{1}_{z_i=v_i}\right) = 1 - \frac{1}{n}\sum_i W_i + \frac{1}{n}\sum_i W_i \mathbf{1}_{z_i \ne v_i} = \epsilon + \frac{1}{n}\sum_i W_i \mathbf{1}_{z_i \ne v_i}$
- (1) $\vdash_2 \frac{1}{n}\sum_i W_i \mathbf{1}_{z_i \ne v_i} \le \frac{1}{n}\sum_i \mathbf{1}_{z_i \ne v_i} = \epsilon$

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^{\top} = 2I - BB^{\top}$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- $\mathcal{A} \vdash_2 \frac{1}{n}\sum_{i=1}^{n}\left(1 - W_i \mathbf{1}_{z_i = v_i}\right)^2 \leq 2\epsilon$
- Thus,

$$\mathcal{A} \vdash_6 \left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - W_i \mathbf{1}_{z_i = v_i}\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle^2\right) \leq 2\epsilon\left(\frac{1}{n}\sum_{i=1}^{n}\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle^2\right)$$

SoS proof of identifiability for robust mean estimation

① $W_i^2 = W_i \quad \forall\, i \in [n]$

② $\sum_{i=1}^{n} W_i = (1-\epsilon)n$

③ $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$

④ $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- Let $b \coloneqq \overline{z} - \overline{Z}$. We have,

$$\langle z_i - Z_i, b\rangle = \langle z_i - Z_i - b + b, b\rangle = \langle z_i - \overline{z}, b\rangle - \left\langle Z_i - \overline{Z}, b\right\rangle + \|b\|^2$$

- By SoS triangle inequality,

$$\vdash_4 \left(\langle z_i - \overline{z}, b\rangle - \left\langle Z_i - \overline{Z}, b\right\rangle + \|b\|^2\right)^2 \le \frac{10}{3}\left(\langle z_i - \overline{z}, b\rangle^2 + \left\langle Z_i - \overline{Z}, b\right\rangle^2 + \|b\|^4\right)$$

SoS proof of identifiability for robust mean estimation

1 $W_i^2 = W_i \quad \forall\, i \in [n]$

2 $\sum_{i=1}^{n} W_i = (1-\epsilon)n$

3 $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$

4 $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Z_i - \overline{Z}\right)^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

$$\vdash_4 \frac{1}{n}\sum_{i=1}^{n}\left\langle z_i - Z_i, \overline{z} - \overline{Z}\right\rangle^2 \le \frac{1}{n}\frac{10}{3}\sum_{i=1}^{n}\left(\langle z_i - \overline{z}, b\rangle^2 + \left\langle Z_i - \overline{Z}, b\right\rangle^2 + \|b\|^4\right)$$
$$= \frac{10}{3}\left(b^\top \Sigma_z b + b^\top \Sigma_Z b + \|b\|^4\right)$$

where $\Sigma_z \coloneqq \mathbf{Cov}(z_1, \dots, z_n)$ and $\Sigma_Z \coloneqq \mathbf{Cov}(Z_1, \dots, Z_n)$

SoS proof of identifiability for robust mean estimation

1 $W_i^2 = W_i \quad \forall\, i \in [n]$	3 $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
2 $\sum_{i=1}^{n} W_i = (1-\epsilon)n$	4 $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \quad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \le \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- 4 $\vdash_4 b^\top \Sigma_Z b \le 2\|b\|^2$ (SoS operator norm)
- For sufficiently large n, $\Sigma_z \preccurlyeq 2I$, which implies $b^\top \Sigma_z b \le 2\|b\|^2$
- Thus, $\mathcal{A} \vdash_4 \frac{1}{n}\sum_{i=1}^{n}\langle z_i - Z_i, \overline{z} - \overline{Z}\rangle^2 \le \frac{10}{3}(4\|b\|^2 + \|b\|^4)$

SoS proof of identifiability for robust mean estimation

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \qquad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Lemma. Let $\overline{z} \coloneqq \frac{1}{n}\sum_i z_i$ be the empirical mean of the uncorrupted samples. So long as $n = \mathrm{poly}(d)$, with high probability,

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left\|\overline{Z} - \overline{z}\right\|^2$$

Proof.

- Putting everything together, we have

$$\mathcal{A} \vdash_6 \left\|\overline{Z} - \overline{z}\right\|^4 \leq \mathcal{O}(\epsilon)\left(4\left\|\overline{Z} - \overline{z}\right\|^2 + \left\|\overline{Z} - \overline{z}\right\|^4\right)$$

■

SoS for robust statistical problem

Identifiability lemma. If $n = \mathrm{poly}(d)$ sufficiently large, and if $S \subset [n]$ of size $|S| = (1-\epsilon)n$ satisfies, for $u_S \coloneqq \frac{1}{|S|}\sum_{i\in S} v_i$,

$$\frac{1}{|S|}\sum_{i\in S}(v_i - u_S)(v_i - u_S)^\top \preccurlyeq 2I,$$

A statement about statistics

then $\|u_S - u\| \le \mathcal{O}(\sqrt{\epsilon})$ with high probability over the samples

1. $W_i^2 = W_i \quad \forall\, i \in [n]$
2. $\sum_{i=1}^{n} W_i = (1-\epsilon)n$
3. $W_i(Z_i - v_i) = 0 \quad \forall\, i \in [n]$
4. $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i\,, \quad \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})^\top = 2I - BB^\top$

Design a polynomial system

Efficient algorithm via pseudoexpectation (+rounding)

Convert to SoS proof

Sum-of-Squares (SoS)

Two pipelines for SoS-based algorithm design: