# Course notes for 6.S977: The Sum of Squares Method

Fall 2022

These are lecture notes I (Noah Golowich) have taken for the course "6.S977: The Sum of Squares Method", taught by Sam Hopkins at MIT in the Fall of 2022. Please note that they are very rough and have not been subjected to any sort of scrutiny or editing (and thus likely to contain errors). Any errors are my own.

## Contents

# 1 September 9, 2022

SoS has connections with multiple areas including: optimization, robotics, optimal control, combinatorics, algebraic geometry, statistics, ML, etc.

## 1.1 Optimization on $\{0,1\}^n$

Given a map $f : \{0,1\}^n \to \mathbb{R}$, find $\min_{x\in\{0,1\}^n} f(x)$. We will assume that we have a succinct representation of $f$ by a polynomial; in particular, $f(x) = \sum_{S\subset[n]} \hat{f}(S)\cdot\prod_{i\in S} x_i = \sum_{S\subset[n]} \hat{f}(S)\cdot x^S$. Here we assume that the number of $S$ with $\hat{f}(S) \neq 0$ is bounded, so that $f$ may be computed efficiently.

It is useful to look at proofs here. What is a proof that $\min f(x) \leq \alpha$: simply some $x$ so that $f(x) \leq \alpha$.

What is a proof that $\min f(x) \geq \alpha$? A trivial (long) witness is simply the truth table of $f$. What is a short witness?

**Definition 1.1.** We say that "$f(x) \geq 0 \forall x \in \{0,1\}^n$" has a degree $d$ SoS proof, namely $\vdash_d f(x) \geq 0$ if there exist $p_1, \ldots, p_M \in \mathbb{R}[x]_{\leq d/2}$ so that

$$f(x) = \sum_{i=1}^{m} p_i(x)^2, \ \forall x \in \{0,1\}^n. \tag{1}$$

Here's a trivial proof:

$$x + y - 2xy = x^2 + y^2 - 2xy = (x-y)^2,$$

so the above is a proof in degree 2 that $\vdash_2 x + y - 2xy \geq 0$.

Here are some basic questions:

1. Do SoS proofs exist for all $f \geq 0$?
2. How large are they?
3. Can we verify them efficiently?
4. Can we find such proofs efficently?

Facts:

1. WLOG we can take $m \leq n^d$ (we will prove this today). Furthermore, if $\vdash_d f(x) \geq 0$, then $\vdash_d f + \epsilon \geq 0$, where the coefficients of the polynomials can be represented in $\text{poly}(n^d, \log 1/\epsilon)$ bits. In particular, the number of $p_i$'s and the number of bits in their representation can both be bounded appropriately, so we are good. We won't worry about the imprecision $\epsilon$. This therefore answers the question of how large the proofs are.
2. Given $p_1, \ldots, p_m$, we can check in $\text{poly}(n^d)$ time if $f(x) = \sum_i p_i(x)^2$ for all $x \in \{0,1\}^n$. This answers the question of whether we can verify the proofs.
3. If $\vdash_d f \geq 0$, we can find $p_1, \ldots, p_m$ certifying $\vdash_d f + \epsilon \geq 0$ in $\text{poly}(n^d, \log 1/\epsilon)$ time. This answers the final question of whether we can find such proofs efficiently.

4. If $f(x) \geq 0$ for all $x \in \{0,1\}^n$, then $\vdash_{2n} f(x) \geq 0$, i.e., there's a proof in degree $2n$. This isn't that surprising, since there's a trivial exponential sized witness of $f \geq 0$; since the degree is linear in $n$, the size of the proof is exponential in $n$.

5. What can we get with only a small amount of computation? For all $f$, there exists $\alpha \in \mathbb{R}$ so that $\vdash_{\deg(f)} f + \alpha \geq 0$, where $\deg(f)$ denotes the smallest even integer larger than the degree of $f$. Note that here $-\alpha$ may not be the minimum degree of $f$. This is on the opposite side of the spectrum of the previous one (very loose bound, but minimum possible degree). Much of the course is focused on getting something in-between: getting much better approximations of $\min f$, but with reasonably small degree.

How much of the above depends on the hypercube? We use the fact that the hypercube is a variaty, namely $\{x \in \mathbb{R}^n \ : \ x_i^2 = x_i \forall i\}$, and is in fact a very nice variety (e.g., has Grobner basis). If the set is all of $\mathbb{R}^n$, all of the facts work, more or less, with some modifications. For TCS purposes, there are some other natural domains where everything carries over nicely: e.g., $\{x \in \mathbb{R}^n : \|x\|^2 \leq 1\}$, or $\{x \in \mathbb{R}^n : x_i \in \{-1,0,1\}, \ \sum_i x_i^2 = k\}$.

## 1.2 Multilinearization

**Fact 1.1.** *Every $f : \{0,1\}^n \to \mathbb{R}$ has a unique representation as $f(x) = \sum_{S \subset [n]} \hat{f}(S) \cdot x^S$.*

This is multilinear in the sense that $x^S$ has no repeated $x_i$ in the product. This fact enables Boolean Fourier analysis.

*Proof.* We can write $f(x) = \sum_y f(y) \cdot \mathbb{1}\{x = y\}$. The indicator function $\mathbb{1}\{x = y\}$ is trivialy a degree $n$ polynomial, and then we multilinearize (in particular, reduce modulo the ideal $(x_i^2 = x_i)_{i \in [n]}$). $\square$

As a consequence of the above fact: to check whether 2 polynomials are equal over the hypercube, it suffices to check whether their unique multilinear representations are equal, and so if they are degree $d$, this takes $O(n^d)$ time.

We prove the 4th fact above:

**Fact 1.2.** *If $f \geq 0$, then $\vdash_{2n} f \geq 0$.*

*Proof.* Let $g(x) = \sqrt{f(x)}$, which is well-defined since $f \geq 0$. By the previous fact $g$ has a representation as am ultilinear polynomial. Now $f(x) = g(x)^2$, and $\deg(g) \leq n$. So, this certifies $\vdash_{2n} f \geq 0$. $\square$

In general, taking square roots is not a great idea since the square root of a function has high degree.

Now we show that the proofs have a certain nice representation:

**Lemma 1.3** (Matrix representation lemma)**.** *$\vdash_{2d} f \geq 0$ iff exists a matrix $M \in \mathbb{R}^{\binom{n}{\leq d} \times \binom{n}{\leq d}}$ so that for all $s \subset [n]$, $|S| \leq 2d$ so that*

$$\hat{f}(S) = \sum_{A,B : A \cup B = S} M_{AB},$$

*and $M \succeq 0$.*

*Proof.* First we show that existence of the matrix implies a $2d$-degree proof. Since $M$ is PSD, we can write $M = BB^\top$ for some square matrix $B$ of the same dimensions as $M$. Let us denote the columns of $B$ as $B_1, \ldots, B_N$. Now define

$$p_i(x) := \langle B_i, X_d(x) \rangle = \sum_{|A| \le d} B_i(A) \cdot x^A,$$

where the vector $X_d(x) = (1, x_1, \ldots, x_n, x_1x_2, \ldots)$ is the vector of monomials of degree at most $d$. The assumption in the lemma implies that

$$f(x) = \sum_{A,B} x^A x^B M_{AB} = \sum_{A,B} X_d(x)_A X_d(x)_B \cdot M_{AB} = \sum \langle B_i, X_d(x) \rangle^2 = \sum_i p_i(x)^2.$$

The other direction in the proof of this lemma reverses the above direction: a SoS proof gives the polynomials $p_i$, then we can define a (non-square) matrix $B$, and set $M = BB^\top$, which gives us the desired representation of $\hat{f}(S)$. □

Next we prove Fact 1 from above: If there is a degree $d$ SoS proof, we can construct the matrix $M$ as in the above lemma, and then applying the reverse direction of the lemma we get that there is a SoS proof with a number of polynomials at most the rank of $M$, which is at most $\binom{n}{n^d}$.

To prove the remaining 2 facts, we need the following convexity fact:

**Lemma 1.4.** *The set* $\{f \ : \vdash_d f \ge 0\}$ *is convex.*

*Proof.* Suppose that polynomials $p_1, \ldots, p_m$ certify $f \ge 0$ and polynomials $q_1, \ldots, q_m$ certify $g \ge 0$. For any $\alpha, \beta \ge 0$, we need to certify that $\vdash_d \alpha f + \beta g \ge 0$. We now check that:

$$\alpha f + \beta g = \sum_i (\sqrt{\alpha} \cdot p_i(x))^2 + \sum_j (\sqrt{\beta} \cdot q_j(x))^2.$$

□

We now prove Fact 5:

**Lemma 1.5.** *For all* $f$ *there is* $\alpha$ *so that* $\vdash_{2\deg(f)} f + \alpha \ge 0$.

(It is actually possible to get $\deg(f)$, but we do $2\deg(f)$ since it's easier.)

*Proof.* By convexity, it is enough to check the fact for monomials $x^S, -x^S$. This is because any polynomial $f$ can be written as a nonnegative linear combination of monomialas $x^S, -x^S$; then we take the corresponding nonnegative linear combination of the corresponding SoS proofs.

For $x^S$, we have that $x^S = (x^S)^2$; this is immediate since $x^S$ takes values in $\{0, 1\}$ for all $x \in \{0, 1\}^n$.

What about for $-x^S$? Let's look at $1 - x^S$ (here we choose $\alpha = 1$). We can write:

$$-x^S + 1 = (x^S)^2 - 2x^S + 1 = (x^S - 1)^2,$$

and so $\vdash_{2|S|} -x^S + 1 \ge 0$.

By inspecting the proof it is evident that we can take $\alpha = \sum_S |\hat{f}(S)|$. □

The above lemma shows that we can certify $\min_x f(X) \geq -\sum_S |\hat{f}(S)|$. This bound is not an interesting bound at all on the minimum value of $f$. In the remainder of today's lecture we will certify a nontrivial bound on an interesting function.

First, we cerify the proof of fact 3:

**Lemma 1.6.** *If $f$ has a degree $d$ SoS proof, we can find $p_1, \ldots, p_m$ for $\vdash_d f+\epsilon \geq 0$ in $\mathrm{poly}(n^d, \log 1/\epsilon)$ time.*

*Proof.* The idea is to search for $M \in \mathbb{R}^{\binom{n}{\leq d} \times \binom{n}{\leq d}}$ so that $\hat{f}(S) = \sum_{A \cup B = S} M_{AB}$, $M \succeq 0$. There are various technical questions about the precision we can get for finding solutions of convex programs, which we gloos over – but by using semidefinite programming (e.g., ellipsoid algorithm), we can find the proof in $\mathrm{poly}(n^d, \log 1/\epsilon)$ time. □

## 1.3 Max-Cut

**Definition 1.2** (Max-Cut)**.** Given a graph $G = (V, E)$, the goal is to find $\max_{S \subset V} \delta(S)$, where $\delta(S)$ denotes the number of edges crossing from $S$ into $V - S$.

Max-Cut is NP-hard (Karp). Best we can hope for is approximations. The approximation algorithm we hope for is that which takes a graph $G$ and outputs a number $ALG(G)$ so that

$$\frac{1}{\beta} \cdot ALG(G) \leq MaxCut \leq ALG(G),$$

for some $\beta > 1$. A trivial approximation algorithm for MaxCut is simply the number of edges in $G$, namely $|E(G)|$; it attains a 2-approximation. The way to cut half of the edges is simply to take a random cut.

It was thought for a while that this was perhaps the best you could do; it took 20 years to get a better algorithm, until Goemans-Williamson showed that you could get a $\beta \approx 1/0.878 \approx 1.139$ approximation algorithm. Qualitatively, it has to do something much better than the trivial algorithm.

**Definition 1.3.** We define the *cut polynomial* of $G$ as:

$$G(x) = \sum_{i \sim j} (x_i - x_j)^2.$$

Clearly, for $x \in \{0,1\}^n$, $G(x)$ measures the number of cut edges; in particular, the problem $\min_x -G(x)$ captures the MaxCut problem.

Here's a proposal to solve the MaxCut problem: fix some $d \in \mathbb{N}$. Find the least $\alpha_d$ so that $\vdash_d \alpha_d - G(x) \geq 0$. (We can check if there is a SoS proof by using semidefinite programming, by Fact 3 above.) The main question is whether this algorithm is any good?

**Theorem 1.7.** *We have $\alpha_2 \leq 1.139\ldots \cdot MaxCut(G)$.*

This gives a very nontrivial bound using a low-degree proof. Quite surprising!

Let's start with a sanity check: namely, that $\alpha_4 \leq |E(G)|$. We need to write $|E(G)| - \sum_{i \sim j}(x_i - x_j)^2$ as a sum of squares. It is enough to write $1 - (x_i - x_j)^2$ as a sum of squares, or equivalently $1 - (x - y)^2$. We use the same trick as we did for $-x^S$ above:

$$(1 - (x-y)^2)^2 = 1 + (x-y)^4 - 2(x-y)^2 = 1 + (x-y)^2 - 2(x-y)^2 = 1 - (x-y)^2.$$

Before proceeding, we discuss some obstructions to SoS proofs:

1. There is some $x$ with $f(x) < 0$; if this were the case, we would have that we can certify every nonnegative function. (Unlikely in low-degree since max-cut is NP-hard.)

2. "Pseudoexpectation".

**Definition 1.4.** A degree-$d$ *pseudoexpectation* is a linear operator: $\widetilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$, so that:

1. $\widetilde{\mathbb{E}}1 = 1$.

2. For every $p \in \mathbb{R}[x]_{\leq d/2}$, $\widetilde{\mathbb{E}}p(x)^2 \geq 0$.

3. For every multiset $S \subset [n]$, $\widetilde{\mathbb{E}}(x^S \cdot (x_i)^2) = \widetilde{\mathbb{E}}(x^S \cdot x_i)$.

The idea here is that for a distribution $\mu$ on $\{0,1\}^n$, $\mathbb{E}_\mu : \mathbb{R}[x] \to \mathbb{R}$ is a degree-$d$ pseudoexpectation for all $d$. One high-level idea that we will think of pseudoexpectations as actuall expectations with respect to actaul distributions (even though they're not).

One basic fact is: a degree $2n$ pseudoexpectation is an actual expectation: namely, there exists a distribution $\mu$ so that $\widetilde{\mathbb{E}}p = \mathbb{E}_\mu p$ for all $p$. (This is closely related to the fact that we can certify nonnegativity of any nonnegative polynomial using a degree $2n$ SoS proof.)

**Lemma 1.8.** *Suppose have some polynomial $f$ and a degree-$d$ PE $\widetilde{\mathbb{E}}$ so that $\widetilde{\mathbb{E}}f < 0$. Then it is not the case that $\vdash_d f \geq q0$.*

*Proof.* If not, then we have $\sum_i p_i(x)^2 = f(x)$. By property 3 of pseudoexpectation (in particular, that we can multilinearize up to degree $d$), $\widetilde{\mathbb{E}} \sum_i p_i(x)^2 = \widetilde{\mathbb{E}}f(x)$. But the LHS is non-negative (property 2 of PE) and the RHS is negative (assumption), a contradiction. □

**Lemma 1.9** (Duality)**.** *For all $f$ and $d$, exactly one of the two holds:*

- *There is a degree-$d$ $\widetilde{\mathbb{E}}$ so that $\widetilde{\mathbb{E}}f < 0$.*
- $\vdash_d f \geq 0$.

The exactly duality holds only on the hypercube; in general, we can relax these things to hold up to $\pm\epsilon$.

We prove the following slight weakening:

**Lemma 1.10.** *For all f,d, exactly 1 of:*

1. *There is a degree $d$ PE so that $\widetilde{\mathbb{E}}f < 0$.*
2. *For all $\epsilon > 0$, $\vdash_d f + \epsilon \geq 0$.*

*Proof.* We saw above that at most 1 of the above occur. Now we show that at least 1 of them occur. Note that $K_d := \{g \ : \vdash_d g \geq 0\}$ is a convex cone. (This cone is closed, which can be proven by arguing carefully, but we will argue about the closure of this cone to avoid having to deal with those details.)

First suppose that $f \in cl(K_d)$. We have that $f_1, f_2, \ldots, \to f$ so that each $f_j$ has a SoS proof. For any $\epsilon > 0$, we can choose $t$ so taht $\vdash_d \epsilon - (f_t - f) \geq 0$. Why can we do this? From fact 5 from earlier today, we can always certify some upper bound that goes to 0 as the size of the coefficients goes to 0. We can do that since $f_t \to f$ here.

We can write $f+\epsilon = f_t + (\epsilon - (f_t - f))$, and we know that $f_t$ and $\epsilon - (f_t - f)$ both have degree-$d$ certificateso f nonnegativity, meaning that $f + \epsilon$ does.

What if $f \notin cl(K_d)$. By the hyperplane separation theorem, we can separate $f$ from $cl(K_d)$ with a hyperplane. In particular, there is a linear map $L : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ so that $Lg \geq 0$ for $g \in K_d$ and $Lf < 0$. We claim that we can rescale $f$ to be a pseudoexpectation. (There is a technical issue here, namely that we need $L$ to act on multilinear polynomials, namely the quotient ring). The idea is that $L(p_i(x))^2 \geq 0$, which holds by definition of $K_d$.

We also want $L1 > 0$. Choose $\alpha$ so that $\vdash_d f + \alpha \geq 0$. We write

$$L1 = L\left(1/\alpha \cdot (\alpha + f - f)\right) = L(1/\alpha(\alpha + f)) - Lf > 0,$$

as desired. We have used that $Lf < 0$ and $L(\alpha + f) \geq 0$ here (since we have already verified that $L$ applied to a sum of squares is non-negative). This means that we can normalize $L$ (by a non-negative real number) as desired. □

Note that $\widetilde{\mathbb{E}}$ is specified by $n^d$ numbers. Similar to the matrix representation theorem we saw for SoS proofs, there is an analogous matrix representation theorem for pesueodexpectations.

We have one more duality theorem:

**Theorem 1.11** (Algorithmic duality)**.** *There is a* $\mathrm{poly}(n^d, \log 1/\epsilon)$*-time alg which takes* $f$ *and returns either:*

1. *A proof that* $\vdash_d f + \epsilon \geq 0$*;*
2. *Some* $\widetilde{\mathbb{E}}$ *so that* $\widetilde{\mathbb{E}} f < \epsilon$*.*

This uses similar ideas from the algorithmic result that finds a degree-$d$ SoS proof, if it exists.

**Back to Max-Cut.** We know that if it is not the case that $\vdash_d \alpha - G(x) \geq 0$, then there is a PE $\widetilde{\mathbb{E}}$ of degree $d$ so that $\widetilde{\mathbb{E}} G(x) \geq \alpha$ (in particular, the above result shows that $\widetilde{\mathbb{E}}\alpha - G(x) < \epsilon$, and then by moving $G$ to the other side and ignoring the $\epsilon$, we get what we want).

We hope that there is some $y \in \{0,1\}^n$ so that $G(y)$ is not too much less than $\alpha$. The general paradigm is as follows: we want to design an algorithm that takes $\mathbb{E}_\mu x^S$ for all $|S| \leq d$, and then produces some $y \in \{0,1\}^n$. We want to analyze this algoirthm using only SoS-provable facts. This will mean that the algorithm works given $\widetilde{\mathbb{E}} x^S$ (since our analysis only depends on SoS-provability).

Imagine we're given the values $\mathbb{E}_\mu x^S$ so that $\mathbb{E}_\mu G(x) \geq \alpha$; we want to find some $y$ so that $G(y)$ is not much smaller than $\alpha$. It is good enough to sample $\mu'$ on $\{0,1\}^n$ so that, for all $i, j$:

$$\mathbb{E}_{\mu'} x_i = \mathbb{E}_\mu x_i, \qquad \mathbb{E}_{\mu'} x_i x_j = \mathbb{E}_\mu x_i x_j. \tag{2}$$

Sicne $G$ has degree 2, that the 2 moments of $\mu'$ match those of $\mu$ means that the expectation under $\mu'$ of $G$ is equal to the expectation under $\mu$ ot $G$. We can't do this exactly (otherwise we could solve MaxCut exactly). But we formulatethe following relaxed goal: construct a distribution $\mu'$ on $\mathbb{R}^n$ so that the moment matching conditions (2) holds. This implies that $\mathbb{E}_{\mu'} G(x) \geq \alpha$.

We will in fact define a Gaussian distribution: its mean vector $v \in \mathbb{R}^n$ is defined by $v_i = \mathbb{E}_\mu x_i$, and its covariance is defined by $\Sigma_{ij} = \mathbb{E}_\mu x_i x_j - (\mathbb{E}_\mu x_i)(\mathbb{E}_\mu x_j)$. Note that we can access the mean and covariance of $\mu$ by looking at degree-2 polynomials. We now set $\mu' = \mathcal{N}(v, \Sigma)$. For this to we

well-defined, we need $\Sigma$ to be PSD (which it is since $\Sigma$ is the covariance of $\mu$, which is a probability distribution).

What if we're given a pseudoexpectation $\widetilde{\mathbb{E}}$? We need to check that $\Sigma \succeq 0$. In particular, we need to check that for all $u$, $u^\top \Sigma u \geq 0$:

$$u^\top \Sigma u = \sum_{i,j} u_i u_j (\widetilde{\mathbb{E}}(x_i x_j) - \widetilde{\mathbb{E}}(x_i)\widetilde{\mathbb{E}}(x_j)) = \widetilde{\mathbb{E}} \sum_{i,j} u_i u_j (x_i x_j - \widetilde{\mathbb{E}} x_i \widetilde{\mathbb{E}} x_j) = \widetilde{\mathbb{E}} \left( \langle u, x \rangle - \widetilde{\mathbb{E}} \langle u, x \rangle \right)^2 \geq 0, \tag{3}$$

where the last step uses the fact that a pseudoexpectation applied to a square is non-negative. (Here, we crucially use that covariance is PSD since expectation of squares is non-negative.)

Consider the following ALG:

1. Sample $g \sim \mathcal{N}(v, \Sigma) \in \mathbb{R}^n$.
2. We need to round $g$ to the hypercube: so set

$$z_i = \begin{cases} 1 & : g_i \geq 1/2 \\ 0 & : g_i \leq 1/2. \end{cases} \tag{4}$$

We finally sketch the analysis that the resulting $z$ has cut value at least $\alpha$ times the GW constant factor.

Goal is to show that $\mathbb{E}[G(z)] \geq c \cdot \alpha$, for some constant $c$. We will do this term by term. Remember that $\mathbb{E}G(z) = \sum_{i,j} \mathbb{E}(z_i - z_j)^2$. We will show that each term (which is simply the probabilitythat $i, j$ is cut by $z$) is at least what it should have been, times $c$. In particular:

$$\mathbb{E}(z_i - z_j)^2 \geq c \cdot \mathbb{E}_\mu (x_i - x_j)^2 = c \cdot \mathbb{E}_{g \sim N(v,\Sigma)} (g_i - g_j)^2.$$

Here we've gotten rid of the graph, which essentially gets rid of the main problem. So the resti s just some analytical calculations. WLOG, we take $v_i = \mathbb{E}_\mu x_i = 1/2$ (since we can always flip 0 and 1, which doesn't change the cut polynomial). Furthermore, $\mathbb{E}_\mu x_i^2 = \mathbb{E}_\mu x_i = 1/2$ since we're on the hypercube.

So, the question can be rephrased as follows: for a Gaussian:

$$N(\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 1/4 & \rho \\ \rho & 1/4 \end{pmatrix}), \tag{5}$$

what is the minimum ratio (over $\rho$):

$$\frac{\Pr_g(g_i > 1/2, g_j < 1/2 \text{ or vice versa})}{\mathbb{E}(g_i - g_j)^2}.$$

This ratio sits above $1/2$ – it is roughly 0.878.

The above argument shows that if there is a PE $\widetilde{\mathbb{E}}$ of degree 2 so that $\widetilde{\mathbb{E}}G(x) \geq \alpha$, then there is $y \in \{0,1\}^n$ so that $G(y) \geq 0.8\alpha$. Thus there is a degree 2 SoS proof that $\vdash_2 \frac{1}{0.8} \cdot MaxCut(G) - G(x) \geq 0$. (If not, then a pseudoexpectation that this is not the case would exist, meaning that there is some $y \in \{0,1\}^n$ so that $G(y)$ exceeds the value of the maximum cut of $G$, which is a contradiction.)

# 2 September 16, 2022

Some history of SoS:

- 1960s: Krivine & Stengle prove that each nonnegative polynomial over a simealgebraic set can be certified nonnegative by an SoS proof.
- 1987: Shor proposes precursor to SoS algorithm (relate polys to semidefinite programs).
- 1990s/2000s: LP/SDP/eigenvalue methods in TCS/optimization.
- 2000s: Lasserre proposes pseudoexpectation SDP, independent work by Parrilo.
- 2010s: SoS as unifying view on LP, SDP, spectral algs; new applications.

## 2.1 Review

Recall that a SoS proof of nonnegativity of a function on the hypercube, denoted $\vdash_d f \geq 0$, us a family of polynomials $p_1, \ldots, p_n$, with $\deg p_i \leq d$, so that $f = \sum_i p_i^2$ over $\{0,1\}^n$.

The dual object to a SoS proof is a pseudoexpectation $\widetilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$, which is linear, repspects $\widetilde{\mathbb{E}}[x^S x_i^2] = \widetilde{\mathbb{E}}[x^S]$ for multisets $S$, $\widetilde{\mathbb{E}}[p^2] \geq 0$, and $\widetilde{\mathbb{E}}[1] = 1$.

We can search for a pseudoexpectation with $\widetilde{\mathbb{E}}[f] < 0$ in time $n^{O(d)}$. We can represent pseudoexpectations as semidefinite matrices. Think of $\widetilde{\mathbb{E}}$ as representing low-degree moments of distributions.

Main duality fact: for ever $f$ of degree at most $d$, either $\vdash_d f \geq 0$, or there is a degree-$d$ PE so that $\widetilde{\mathbb{E}}[f] < 0$.

The idea of taking a pseudoexpectation and sampling a Gaussian whose mean and covariance are the same as a PE (which we used for max-cut) is super useful. The same idea gives approximation algorithms for: $\max_x x^\top A x$ for $A \succeq 0$, and also $\max_{x,y} x^\top A y$ (the cut norm/Grothendieck's ienquality). This idea also forms the basis for best-known approximations algos for graph expansion (Arora-Rao-Vazirani).

## 2.2 Structured instances for max-cut: hyperfiniteness

We first talk about hyperfinite graphs.

**Definition 2.1.** A graph $G$ is $(C, \epsilon)$ *hyperfinite* if we can remove $\epsilon \cdot |E|$ edges so that all connected components have size $\leq C$.

Why is max-cut easy here? Suppose we knew the decomposition into small components. On each constant-size connected component, do brute force search. Then you glue the cuts for each component together arbitrarily. This structured cut cuts at least $(1 - O(\epsilon)) \cdot OPT$ edges; this is because we have ignored only $\epsilon$ fraction of the edges, and that $\epsilon$ fraction is at most a $2\epsilon$ fraction of OPT (since $OPT \in [|E|/2, |E|]$).

Main point: SoS doesn't care if we don't know the decomposition, but it will function as if we did.

**Theorem 2.1.** *Suppose $G$ is a $(C, \epsilon)$-hyperfinite graph. It holds that*

$$\max_{\deg\ O(c)\ \widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}[G(x)] \leq (1 + O(\epsilon)) \cdot \max_y G(y).$$

By duality, there are SoS proofs which certify an upper bound of $(1+O(\epsilon))$ times the maximum value. By doing binary search, you can get an approximation of the max value; Sam doesn't know if there is a rounding algo to get the cut (probably there is).

*Proof.* Consider a PE $\widetilde{\mathbb{E}}$. Then by defn $\widetilde{\mathbb{E}}G(x) = \widetilde{\mathbb{E}}\sum_{i\sim j}(x_i - x_j)^2$. Now let's decompose this according to hyperfiniteness:

$$\widetilde{\mathbb{E}}G(x) = \sum_C \widetilde{\mathbb{E}}G_C(x) + \sum_{(i,j)\in S} \widetilde{\mathbb{E}}(x_i - x_j)^2,$$

where $G_C(x)$ is the cut polynomial restricted to component $C$. Here $|S| \le \epsilon|E|$.

We didn't prove this last time, but a degree-$2n$ PE on $n$ variables corresponds to an actual expectation over a distribution. Therefore, since (by assumption) $G_C$ is a degree-$2C$ PE on the component $C$, we have $\widetilde{\mathbb{E}}G_C(x) \le OPT_C$.

Furthermore, we can find an explicit SoS proof that $\widetilde{\mathbb{E}}(x_i - x_j)^2 \le 1$. Thus, we see that

$$\widetilde{\mathbb{E}}G(x) \le \sum_C OPT_C + \epsilon|E| \le \sum_C OPT_C + 2\epsilon \cdot OPT.$$

Furthermore, $OPT \ge \sum_C OPT_C$. So, we get $\widetilde{\mathbb{E}}G(x) \le (1+2\epsilon)\cdot OPT$. □

**Remarks.** If you believe unique games conjecture, then the worst-case approximation ratio doesn't improve for SoS algo for worst-case max cut even if you go up to degree-$C$. An important direction is to understand whether SoS can do this sort of thing, without relying on unproven conjectures. Furthermore, it is known that approximating max-cut up to $1+\epsilon$ is NP-hard.

### 2.3 Structured instances for max-cut: dense graphs

Below theorem is roughly due to Barak-Raghavendra-Steurer.

**Theorem 2.2.** *If $G$ is dense (i.e., it has $\Omega(n^2)$ edges), then*

$$\max_{\deg\widetilde{\mathbb{E}}\le \mathrm{poly}(1/\epsilon)} \widetilde{\mathbb{E}}G(x) \le (1+O(\epsilon))\cdot OPT_G.$$

To prove the above theorem, need some more prelims.

**Local distributions.** Suppose that $\deg\widetilde{\mathbb{E}} \ge d$. Then for all $S \subset [n]$, $|S| \le d/2$, there exists $\mu_S : \{0,1\}^{|S|} \to \mathbb{R}_+$ so that for all $T \subseteq S$, $\mathbb{E}_{\mu_S} x^T = \widetilde{\mathbb{E}}x^T$. If we look at two such subsets $S, S'$, then the distributions $\mu_S, \mu_{S'}$ agree on the subset $S \cap S'$ since $\widetilde{\mathbb{E}}x^T$ is constant for $T \subset S \cap S'$. The reason that this holds is the same as that which we said before: the degree of $\widetilde{\mathbb{E}}$ is at least twice the size of any such subset $S$.

Example: consider a triangle. We create a distribution on each edge: the distribution on $\{x_i, x_j\}$ takes $(1,0)$ with probability $1/2$ and $(0,1)$ with probability $1/2$. Why do these agree locally: the marginal distribution on $\{x_i\}$ is $Ber(1/2)$. So, no matter which 2-variable local distribution you start with, the marginal on the vertex forming the intersection is $Ber(1/2)$. But these are not globally consistent:

$$\mathbb{E}(x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_1)^2.$$

Above is not defined, but we can define it by decomposing into each pair of 2 variables and using the coresponding local distribution:

$$\mathbb{E}_{\mu_{12}}(x_1 - x_2)^2 + \mathbb{E}_{\mu_{23}}(x_2 - x_3)^2 + \mathbb{E}_{\mu_{13}}(x_1 - x_3)^2 = 3.$$

And there is no distribution which cuts 3 edges in expectation.

Let's consider the matrix representation of a pseudoexpectation which corresponds to the local distributions. Namely, for $\widetilde{\mathbb{E}}[x_i x_j]$, we write down the corresponding local moment. The corresponding table is:

$$\begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}.$$

The above is not PSD. But you can replace the 0s with $1/8$, and the resulting matrix is PSD, and has pseudoexpectation 2.5. (But SoS can certify upper bound better than 3, so you can't push this counterexample all the way up to 3.)

**Remark.** The local distribution $\mu_S$ is unique, given $\widetilde{\mathbb{E}}$ of degree $d$ and $S \subset [n]$, $|S| = t \ll d$. This is because $\mu_S$ is defined uniquely by $\{\widetilde{\mathbb{E}}x^T : T \subset S\}$, which is fixed.

**Remark.** It turns out that local distributions are strictly weaker than SoS: difference between Sherali-Adams and Lasserre-Parrillo.

**Conditioning.** Think about conditioning on $x_i = 1$ or $x_i = 0$. Given $\widetilde{\mathbb{E}}$ so that $\widetilde{\mathbb{E}}x_i > 0$ (for actual distributions, if expectation is 0, then it never can take the value 0).

Define a new linear opterator $\widetilde{\mathbb{E}}[\cdot | x_i = 1] : \mathbb{R}[x]_{\leq \deg \widetilde{\mathbb{E}} - 2} \to \mathbb{R}$ as follows:

$$\widetilde{\mathbb{E}}[p(x) | x_i = 1] := \frac{\widetilde{\mathbb{E}}[p(x) \cdot x_i]}{\widetilde{\mathbb{E}}x_i}.$$

This is completely analogous to how you define a conditional distribution. It is a simple exercise to check that this is a pseudoexpectation of degree $\deg \widetilde{\mathbb{E}} - 2$. Normalization and linearity are trivial. To check positivity:

$$\widetilde{\mathbb{E}}[q(x)^2 | x_i = 1] = \frac{\widetilde{\mathbb{E}}[q(x)^2 x_i]}{\widetilde{\mathbb{E}}x_i} = \frac{\widetilde{\mathbb{E}}[q(x)^2 \cdot x_i^2]}{\widetilde{\mathbb{E}}x_i} \geq 0,$$

where the last equality uses linearity. Similarly, we can define

$$\widetilde{\mathbb{E}}[p(x) | x_i = 0] := \frac{\widetilde{\mathbb{E}}p(x) \cdot (1 - x_i)}{\widetilde{\mathbb{E}}(1 - x_i)}.$$

Note that we have used that $\widetilde{\mathbb{E}}(1 - x_i) \leq 1$, as can be shown by noting that $\widetilde{\mathbb{E}}(1 - x_i) \leq \frac{\widetilde{\mathbb{E}}1 + \widetilde{\mathbb{E}}(1 - x_i)^2}{2} = \frac{1 + \widetilde{\mathbb{E}}(1 - x_i)}{2}$. Similarly we have used that $\widetilde{\mathbb{E}}x_i \leq 1$.

**Remark.** Note that conditioning can be generalized beyond the hypercube as follows: if you replace $x_i$ or $1 - x_i$ above with some square, you boost the probability of the input polynomial appropriately. This is called "tilting".

**Proving Theorem 2.2.** Let's now prove our theorem about dense graphs. We will even give a rounding algorithm that gives us a cut. One really obvious idea to do rounding is independent rounding: given $\widetilde{\mathbb{E}}$, sample $y_i \sim \begin{cases} 0 & \text{wp } 1 - \widetilde{\mathbb{E}}x_i \\ 1 & \text{wp } \widetilde{\mathbb{E}}x_i \end{cases}$. This is the same as taking the local distribution induced by $\widetilde{\mathbb{E}}$ on $x_i$ and sampling according to that local distribution.

Is this a good idea? Note that it is basically the idea behind randomized rounding of LPs (since LPs are basically degree-1 PEs with some linear constraints).

Thought experiment: imagine the PE has local distributions that are close to independent. In particular, look at

$$\mathbb{E}_{i,j\sim[n]}|\mu_{ij} - \mu_i \otimes \mu_j|_{TV} \leq \delta,$$

where we take $i, j$ uniformly and independently from $[n]$. The claim here is that independent rounding works very well:

**Lemma 2.3.** *If the above assumption holds, then*

$$\mathbb{E}G(y) \geq \widetilde{\mathbb{E}}G(x) - \delta n^2.$$

The key here is that if $G$ is dense, then $\delta n^2$ is very small.

*Proof.* Note that

$$\begin{aligned}
\mathbb{E}G(y) &= \sum_{i\sim j} \Pr(y_i \neq y_j) = \sum_{i\sim j} \Pr_{x_i\sim\mu_i, x_j\sim\mu_j}(x_i \neq x_j) \\
&\geq \sum_{i\sim j} \left(\Pr_{(x_i,x_j)\sim\mu_{ij}}(x_i \neq x_j) - |\mu_{ij} - \mu_i \otimes \mu_{ij}|_{TV}\right) \\
&\geq \sum_{i\sim j} \widetilde{\mathbb{E}}(x_i - x_j)^2 - \sum_{i,j} |\mu_{ij} - \mu_i \otimes \mu_j|_{TV} \geq \widetilde{\mathbb{E}}G(x) - \delta n^2.
\end{aligned}$$

Here $\mu_{ij}$ is the local distribution on $(i, j)$ induced by $\widetilde{\mathbb{E}}$. □

We still have to show approximate independence; to do so, we use the pinning lemma:

**Lemma 2.4** (Pinning lemma)**.** *Let $\widetilde{\mathbb{E}}$ be degree $d$, with $d \ll n$. Then there is $t \leq d/2 - 2$ so that if $S \subset [n]$ is random with $|S| = t$ and $y_S \sim \mu_S$, then by pinning the variables in $S$ to $y_S$, we have*

$$\mathbb{E}_{S,y_S}\mathbb{E}_{i,j}\|\mu_{i,j|y_S} - \mu_{i|y_S} \otimes \mu_{j|y_S}\|_{TV} \leq O(1/\sqrt{d}).$$

Here $\mu_{ij|y_S}$ is defined by taking the conditional pseudo-expectation conditioned on $x_S = y_S$ and then defining the corresponding local distribution.

To prove the lemma, what happens if $\mathbb{E}_i\mathbb{E}_j\|\mu_{ij} - \mu_i \otimes \mu_j\|_{TV} \gg \delta$? If we take a typical coordinate $i$, then it is correlated with most other coordinates $j$. In particular, if we know $i$, then we learn a lot about most other coordinates. If this quantity stayed large, we could do the same thing again, and so on. But the idea is there is only so much to learn about all the coordinates in the graph. At this stage you've broken the local correlations. To do it formally, we use information theory.

**Information theory background.** We use $H$ to denote entropy and $I$ to denote mutual information.

1. If $X, Y \in \{0,1\}$, then $H(X), I(X;Y) \in [0,1]$.
2. $I(X;Y) = H(X) - H(X|Y)$.
3. Pinsker's inequality: $|\mu_{XY} - \mu_X \times \mu_Y|_{TV} \leq \sqrt{I(X;Y)/2}$. All we need is that this holds up to a polynomial.

Now we prove the pinning lemma.

*Proof of Lemma 2.4.* Given $\widetilde{\mathbb{E}}$, define a random sequence of PEs as follows: $\widetilde{\mathbb{E}}^0 = \widetilde{\mathbb{E}}, \widetilde{\mathbb{E}}^1, \ldots, \widetilde{\mathbb{E}}^{d/2}$. Here $\widetilde{\mathbb{E}}^s$ is given as follows: draw $i \sim [n] \backslash \{$ indices used so far $\}$, draw $y_i \sim \mu_i^{s-1}$ (the local distribution on coordinate $i$ under $\widetilde{\mathbb{E}}^{s-1}$), and then set $\widetilde{\mathbb{E}}^s = \widetilde{\mathbb{E}}^{s-1}|(x_i = y_i)$.

**Exercise.** The above distribution over PEs leads to some $\widetilde{\mathbb{E}}^{d/2}$ which has the same distribution as if we were to choose a random subset of size $d/2$ and condition on the variables in that subset. A related (more basic) exercise is that we can define conditioning on $\{x_i\}_{i \in S} = \{y_i\}_{i \in S}$ by $\widetilde{\mathbb{E}}[\cdot | y_S = x_S] = \frac{\widetilde{\mathbb{E}}[\cdot \mathbb{1}\{y_S = x_S\}]}{\widetilde{\mathbb{E}} 1 \cdots}$.

Define $glob_S = \mathbb{E}_{i,j} I(X_i^s; X_j^s)$ joint from $\mu_{ij}^s$, where $X_{ij}^s$ is a sample from the local distribution for coordinates $(i,j)$ for the pseudoexpectation $\widetilde{\mathbb{E}}^s$.

Fix a time $s$ where you stop. If $\mathbb{E} glob_s \leq \delta$ (outer expectation is over all the randomness that defines expectations), then $\mathbb{E}\mathbb{E}_{ij} |\mu_{ij}^s - \mu_i^s \otimes \mu_j^s|_{TV} \leq O(\sqrt{\delta})$. This is just pinsker's inequality.

So, it suffices to show that there is some $s$ so that $\mathbb{E} glob_S \leq O(1/d)$.

Define a potential function:

$$\Phi^s := \mathbb{E}\mathbb{E}_{i \sim [n]} H(X_i^s),$$

where $X_i^s$ is drawn from its 1-wise marginal under the local distribution for $\widetilde{\mathbb{E}}^s$. The idea is that this potential function must drop if there's a lot of local correlation:

**Claim 2.5.** $\Phi^s - \Phi^{s+1} \geq \Omega(\mathbb{E} glob_S)$.

Note that when we condition on some variables we will no longer sample from them in the future, and so we have to be careful about $i \sim [n]$ (the istribution isn't completely uniform, so we lose some constant factor in the above claim, which is ok).

*Proof of lcaim.* Recall that

$$I(X_i^s; X_j^s) = H(X_i^s) - H(X_i^s | X_j^s).$$

Now let's average both sides over all $i, j$. Certainly $\mathbb{E}_{i \sim [n]} H(X_i^s) = \Phi^s$. Moreover,

$$\mathbb{E}_j \mathbb{E}_i H(X_i^s | X_j^s) = \Phi^{s+1}. \tag{6}$$

This holds because to get $\widetilde{\mathbb{E}}^{s+1}$ we choose a coordiante $j$ at random and then condition on its value drawn from the distribution $\mu_j^s$. The average entropy of $X_i^s$ under this conditioning is exactly the definition of conditional entropy.

Moreover, $\mathbb{E}_{ij} I(X_i^s; X_j^s) = glob_S$ by definition. □

As a result of the above claim, we get that

$$1 \geq \Phi^0 - \Phi^{d/2} \geq \Omega\left(\sum_{s=0}^{d/2} \mathbb{E} glob_S\right) \geq 0,$$

where the final inequality followos since mutual information is non-negative. Then we get that there is some $s$ so that $\mathbb{E} glob_S \leq O(1/d)$ by an averaging argument. □

Now we prove the dense max-cut theorem.

*Proof of Theorem 2.2.* Note that we can find the low global correlation distributions guaranteed by the pinning lemma in polynomial time: we can simply do brute force search over all subsets $S$ of size at most $d/2$, and compute the conditional distributions: all of these steps take time $n^{O(d)}$, and we already need that much time for solving SDPs.

Given $\widetilde{\mathbb{E}}$, we know from the pinning lemma that there exists $s \leq \text{poly}(1/\epsilon)$ so that

$$\mathbb{E}_{\widetilde{\mathbb{E}}' \sim pinning} \widetilde{\mathbb{E}}_{ij} \|\mu'_{ij} - \mu'_i \otimes \mu'_j\|_T V \leq \epsilon^{10}, \tag{7}$$

where here $\widetilde{\mathbb{E}}', \mu'$ denote the pinned pseudoexpectation and local distributions. Furthermore, we have that

$$\mathbb{E}_{pinning} \widetilde{\mathbb{E}}' G(x) = \widetilde{\mathbb{E}} G(x).$$

This second statement is an exercise; the idea is that when we do the conditioning we sample things from the correct distribution at each step.

By Markov's inequality, with probability $1 - \epsilon^5$, $\mathbb{E}_{ij} \|\mu'_{ij} - \mu_i \otimes \mu'_j\|_{TV} \leq \epsilon^5$.

Similarly, with probability at least $10\epsilon^5$, $\widetilde{\mathbb{E}}' G(x) \geq (1-\epsilon) \cdot \widetilde{\mathbb{E}} G(x)$, where we use that $\widetilde{\mathbb{E}} G(x), \widetilde{\mathbb{E}}' G(x)$ are between 0 and $n^2$; in particular, we're using Paley-Zygmund here.

Thus, there exists $\widetilde{\mathbb{E}}'$ so that independent rounding gives $\mathbb{E}_{y \sim \mu'} G(y) \geq (1 - O(\epsilon)) \cdot \widetilde{\mathbb{E}} G(x)$. □

Note that the algorithm is very simple: solve for a PE $\widetilde{\mathbb{E}}$ maximizing $\widetilde{\mathbb{E}} G(x)$ using SDP. Consider all subsets of at most $\text{poly}(1/\epsilon)$ variables and assignments of those variables, and consider the resulting pinned pseudoexpectation. Then for that pseudoexpectation, do independent rounding, which will give a large cut value in expectation (Lemma 2.3).

## 2.4 Max-cut on structured instances

So far, we have only used Sherali-Adams: we have only looked at the local distributions. As we saw, this is weaker than SoS since local distributions (even on a triangle) can fool you more than SoS in terms of the max-cut value.

Suppose $G$ is $\Delta$-regular, and let $A_G$ be the normalized adjacency matrix. Let the eigenvalues of $A_G$ be $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We have that $\lambda_2 \approx \epsilon \ll 1$ if and only if $G$ is an expander. Roughly speaking, expanders are sparse approximations of the complete graph.

**Theorem 2.6.** *For every $\Delta$-regular $G$,*

$$\vdash_d G(x) \leq \left(1 + d^{-\Omega(1)} + \lambda_2^{\Omega(1)}\right) \cdot \max_y G(y).$$

The above is a generalization of the fact for dense graphs. The diea is that if $\lambda_2$ is small, then the approximation factor is close to 1.

To prove the above theorem, we use a similar approach to previously. What went badly with independent rounding? We cared about the error:

$$\sum_{i\sim j} |\Pr(x_i \neq x_j \sim \mu_i \otimes \mu_j) - \Pr(x_i \neq x_j \sim \mu_{ij})|_{TV} \leq \sum_{i\sim j} \|\mu_{ij} - \mu_i \otimes \mu_j\|_{TV} \leq \sum_{i,j\in[n]} \|\mu_{ij} - \mu_i \otimes \mu_j\|_{TV}.$$

In a dense graph, the second inequality is fine, but in a sparse graph, it is very loose.

For expander graphs we have a local to global phenomenon.

**Definition 2.2.** Given a degree-2 PE $\widetilde{\mathbb{E}}$ over $\{0,1\}^n$, define $y_i(x_i) = 2x_i - 1$, and it is straightforward to check that $p(y) \mapsto \widetilde{\mathbb{E}}p(y)$ is a pseudoexpectation over $\{-1,1\}^n$. We define

$$Cov_{ij} = \widetilde{\mathbb{E}}y_iy_j - \widetilde{\mathbb{E}}y_i\widetilde{\mathbb{E}}y_j.$$

Define $globCorr = \mathbb{E}_{i,j}Cov_{ij}^2$ (expectation over all pairs $(i,j)$). Further, define $localCorr = \mathbb{E}_{i\sim j}Cov_{ij}^2$. In particular, local correlation is over all edges of the graph.

**Lemma 2.7.** *For $\{0,1\}$-valued random variables $X, Y$, $Cov(X,Y)^2 \leq O(I(X;Y))$. Furthermore, $Cov(X,Y) \geq \text{poly}(\|P_X \otimes P_Y - P_{XY}\|_{TV})$.*

We don't prove the above (standard/easy fact).

the idea is as follows: we can get upper bound on global information, and thus global correlation, using pinning. We can also get a lower bound on local correlation. So we want to relate local and global correlation. To do so, we prove:

**Lemma 2.8.** *Let $\widetilde{\mathbb{E}}$ be a degree-4 PE. Then $localCorr \leq globalCorr + O(\lambda)$.*

Given the above, we can show that the above lemma can be plugged into everything we did today about pinning and global correlation to get the desired result about max-cut in expanders.

*Proof.* We ahve

$$v^\top A_G v = \frac{1}{\Delta}\sum_{i\sim j} v_iv_j = \frac{1}{n}\sum_{i,j} v_iv_j + \sum_{2\leq j\leq n} \lambda_j \langle v, w_j\rangle^2,$$

where $w_j$ is the $j$th eigenvector of $A_G$, normalized so that $\|w_j\| = 1$.

Then let's look at local correlation:

$$\begin{aligned}\mathbb{E}_{i\sim j}(\widetilde{\mathbb{E}}y_iy_j - \widetilde{\mathbb{E}}y_i\widetilde{\mathbb{E}}y_j)^2 &= \mathbb{E}_{i\sim j}(\widetilde{\mathbb{E}}[(y_i - \widetilde{\mathbb{E}}y_i)(y_j - \widetilde{\mathbb{E}}y_j)])^2 \\ &= \mathbb{E}_{i\sim j}\widetilde{\mathbb{E}}(y_i - \widetilde{\mathbb{E}}y_i)(y_i' - \widetilde{\mathbb{E}}y_i')(y_j - \widetilde{\mathbb{E}}y_j)(y_j' - \widetilde{\mathbb{E}}y_j') \\ &= \widetilde{\mathbb{E}}\mathbb{E}_{i\sim j}(y_i - \widetilde{\mathbb{E}}y_i)(y_i' - \widetilde{\mathbb{E}}y_i')(y_j - \widetilde{\mathbb{E}}y_j)(y_j' - \widetilde{\mathbb{E}}y_j').\end{aligned}$$

The first equality is because everything is only on $O(1)$ variables, so we need the pseudoexpectation to be of constant degree.

Define $V_i(y,y') = (y_i - \widetilde{\mathbb{E}} y_i)(y_i' - \widetilde{\mathbb{E}} y_i')$. Then we have that the above is equal to

$$\frac{1}{n\Delta}\widetilde{\mathbb{E}}\sum_{i\sim j} V_i(y,y')V_j(y,y') = \frac{1}{n}\widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i,j} V_i(y,y')V_j(y,y') + \sum_{2\le j\le n}\lambda_j\langle v, w_j\rangle^2\right]$$
$$= \widetilde{\mathbb{E}}\mathbb{E}_{i,j} V_i(y,y')V_j(y,y') + \frac{1}{n}\widetilde{\mathbb{E}}\sum_{2\le j\le n}\lambda_j\langle v, w_j\rangle^2.$$

We claim that the first term above is the global correlation (do the same manipulations as previously in reverse). To deal with the second (error) term, we want ti to be at most $\lambda_2$. Here we go beyond PEs: we use pseudoexpectations applied to squares of polynomials that depend on more than a few variables! In particular, we write

$$\lambda_2 \cdot I - \sum_{2\le j\le n}\lambda_j \cdot w_j w_j^\top \succeq 0.$$

This is because the $w_j$ are orthogonal and $\lambda_2 \ge \lambda_j$ for $j \ge 2$. Let's take the Cholesky decomposition of the above matrix, and in particular we have

$$\widetilde{\mathbb{E}}\left[v^\top(\lambda_2 \cdot I - \sum_{2\le j\le n}\lambda_j w_j w_j^\top)v\right] = \widetilde{\mathbb{E}}\sum_i p_i(v)^2 \ge 0,$$

where the final inequality follows because we have a sum of squares. Rearranging, we get that

$$\widetilde{\mathbb{E}}\sum_{2\le j\le n}\lambda_j v^\top w_j w_j^\top v \le \frac{1}{n}\lambda_2 \cdot \widetilde{\mathbb{E}} V(y,y')^\top I V(y,y') = \frac{\lambda_2}{n}\widetilde{\mathbb{E}}\sum_i V_i(y,y')^2 \le \lambda_2,$$

where the final inequality follows since the $V_i$ are bounded by at most 1 (which holds in pseudoexpectation, by locality). □

# 3 September 30, 2022

Today we will tlak about refuting random CSPs (consttraint satisfaction problems). A *CSP* is defined by a predicate $\phi : \{-1,1\}^k \to \{0,1\}$. We will assume that $\mathbb{E}_{x\sim\{-1,1\}^k}[\phi(x)] < 1$; in particular, $\phi$ is not satisfied with probability 1. In particular, an instance of a CSP consists of:

1. Variables $x_1, \dots, x_n$.
2. Tuples $S_1, \dots, S_m \in [n]^k$, which say which variables are contained in each predicate.
3. Vectors $y_1, \dots, y_m \in \{-1,1\}^k$, which specify the negation pattern for each predicate.

The constraints of a CSP then consist of the following $m$ conditions:

$$\{\phi(x_{S_i} \circ y_i) = 1\}_{i=1}^m.$$

For instance, if $S = (1,2)$ and $y = (1,-1)$, then $x_{S_1} \circ y_1$ is the string $(x_1, -x_2)$. The individual constraints above are often called *clauses*.

Examples of CSPs:

- Max-Cut: $\phi(x_0, x_1)$ is the not-equal predicate for $x_0, x_1 \in \{-1, 1\}$.
- SAT: $\phi = OR$.
- NAE-SAT: $\phi$ is 1 if at least one of the $x_i$ is 1, and not all are 1.

In the worst case, CSPs are notoriously hard, but there are empirical solvers that work well. So, there is something tractable going on. This is some motivation to study random instances of CSPs.

### 3.1 Random CSPs

To generate a an $m$-clause instance, we simply sample $S_1, \dots, S_m, y_1, \dots, y_m$ uniformly at random. Today, we will include every $S \in [n]^k$ with probability $m/n^k$.[1] So, the expected number of clauses is $m$. We also draw $y_1, \dots, y_m \sim \{-1, 1\}^k$ independently at random.

Given $\varphi \sim CSP_\phi^{n,m}$, we could ask the question, what is:

$$\max_{x \sim \{-1,1\}^n} \varphi(x),$$

where $\varphi(x) = \sum_{i=1}^m \phi(x_{S_i} \circ y_i)$, i.e., what is the maximum number of clauses we can satisfy? It turns out that due to concentration, we have $\max_x \varphi(x) \approx \mathbb{E}_\varphi \max_x \varphi(x)$ with high probability, so we can simply output the expected value, so it's not a very intereseting problem.

Another possible question we could ask is to find $x$ which is the argmax of the above maximization problem. It turns out that therea re two regimes here: if $m \gg n$, there are typically no interesting $x$s, in the sense that $\max_x \varphi(x) \approx \mathbb{E}_{x \sim \{-1,1\}^n} \varphi(x)$, which is an easy concentration argument.

On the other hand, if $m \ll n$, the CSP is fully satisfiable, in which case there are many $x$s so that $\varphi(x) = m$ (i.e., can satisfy all the constraints). You can find such $x$ via local search. There's a very small window when you transition between the two regimes. This is an interesting regime, and algorithms to solve this problem in that regime are not well understood.

So, for us, even finding a good satisfying assignment doesn't lead to interesting problems.

### 3.2 Refutation

We will get interesting problems by considering proofs of certain statements. In particular, we consider the problem of *(strong) refutation*: given $\varphi$, output $ALG(\varphi) \in [0, m]$ so that:

1. For all $\varphi$, $ALG(\varphi) \geq \max_x \varphi(x)$. (This is interesting because it has to hold for all $\varphi$.)
2. $\mathbb{E} ALG(\varphi) \leq (1 - \delta)m$ where $\delta = \Omega(1)$ as $n, m \to \infty$.

We remark that weak refutation allows for $\delta = o(1)$ (e.g., $1/m$), and we want with high probability $ALG(\varphi) \leq (1 - \delta)m$.

This is interesting in the regime that $m \gg n$. Note that if we design such an algorithm, then we've automatically generated a proof system that can prove interesting upper bounds on the satisfiability of certain CSPs. This is because the trace of the ALG's computation is of polynomial length, and so that computation trace run on $\varphi$, which generates an output $\leq (1 - \delta)m$, is a proof

[1] Technically, you should actually include each pair $(S, y)$ with probability $m/(2n)^k$. This allows for the same $S$ to appear with different negation patterns $y$.

that $\varphi(x) \leq (1-\delta)m$. This is interesting since often it's nontrivial to do better than exponential length proofs.

Notee that as $m$ gets larger (e.g., in max-cut, more edges), it gets harder to completely satisfy the CSP, and so the problem of refuting CSPs gets easier. Note that as long as $m \gg n$, the problem is nontrivial, and we always have $m \leq n^k$. In particular, for $m = n^k$, it turns out that the strong refutation problem is easy (we say this for max cut, for dense graphs, and it holds more generally). So the question is where in the range $m \in [n, n^k]$ we can solve the strong refutation problem.

### 3.3 A potential algorithm

For some even integer $d \in \mathbb{N}$, let's see what value SoS can certify. In particular, given $\varphi$, find the least $c$ so that $\vdash_d \varphi(x) \leq c$. This takes time $n^{O(d)}$ (we think of $k$ as a constant).

**Theorem 3.1.** *For all $\phi$ (nontrivial), if $m \gg n^{k/2} \cdot \log^{O(1)} n$, there exists $\delta > 0$ so that with high probability over $\varphi$, $\vdash_{O(k)} \varphi(x) \leq (1-\delta)m$.*

The first step in the proof is to reduce to one particular CSP, in particular $k$-XOR, which is defined as follows:

$$\phi(x) = \begin{cases} 1 & : \text{ if } \prod_{i=1}^k z_i = 1 \\ 0 & : \text{ if } \prod_{i=1}^k z_i = -1 \end{cases}.$$

Moreover, for the XOR predicate, $\phi(x_S \circ y)$ depends only on $\prod_{i \in S} x_i \cdot \prod_{i \leq k} y_i$. Thus, the only feature of the literal pattern $y$ that matters is its parity. Thus, WLOG, we can replace $y_i \in \{-1,1\}^k$ with a single boolean value $a_i \in \{-1,1\}$.

**Definition 3.1.** Given a $k$-XOR instance, we define a polynomial:

$$\psi(x) = \sum_{i=1}^m a_i \cdot \prod_{j \in S_i} x_j = \text{ \# sat - \# unsat }.$$

This is not the polynomial we had before, which was counting the number of satisfied assignments. In contrast, $\psi(x)$ counts the number of satisfied minus unsatisfied assignments.

Note that $k$-XOR is interesting because the Fourier decomposition of boolean functions. In particular, a general $k$-CSP problem breaks down into a number of $k'$-XOR CSPs.

Returning to the case of general $\phi$, we write the Fourier decomposition $\phi(z) = \sum_{T \subset [k]} \hat{\phi}_T \cdot z_T$, where $z_T = \prod_{i \in T} z_i$. Now take a random $\varphi \sim CSP_\phi^{n,m}$. We build $2^k$ $k'$-XOR instances, for $k' \leq k$. Now we can write

$$\varphi(x) = \sum \phi(x_{S_i} \circ y_i) = \sum_{T \subset [k]} \hat{\phi}_T \sum_i \prod_{j \in T} y_{ij} \prod_{j \in T} (x_{S_i})_j.$$

where the $z$s are the $x$ times $y$ values. Each $T$ now gives some function $\psi_T$ so that $\varphi(x) = \sum_T \hat{\phi}_T \cdot \psi_T(x)$, where we have defined $\psi_T = \sum_i \prod_{j \in T} y_{ij} \prod_{j \in T} (x_{S_i})_j$.

Note that the $\psi_T$ are indeed random $k'$-XOR instances, for $k' = |T|$. This is because we construct $2^k$ different $k'$-XOR instances for each (random) set $S_i$.

If we have $\vdash_{O(k)} \psi_T(x) \le \epsilon m$ and $\vdash_{O(k)} \psi_T(x) \ge -\epsilon m$ (we need both signs since the $\hat{\phi}_T$ can be negative), then by summing up the proofs, we get that

$$\vdash_{O(k)} \varphi(x) \le \hat{\phi}_\emptyset + \sum_{T \subset [k], T \neq \emptyset} \hat{\phi}_T \cdot \psi_T(x) \le m \cdot \left( \hat{\phi}_\emptyset + 2^k \cdot \epsilon \right).$$

Our assumption on nontriviality gives that $\hat{\phi}_\emptyset < 1$. If we can take $\epsilon \ll 2^{-k}$, then we can refute $\varphi(x)$.

Note that by symmetry in the distribution (since $\psi$ and $-\psi$ show up with the same probability), if we can prove $\vdash_{O(k)} \psi_T(x) \le \epsilon m$ with high probability, then we can prove $\vdash_{O(k)} \psi_T(x) \ge -\epsilon m$ with the same probability.

**Remark.** If it turns out that the Fourier expansion of $\phi(x)$ has all terms of degree at most $k_0$, then we can do everything with $k$ replaced by $k_0$. As an example, for $\phi = 3 - NAE - SAT$, then you can check that $\hat{\phi}_{\{1,2,3\}} = 0$, and os you can refute with $m \approx n$ (as opposed to $m \approx n^{1.5}$) clauses.

**Main tools.**

1. **Spectral SoS certificates**. The following is a very powerful way to generate SoS proofs.

   **Lemma 3.2.** *Suppose that*

   $$f(x) = (x^{\otimes d/2})^\top M x^{\otimes d/2}$$

   *over $\{-1,1\}^n$, for some symmetric matrix $M$. Then $\vdash_d f(x) \le n^{d/2} \cdot \|M\|_\sigma$, where $\|M\|_\sigma$ denotes the maximum singular value of $M$.*

   *Proof.* Note that $\|M\|_\sigma I - M \succeq 0$. Then any PSD matrix can be factorized as a square, as follows:

   $$(x^{\otimes d/2})^\top \cdot (\|M\| I - M) x^{\otimes d/2} \succeq 0,$$

   where we abuse the notation $\succeq 0$ to mean that it can be written as a sum of squares. The LHS of the above is $\|M\| \cdot \|x^{\otimes d/2}\|^2 - f(x)$, which has a SoS proof that it is equal to $\|M\| \cdot n^{d/2} - f(x)$ (since we're working over the $\{-1,1\}$-hypercube), and this is what we want to show. □

2. **Matrix bernstein inequality.** Since are dealing with random CSP instances, we will need to argue about random matrices. We first review the standard Bernstein inequality.

   The idea is as follows: Suppose that $A_1, \dots, A_n$ are random variables so that $|A_i| \le R$, they are independent, and $\mathbb{E} A_i = 0$ for each $i$. Then $\sum_i A_i$ should be roughly a Gaussian with variance $\mathbb{E} \sum_i A_i^2$. This is true in some regime (which doesn't hold when one of the $A_i$ dominates).

   Here's the formal statement of the matrix version:

   **Lemma 3.3.** *Suppose that $A_1, \dots, A_n \in \mathbb{R}^{d \times d}$ are symmetric, random matrices with $\mathbb{E} A_i = 0$ and $\|A_i\| \le R$ with probability 1 for all $i$. Then*

   $$\mathbb{E} \| \sum_i A_i \| \le O\left( \left\| \mathbb{E} \sum_i A_i^2 \right\|^{1/2} \cdot \sqrt{\log d} + R \log d \right),$$

   *where we use $\| \cdot \|$ to denote spectral norm.*

### 3.4 Refutation for 2-XOR

We do the cases $k = 2, 3, 4$. Let's start with $k = 2$. Here we have $\psi(x) = \sum_{i,j} a_{ij} x_i x_j$, where $a_{ij} \in \{-1, 1\}$ if $(i, j)$ is a clause, and 0 otherwise. There is a vanishing probability where some $(i, j)$ shows up both in a negative and positive clause, which we ignore (it's negligible).

We can write $\psi(x) = x^\top (\sum_{i,j} a_{ij} E_{ij}) x$, where $E_{ij}$ is the matrix in $\mathbb{R}^{n \times n}$ with all 0s except for $1/2$ entry at the $(i, j)$ and $(j, i)$ positions.

Some intuition for random (symmetric) matrices. Recall that $\|M\|_F = \sqrt{\sum_{i=1}^d \lambda_i^2}$, where $\lambda_i$ are the eigenvalues of $M$. If $M$ is unstructured then all eigenvalues should be roughly equal in the sense that the top eigenvalue behaves like the average one, so we roughly have $\max_i \lambda_i \approx \|M\|_F / \sqrt{d}$.

Now why should $\sum_{i,j} a_{ij} E_{ij}$ be unstructured in the above sense? In each entry, we have put 0 or $\pm 1$ with some appropriate probability (namely $m/n^2$). As large as $m \gg n$, each row has at least a super-constant number of nonzero entries, which are different whp, so it seems that it should be of full rank. To reason formally about this, we use Matrix Bernstein.

In particular, we need to bound $\mathbb{E} \sum_{i,j} a_{ij}^2 E_{ij}^2$: ignoring constants, we have

$$\mathbb{E} \sum_{i,j} a_{ij}^2 E_{ij}^2 \approx \sum_{i,j} \frac{m}{n^2} E_{ij}^2 \approx \frac{m}{n^2} \sum_{i,j} E_{ii} + E_{jj} \preceq O\left(\frac{m}{n^2} \cdot n \cdot I\right) = O(\frac{m}{n} I),$$

where we have used that $E_{ij}^2 = 1/4 \cdot (E_{ii} + E_{jj})$. Furthermore, we have used the fact that in the final inequality that each $E_{ii}$ appears $n$ times (since there are $n$ values of $j$). We also note that $\|E_{i,j}\| \leq 1$ and $|a_{ij}| \leq 1$, so by Matrix bernstein, we get that

$$\mathbb{E} \| \sum a_{ij} E_{ij} \| \leq O\left(\sqrt{m/n} \cdot \sqrt{\log n} + \log n\right) \leq O(\sqrt{m/n} \cdot \sqrt{\log n}),$$

where we have used that we will take $m \geq n \log n$. By Markov's inequality, we can get that the above holds up to a factor of $2^k$ with probability $1 - 2^{-k}$.

Now we can use spectral SoS certificates with $d = 2$: we have $\vdash_{O(1)} n \cdot \sqrt{m/n} \cdot \sqrt{\log n}$. When is this $\ll \epsilon \cdot m$?, for $\epsilon$ a tiny constant?

This holds as long as $\sqrt{nm \log n} \ll \epsilon m$, i.e., as long as $\frac{n \log n}{\epsilon^2} \ll m$. Thus, with high probability, degree $O(1)$ SoS refutes 2-XOR. (It's actually degree 2 here.)

### 3.5 Refutation for 4-XOR

Now let's do $k = 4$. Now we have $\psi(x) = \sum a_{ijkl} x_i x_j x_k x_\ell$. One thing we can do is view it as a 2-XOR instance in $n^2$ variables where we consider the variables to be $x_i x_j$, for all pairs $(i, j)$.

We will actually argue more directly and sketch out how things can be generalized from above. In particular, we can write:

$$\psi(x) = (x^{\otimes 2})^\top (\sum a_{ijkl} E_{ijkl}) x^{\otimes 2},$$

where $E_{ijkl}$ hs an $n^2 \times n^2$ matrix with nonzero entries at places like $(ij, k\ell), (jk, i\ell)$, and so on. Now we use the same argument as for $k = 2$: we have that

$$\|\mathbb{E} \sum a_{ijkl} E_{ijkl}\|_F^2 \leq m/n^2,$$

and so by Matrix Bernstein we get that $\| \sum a_{ijkl} E_{ijkl} \| \leq \sqrt{m/n^2}$ with high probability. Thus, we have $\vdash_{O(1)} \psi(x) \leq n^2 \cdot m/n = n\sqrt{m} \ll m$ if $m \gg n^2$.

### 3.6 What about $k = 3$?

Turns out the case of odd $k$ is somewhat more challenging. One possibility is to write $\psi(x) = \sum a_{ijk} x_i x_j x_k = x^\top \cdot M \cdot x^{\otimes 2}$, where $M$ is an $n \times n^2$ matrix. We want the matrix to be symmetric, so we can write it as: $(x, x^{\otimes 2})^\top \cdot M' \cdot (x, x^{\otimes 2})$, where $M'$ is $(n + n^2) \times (n + n^2)$ and has the $n \times n^2$ blocks equal to $M$. The issue is if we try to do this, we run into issues because we can't spread out the eigenvlaues over all $n^2$ entries, as $M'$ only has rank at most $2n$. This approach will give refutations using $m = n^2$ clauses, but we want $m = n^{1.5}$, so off by a polynomial factor.

We will search instead for a square matrix: try to use SoS reasoning on $\psi$ to find a different polynomial which is more amenable to being arranged as a square matrix. Let's write

$$\psi(x) = \sum_{ijk} a_{ijk} x_i x_j x_k = \sum_i x_i \sum_{j,k} a_{ijk} x_j x_k.$$

Let's do Cauchy-Schwarz:

$$\psi(x)^2 = (\sum_{ijk} a_{ijk} x_i x_j x_k)^2 \preceq (\sum_i x_i^2) \cdot (\sum_i (\sum_{j,k} a_{ijk} x_j x_k)^2),$$

where the inequality is a SoS proof (this is on the PSET). Note that $\sum_i x_i^2 = n$ since we're on the $\{-1, 1\}^n$ hypercube. So we need to focus on the second polynomial on the RHS. Let's think of the coefficients $a_{ijk}$ as arranged in some 3-tensor. For each $i$, we consider the $i$th slice which is $a_{ijk}$, for $j, k \in [n]$. We write $A_i = (a_{ijk})_{j,k}$ As this matrix, so the whole term is

$$\sum_i (x^\top A_i x)^2.$$

Now we want to write the above as a quadratic form. We have two options:

1. First, we can write

   $$(x^\top A_i x)^2 = (x^{\otimes 2})^\top A_i^\flat (A_i^\flat)^\top (x^{\otimes 2}),$$

   where $A_i^\flat$ is the flattened version of $A_i$, written as an $n^2$-dimensioanl vector. In particular, $(A_i^\flat)_{(j,k)} = A_i(j, k)$. Note that the rank of the matrix $(A_i^\flat (A_i^\flat)^\top)$ is 1, and we sum $n$ of them up, so get rank at most $n$. This is bad, since we want things to be of higher rank. It turns out that if we use matrix Bernstein here, it won't improve on $m = n^2$ clauses.

2. Alternatively, we can write

   $$(x^\top A_i x)^2 = (x^{\otimes 2})^\top (A_i \otimes A_i) x^{\otimes 2}.$$

   The matrix $A_i \otimes A_i$ has rank equal $rank(A_i)^2$, which could be larger (good!). We will go with this approach.

So we will write

$$\sum (x^\top A_i x)^2 = (x^{\otimes 2})^\top (\sum_i A_i \otimes A_i) x^{\otimes 2}.$$

We can't quite use Matrix bernstein sincee we don't have 0 expectation. First, we have to compute:

$$\sum_i (x^{\otimes 2})^\top (\mathbb{E} A_i \otimes A_i) x^{\otimes 2} = p(m,n) \cdot \|x\|^4,$$

where $p(m,n)$ is some small polynomial: just need to check which entries are 0 after expectation.

Now, by matrix Bernstein, we have to make some careful choices to get independent matrices, and we get an upper bound on

$$\| \sum_i A_i \otimes A_i - \mathbb{E} A_i \otimes A_i \|,$$

and then you can run the same argument as $k = 2, 4$. It turns out that $\sum_i A_i \otimes A_i$ has roughly $m$ nonzero rows (and roughly rank $m$), which is better than $n$ nonzero rows which we got from before.

So, so far we have gotten that $\vdash_6 \psi(x)^2 \le \epsilon m$ whp if $m \ge n^{1.5} \cdot \mathrm{poly}(\log n, 1/\epsilon)$.

How do we actually get an upper bound on $\psi(x)$? We use the following basic fact:

**Lemma 3.4.** *If $\vdash_d f^2 \le B$, then $\vdash_d f \le \sqrt{B}$.*

*Proof.* We use that

$$f = f \cdot \frac{1}{B^{1/4}} \cdot B^{1/4} \preceq \frac{1}{2} \cdot \left( \frac{f^2}{\sqrt{B}} + \sqrt{B} \right) \preceq \frac{1}{2} \cdot \left( \frac{B}{\sqrt{B}} + \sqrt{B} \right) = \sqrt{B}.$$

□

### 3.7 Application: tensor completion

Let's first recall the problem of matrix completion. We are given some matrix $A \in \mathbb{R}^{U \times M}$ (say users and movies), where $A_{um} \in [-1, 1]$ denotes whethe user $u$ likes movie $m$. We are given some entries, and we want to fill in the matrix so that it is low rank. In particular, we assume that $M = \sum_{i=1}^r u_i v_i^\top$ for some small $r$. The idea is that $u_i$ represents features of users and $v_i$ represents features of movies.

The number of parameters need to specify $M$ if it is of rank $r$ is $O(r \cdot (m+n))$, where we write the dimensions as $A \in \mathbb{R}^{n \times m}$ now. It turns out that you can do this if you're given $O(n+m)$ entries of $A$ with $r = O(1)$, in polynomial time.

What about tensor completion? Let's consider a 3-tensor (think users, restaurants, and times of day – Sam calls this the "Yelp problem"). Let's now assume we have a low-rank tensor $T = \sum_{i=1}^R u_i \otimes v_i \otimes w_i$.

If we flatten $T$ into a matrix of the form $\sum_i u_i (v_i \otimes w_i)^\top$, the resulting matrix has rank $r$ as well. How well does this do? If all dimensions are $n$, then $T$ is specified by $O(rn)$ entries. The matrix completion problem requires that we need $O(r \cdot (n + n^2))$ entries since we need to scale with the long dimension.

Can we do this with fewer than $O(rn^2)$ entries?

**Theorem 3.5** (Barak & Moitra)**.** *There exists a poly-time algorithm to complete $T$ using $\tilde{O}(n^{1.5})$ entries.*

The $n^{1.5}$ should look a lot like the number of clauses needed to refute a random 3-CSP.

Let's say more formally what the above theorem means. Let's for simpliciy make the symmetry assumption $T = \sum_{i=1}^r u_i \otimes u_i \otimes u_i$, where $\|u_i\|_\infty \leq 1$. Then if you are given $\tilde{O}(n^{1.5})$ random entries of the tensor $T$, you can find some $X \in \mathbb{R}^{n \times n \times n}$, so that

$$\mathbb{E}\|X - T\|_2^2 \leq \text{poly}(r) \cdot n^3 \cdot \left(\frac{n^{1.5} \log n}{m}\right)^{\Omega(1)},$$

where the poly$(r)$ is a constant, $n^3$ is for normalization, and the final term becomes $o(1)$ as $m \gg n^{1.5} \cdot \log n$. The expectation is over the random set $\Omega \subset [n]^3$ of entries of size $|\Omega| = m$ that we see.

Now observe that $\frac{1}{r} \cdot T = \mathbb{E}_\mu x^{\otimes 3}$. What is the distribution $x \sim_\mu \{-1, 1\}^n$ that makes this true? first sample $i \sim [r]$ uniformly, and then sample $x_j \sim bias(u_i(j))$. Note that

$$T_{jkl} = \frac{1}{r} \cdot \sum_i \mathbb{E}[x^{\otimes 3}|i]_{jkl} = \frac{1}{r} \sum_i u_i(j) u_i(k) u_i(l),$$

which holds for $j, k, l$ distinct. There's a slight issue with $j, k, l$ not distinct which will be fixed in the notes.

If we could find a distribution $\mu$ so that $\mathbb{E}_\mu x^{\otimes 3} \propto T$, then we'd be good in terms of getting a tensor decomposition. We can't do this, but can search for a pseudodistribution. So the alg is as follows:

1. Find a degree $O(1)$ $\widetilde{\mathbb{E}}$ so that $\frac{1}{r} T_{ijk} = \widetilde{\mathbb{E}} x_i x_j x_k$ for all $(i, j, k) \in \Omega$.
2. Output $r \cdot \widetilde{\mathbb{E}} x^{\otimes 3}$.

We want to show that if a pseudoexpectation agrees with $T$ in the entries in $\Omega$, then it has to agree with all entries of the tensor (up to some approximation error). In particular, we want to upper bound

$$\mathbb{E}_\Omega \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk} \left(T_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k\right)^2 - \mathbb{E}_{ijk \sim \Omega} (T_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2.$$

We will find $\widetilde{\mathbb{E}}$ so that the second term is 0. We want to show that for all pseudoexpectations, the second term is bounded above by the first term plus some small error. Note that this begins to look like uniform convergence from learning theory.

Let us write $T' = T/r$. To bound the above, we introduce a ghost sample $\Omega'$:

$$\mathbb{E}_\Omega \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{\Omega'} \mathbb{E}_{ijk \sim \Omega'} (T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2 - \mathbb{E}_{ijk \sim \Omega} (T_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2.$$

By Jensen, the above can be upper bounded by:

$$\mathbb{E}_{\Omega, \Omega'} \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk \sim \Omega'} (T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2 - \mathbb{E}_{ijk \sim \Omega} (T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2.$$

Now we do symmetrization, which allows us to rewrite the above as

$$\mathbb{E}_{\Omega, \Omega', \sigma_{ijk}} \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk \sim \Omega'} \sigma_{ijk} (T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2 - \mathbb{E}_{ijk \sim \Omega} \sigma_{ijk} (T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2,$$

where $\sigma_{ijk} \in \{\pm 1\}$ are uniformly random signs. By triangle ienquality, the above can be bounded above by

$$2\mathbb{E}_{\Omega,\sigma_{ijk}} \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk\sim\Omega} \sigma_{ijk}(T'_{ijk} - \widetilde{\mathbb{E}} x_i x_j x_k)^2.$$

Now let's expand the square:

$$\mathbb{E}_{\Omega,\sigma} \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk\sim\Omega}\ \sigma_{ijk}(\widetilde{\mathbb{E}} x_i x_j x_k)^2 + \cdots.$$

Now we do a trip that we've used before: we define a new pseudoexpectation over "independent copies", variables $x'$, that allow us to rewrite the square:

$$\mathbb{E}_{\Omega,\sigma} \sup_{\widetilde{\mathbb{E}}} \mathbb{E}_{ijk\sim\Omega}\ \sigma_{ijk}\widetilde{\mathbb{E}} x_i x'_i x_j x'_j x_k x'_k + \cdots.$$

Now let us define variables $y_i = x_i x'_i$, and for any $\widetilde{\mathbb{E}}$ on $x, x'$, it is also a pseudoexpectation on the $y_i$ (and is in particular a PE on the cube). So we have:

$$\mathbb{E}_{\Omega,\sigma} \sup_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}} \frac{1}{\Omega} \sum_{i,j,k\in\Omega} \sigma_{ijk} y_i y_j y_k + \cdots.$$

This is exactly a random polynomial of exactly the form we considered above, and you can get an upper bound on the above as we showed.

# 4 October 7, 2022

Today we talk about SoS beyond the hypercube, and an application to robust mean estimation. In particular, let's consider any subset $\Omega \subset \mathbb{R}^n$: we want subsets that have some sort of short description. In particular, we focus on sobusets that are defined by polynomial inequalities (called *semialgebraic*).

Consider some $m \leq \text{poly}(n)$ and a set of inequalities

$$A = \{f_1(x) \geq 0, \ldots, f_m(x) \geq 0\}.$$

Then define $\Omega = \{x \in \mathbb{R}^n \ : \ f_i(x) \geq 0\ \forall f_i \in A\}$. Note that to encode an equality $g_i(x) = 0$, we can encode it as the two equalities $g_i \geq 0, g_i \leq 0$.

Two kinds of questions we care about:

- Is the set $\Omega$ nonempty?
- Given some other function $g : \mathbb{R}^n \to \mathbb{R}$, what is $\max_{x\in\Omega} g(x)$?

This is a direct generalization of our discussion pertaining to the hypercube: there we have $A = \{x_i^2 - x_i = 0\}_{i=1}^n$. For the above questions, each has a trivial form of witness in one direction but not in the other direction (e.g., lower bounds on $\max g$ are trivial to witness, upper bound sare harder).

## 4.1 Pseudoexpectations on general sets

**Definition 4.1.** Given a set of inequalities $A$, we say that $A \vdash_d g \geq 0$ if there exists a set of inequalities $p_S$, $S \sqsubset [m]$ (i.e., $S$ ranges over multi-subsets of $[m]$), so that each $p_S$ is a SoS polynomial, and for all $x \in \mathbb{R}^n$,

$$g(x) = \sum_{S \sqsubset [m]} p_S(x) \cdot \prod_{i \in S} f_i(x). \tag{8}$$

Furthermore, each term $p_S(x) \cdot \prod_{i \in S} f_i(x)$ must be a degree $\leq d$ polynomial.

Clearly, if $A \vdash_d g \geq 0$, then we must have $g(x) \geq 0$ for all $x \in \Omega$: this is because $f_i(x) \geq 0$ for all $x \in \Omega$. Typically we have $m$ at most polynomial in $n$, and $d$ at most constant or growing very slowly in $n$. So, the only subsets $S \sqsubset [m]$ which show up will be very small ones (e.g., constant size, if $d$ is constant). Thus the sum is over a polynomial-size set.

Given the set $\{p_S\}$, how do we verify (8)? We can verify in polynomial time that the equality holds by looking at the coefficients of both sides of the equation. The above definition can be seen to coincide with our definition over the hypercube: we have to use that anything which is identically 0 over the hypercube lies in the ideal generated by the polynomials $x_i^2 - x_i$.

Some basic questions are:

- Do SoS proofs exist?
- Do small-degree proofs exist?
- Do small size proofs exist?

Before answering the above question, we need one more definition:

**Definition 4.2.** A *refutation* is the constant polynomial $g(x) \equiv -1$, where $A$ satisfies that $A \vdash_d g \geq 0$. This tells you that $\Omega$ must be empty: the axioms $A$ cannot simultaneously be satisfied by any $x$.

## 4.2 Basics for SoS proofs over general domains

**Theorem 4.1** (Positivstellensatz)**.** *For all $A$, either $\Omega$ is nonempty, or there exists some $d$ so that $A \vdash_d -1 \geq 0$.*

In particular, for all semialgebraic sets, either it is nonempty or there is some degree (perhaps super large as a function of axioms) so that there is a SoS proof that the set is empty. The above is phrased in the retutation setting: if we're interested for the optimization setting, we can always add some constraint on the function $g$ we're optimizing to the set of axioms. In particular, we want to sknow if $g(x) \leq C$ for all $x \in \Omega$: then we can consider the set $A' = A \cup \{g(x) - c \geq 0\}$, and try to refute $A'$.

In terms of whether small size proofs exist: for the hypercube, we knew that low-degree proofs can be expressed in a small (polynomial) number of bits (HW problem). This turns out to no longer be true for the setting of general inequalities $A$. In particular, there are examples where the degree of the axioms $f_i$ is constant, the degree of the proof (called $d$) is constant, yet the bit complexity of the proofs is too big. The good news is such examples are pathological and won't

really hurt us (so these types of examples aren't really important for us, at least for known TCS applications).

Over the hypercube, we were able to give generic statements about polynomial-time algorithms: the reason for this is that we knew bounds on the coefficientts. This becomes messier for general $A$.

The bad news for us is that the degree of any proof might be very large. This is what we spend lots of time on.

### 4.3 Composability

Now we develop some tools for coming up with SoS proofs of inequalities over general domains: given a set $A$ of axioms:

- If $A \left|\frac{}{d}\right. f \geq 0$ and $A \left|\frac{}{d}\right. g \geq 0$, then $A \left|\frac{}{d}\right. f + g \geq 0$. Why? Just add the two proofs.
- If $A \left|\frac{}{d}\right. f, g \geq 0$, then $A \left|\frac{}{2d}\right. fg \geq 0$. Why? Just multiply the two proofs (since a product of SoS polynomials is a SoS polynomial). We write this one out in more detail: If $A = \{f_1 \geq 0, \ldots, f_m \geq 0\}$, $f = \sum_S p_S f_S$, $g = \sum_S p_S g_S$, then

$$fg = \sum_{S,S'} p_S q_{S'} f_S f_{S'}.$$

  Now $p_S q_{S'}$ is a SoS, and $f_S f_{S'}$ is a subset of axioms, corresponding to $SsqcupS'$.
- Consider sets $A, B, C$ of axioms. If $A \left|\frac{}{d}\right. B, B \left|\frac{}{d'}\right. C$, then $A \left|\frac{}{dd'}\right. C$.
- Write $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_N)$. If $A(x) \left|\frac{x}{d}\right. g(x) \geq 0$, and each $x_i$ is a result of evaluating some polynomial in the $y$, i.e., $x_i(y_1, \ldots, y_N)$, then we have $A(x(y)) \left|\frac{y}{d \cdot \max_i \deg(x_i)}\right. g(x(y)) \geq 0$. This is trivial to prove (by substitution). Here to prove some statement about the $y$'s, we construct some quantity $x_i$'s which are a function of the $y$'s, and then reason about the $x_i$'s.

There are many other axioms too; we will use such things very freely.

**Proposition 4.2.** *If $f(x) : \mathbb{R} \to \mathbb{R}$, and if $f(x) \geq 0$ for all $x$, then $f$ is a SoS.*

The above can be proved easily by the fundamental theorem of algebra. Proposition 4.2 can be used in the following way: if we want to prove that $F(x_1, \ldots, x_n)$ is non-negative, we can sometimes reduce it to non-negativity for a polynomial $f(g(x_1, \ldots, x_n))$, where $f$ is a single-variable function. Then, if $f$ is actually non-negative, we can use it can be written as a SoS (by the above proposition).

**Proposition 4.3.** *For all $f$, there exists $M \in \mathbb{R}$ so that $\{\|x\|^2 \leq 1\} \left|\frac{}{\deg f}\right. \leq M$.*

The above proposition isn't that useful for constructing SoS proofs; but it will be useful to prove duality theorems in this general setting.

## 4.4 Pseudoexpectations

We will now redefine pseudoexpectations to hold over more general domains.

**Definition 4.3.** A pseudoexpectation is a map $\widetilde{\mathbb{E}} : \mathbb{R}[x_1, \ldots, x_n]_{\leq d} \to \mathbb{R}$ which is:

- Linear.
- Positive semi-definite, which means that for all $p$ satisfying $\deg p \leq d/2$, $\widetilde{\mathbb{E}} p^2 \geq 0$.
- Normalized: $\widetilde{\mathbb{E}} 1 = 1$.

A pseudoexpectation over the hypercube also respected the multilinearity operation. Note that:

- A degree-$d$ PE can be represented with $n^d$ numbers.
- The non-negativity constraint (which sometimes is written as $\widetilde{\mathbb{E}} \succeq 0$) is equivalent to

$$\widetilde{\mathbb{E}}(x^{\otimes \leq d/2})(x^{\otimes \leq d/2})^\top \succeq 0,$$

  where $x^{\otimes \leq d}$ is the vector which contains all monomials of degree up to $d/2$. (Why is this the case: given a SoS polynomial write out the polynomial as a sum of squares, write the PE evaluated at each square of the form $v^\top M v$, where $M$ is the above matrix and $v$ is the vector of coefficients of the thing being squared. The reverse direction is similar.)

**Definition 4.4** (Satisfying a system of inequalities)**.** We say that $\widetilde{\mathbb{E}}$ satisfies $A$ if for all $S \sqsubset A$, for all $p$, if $\deg(p^2 \cdot f_S) \leq d$, then $\widetilde{\mathbb{E}}[p^2 f_S] \geq 0$. We will write $\widetilde{\mathbb{E}} \vDash A$.

This generalizes the condition we had for the hypercube which said that $\widetilde{\mathbb{E}}[(x_i^2 - x_i)p] = 0$.

Some intuition for the above: if we had a distribution $\mu$ over $\Omega$, then we have $\mathbb{E}_\mu[p^2 f_S] \geq 0$. Note that a much wekaer statement is that $\widetilde{\mathbb{E}}[f_i(x)] \geq 0$ for all $f_i \in A$.

Here's an example: consider the distribution $\mathrm{Unif}(\{-1, 1\})$. Here we have $\mathbb{E}[x] = 0$, yet it does not satisfy the equality $x = 0$; in particular, $\mathbb{E}[x^2] = 1$.

## 4.5 Duality

**Theorem 4.4.** *Let $A$ containt $\|x\|^2 \leq M$. For all even $d$, for all $f \in \mathbb{R}[x_1, \ldots, x_n]_{\leq d}$, exactly one of the following occurs:*

1. *for all $\epsilon > 0$, $A \vdash_d f \geq -\epsilon$.*
2. *There exists $\widetilde{\mathbb{E}}$ of degree $d$ so that $\widetilde{\mathbb{E}} \vDash A$ and $\widetilde{\mathbb{E}} f \leq 0$.*

Next we discuss an algorithmic version of duality.

**Theorem 4.5.** *There exists an $n^{O(d)}$ time algorithm given $A$ which is satisfiable (i.e., $\Omega$ is nonempty) and for which the number $m$ of constraints is $m \leq n^{O(d)}$ and for which the bit complexity is similarly bounded by $n^{O(d)}$, and outputs a pseudoexpectation $\widetilde{\mathbb{E}}$, so that $\widetilde{\mathbb{E}} \vDash^{2^{-n}} A$.*

Above, $\widetilde{\mathbb{E}} \vDash^{\epsilon} A$ means that for all $p$ so that $\|p\| = 1$ (here $\|\cdot\|$ refers to the 2-norm when coefficients are written as a vector), $\widetilde{\mathbb{E}}[p \cdot f_S] \geq -\epsilon$.

The ideal statement is that $\widetilde{\mathbb{E}}$ satisfies $A$, but we don't quite get it. Note that if $A \vdash_d \|x\|^2 \leq M$, then we can improve the above theorem statement to get that $TE \vDash A$. But the approximation error won't really matter: we typically solve for a PE that satisfies some system of inequalities (approximately), and then round that PE back to get some $x^\star$. When we analyze the rounding algorithm, we typically have some sequence of inequalities $\widetilde{\mathbb{E}}p_1 \geq \widetilde{\mathbb{E}}p_2 \geq \cdots$, derived using the axioms $A$. But these inequalities are true if we only have $\widetilde{\mathbb{E}} \vDash^{2^{-n}} A$, as long as the polynomials $\|p_i\| \ll 2^n$. This will never amplify the additive errors in any way that hurst us. So, all we have to ensure is that when we analyze the algorithms, we don't analyze them using enormous numbers.

### 4.6 Proofs to algorithms

Typically we have some data-generating process that depends on some parameter $\theta$, which produces some samples $X_1, \ldots, X_n$. We want to find some algorithm $\hat{\theta}$ so that $\|\hat{\theta}(X_1, \ldots, X_n) - \theta\|$ is small. Often to prove identifiability, we want to show that going backwards (from "data to theta") is possible. Unfortunately this map may be computationally intractable/very complex.

The idea of SoS-proof based algorithms: if we can prove in SoS that identifiability holds, then we get a poly-time algorithm $\hat{\theta}$ mapping from the data $X_1, \ldots, X_n$ back to the correct parameter $\theta$.

Today we instantiate this framework with a very basic example of a robust statistics problem. Often in statistics, we make some assumption that a population has some properties (e.g., is Gaussian) and then data is drawn from this distribution. In *robust statistics*, we relax the assumption on the data distribution:

**Definition 4.5** (Strong contamination)**.** Given a distribution $D$, we say that samples $X_1, \ldots, X_n \sim_\epsilon D$ are drawn in the *strong contamination model* if:

1. $X_1^\star, \ldots, X_n^\star \sim D$ iid.
2. The adversary looks at the samples $X_{1:n}^\star$ and then modifies any $\epsilon n$ of them, and hands the resulting $X_1, \ldots, X_n$ to the learning algorithm.

The strong contamination model certainly can model situations where the data is actually generated adversarially. But it can also model situations where there are small ($\epsilon$-fraction) fraction of sub-populations that are not modeled by the family of distributions containing $D$, namely where you have model-misspecification up to $\epsilon$ error.

In low-dimensional settings, robust statistics heavily studied in 1960s-1970s. But, those algorithms don't scale to high-dimensional settings. The most basic high-dimensional statistics problem is *robust mean estimation*.

**Definition 4.6** (Robust mean estimation)**.** Suppose that $D$ is a distribution on $\mathbb{R}^d$, and assume that $\mathbb{E}_{X \sim D}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top \preceq I$. We receive $\epsilon$-contaminated samples $X_1, \ldots, X_n \sim_\epsilon D$, and the goal is to find $\hat{\mu}$ so that $\|\hat{\mu}(X_{1:n}) - \mathbb{E}_{X \sim D} X\|$ is small.

In the standard setting, taking the empirical average is optimal; but in the adversarial setting, the adversary can kill this algorithm by taking samples to infinity. Note that, in the presence of corruptions, we need some assumption on the data generating distribution since it could be that the distribution the adversary generates is the true distribution.

**Theorem 4.6.** *With $n \gg d/\epsilon$ samples, we can find $\hat{\mu}$ so that $\|\hat{\mu} - \mu\| \le O(\sqrt{\epsilon})$ with high probability in $n^{O(1)}$ time.*

Note that, for small $\epsilon$, we can actually get a refined guarantee os $\|\hat{\mu} - \mu\| \le O\left(\sqrt{d/n} + \sqrt{\epsilon}\right)$.

Note that the $\sqrt{\epsilon}$ is tight since the adversary can confuse you a bit with $\epsilon$-contamination. The above theorem was originally proved without SoS. But soon thereafter it was shown how to achieve the above using SoS, which leads to other algorithms in robust statistics.

Note that for an adversary who moves some small fraction of samples to infinity can be taken care of *naive outlier removal*: in particular, discard any samples $X_i$ for which $\|X_i - X_j\| \gg \sqrt{d}$ for many $X_j$ (here $\sqrt{d}$ is chosen as an upper bound on the square root of the trace of the covariance). Then we can just average the remaining samples. Note that with this strategy, the adversary can still mess you up: if they move all samples to the same point which is distancee $\sqrt{d}$ from the true mean. Then we have, letting $[n] = G \cup B$ consist of the good/bad decomposition of samples,

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(\sum_{i\in G} X_i + \sum_{i \in B} X_i\right) \approx_{\epsilon\sqrt{d}} \mu.$$

In particular, the empirical mean is off by $\epsilon\sqrt{d}$, so we're off by a $\sqrt{d}$ factor.

## 4.7 Proofs of identifiability for robust mean estimation

Let's begin by proving identifiability in a simple way that can be encoded as a SoS proof.

**Lemma 4.7.** *Suppose $X, X'$ are random variables on $\mathbb{R}$. Suppose that $\mathrm{TV}(X, X') \le \epsilon$. Moreover suppose that $\mathbb{V}[X], \mathbb{V}[X'] \le 1$. Then $|\mathbb{E}X - \mathbb{E}X'| \le O(\sqrt{\epsilon})$.*

Roughly speaking: the idea is that we are constrained to only move $\epsilon$ mass from $X$ to $X'$, and we can't move it too far by the variance bound.

*Proof.* Let us consider a coupling $(X, X')$ so that $X = X'$ with probability at least $1 - \epsilon$. Then

$$|\mathbb{E}(X - X')| = |\mathbb{E}[\mathbb{1}\{X \neq X'\}(X - X')]| \le \sqrt{\mathbb{E}\mathbb{1}\{X \neq X'\}} \cdot \sqrt{\mathbb{E}(X - X')^2} \le \sqrt{\epsilon} \cdot \sqrt{\mathbb{E}(X - X')^2},$$

where we have used Cauchy-Schwarz and the coupling assumption.

Now we write

$$\begin{aligned}\mathbb{E}(X - X')^2 =& \mathbb{E}(X - \mathbb{E}X - (X' - \mathbb{E}X') + \mathbb{E}X - \mathbb{E}X')^2 \le O(1) \cdot \left(\mathbb{E}(X - \mathbb{E}X)^2 + \mathbb{E}(X' - \mathbb{E}X')^2 + (\mathbb{E}X - \mathbb{E}X')^2\right) \\ \le& O(1) \cdot (1 + 1 + (\mathbb{E}X - \mathbb{E}X')^2).\end{aligned}$$

Putting it all together, we get

$$|\mathbb{E}X - \mathbb{E}X'| \le O(\sqrt{\epsilon}) \cdot (1 + |\mathbb{E}X - \mathbb{E}X'|),$$

and thus $|\mathbb{E}X - \mathbb{E}X'| \le O(\sqrt{\epsilon}/(1 - \sqrt{\epsilon})) \le O(\sqrt{\epsilon})$. □

**High-dimensional variant.** Now we generalize the previous lemma to high dimensions.

**Lemma 4.8.** *Suppose $X, X'$ are random variables on $\mathbb{R}^d$. Suppose that $\mathrm{TV}(X, X') \le \epsilon$. Moreover suppose that $\mathrm{Cov}(X), \mathrm{Cov}(X') \preceq 1$. Then $\|\mathbb{E}X - \mathbb{E}X'\| \le O(\sqrt{\epsilon})$.*

*Proof.* Take $X, X' \in \mathbb{R}^d$. Define $Y = \frac{\langle X, \mathbb{E}X - \mathbb{E}X' \rangle}{\|EX - \mathbb{E}X'\|}$ and $Y' = \frac{\langle X', \mathbb{E}X - \mathbb{E}X' \rangle}{\|\mathbb{E}X - \mathbb{E}X'\|}$. We claim that $Y, Y'$ satisfy the conditions of the 1-dimensional lemma. The data processing inequality gives the total variation distance. The variance is at most 1 in any (unit) direction by the covariance assumption, and the vector we're hitting $X, X'$ with to get $Y, Y'$ is a unit vector.

So, by the previous lemma, we get that

$$O(\sqrt{\epsilon}) \ge |\mathbb{E}Y - \mathbb{E}Y'| = \frac{\|\mathbb{E}X - \mathbb{E}X'\|^2}{\|EX - \mathbb{E}X'\|} = \|EX - \mathbb{E}X'\|.$$

□

**Getting identifiability.** In our dream world, the ground-truth samples $X_i^\star$ have the property that $\mathrm{Cov}(X_i)^\star \le 2$ and $\mathbb{E}X^\star \approx \mu$. In the dream world, we take the corrupted samples, find a large subset whose empirical covariance is also bounded: namely, find $T \subset \{X_{1:n}\}$, $|T| \ge (1-\epsilon)n$, and so that $\mathrm{Cov}(T) \preceq 2I$. Then we apply the lemma where the two distributions are (1) the empirical distribution over $T$, and (2) the true distribution of $X^\star$. The idea is that there are only $\epsilon$ points that are tampered with (so TV is at most $O(\epsilon)$), and both distributions have bounded covariance, so the lemma tells us that the distributions have means that are apart by at most $O(\sqrt{\epsilon})$: i.e., $\|\mathbb{E}_{X\sim T}[X] - \mathbb{E}_D X^\star\| \le O(\sqrt{\epsilon})$. Note that we have to remove the $\epsilon n$ fraction samples for two reasons: (1) because of the adversary, (2) since we don't make any concentration assumption, the true empirical covariance might not be bounded with high probability (this second issue goes away if we assume, e.g., sub-Gaussianity).

Formalizing the above, to establish identifiability, we use the following lemma:

**Lemma 4.9.** *If $n \gg d/\epsilon$ then there exist a subset $S \subset \{X_{1:n}^\star\}$ with $|S| \ge (1-\epsilon)n$, $\mathrm{Cov}(S) \preceq 2 \cdot I$ (where $\mathrm{Cov}(S)$ denotes the empirical covariance of samples in $S$), and $|\mathbb{E}_{X\sim S}X - \mu| \le O(\sqrt{\epsilon})$.*

To find the subset $S$ (which allows us to estimate $\mu$ to within $\sqrt{\epsilon}$), note that it suffices to find some subset $S$ of points so that $\mathrm{Cov}(S) \preceq 2 \cdot I$. then in fact using the previous lemmas it follows that $\mathbb{E}_{X\sim S}X - \mu| \le O(\sqrt{\epsilon})$. The naive thing to do is to use brute force search. A second more clever thing to do is to find a better algorithm to find $T \subset \{X_{1:n}\}$ with $\mathrm{Cov}(T) \preceq 2I$ and so that $|T| \ge (1-\epsilon)n$.

We will instead take a third option which is to directly go to estimating the mean.

## 4.8 Proofs-to-algorithms for robust mean estimation

We will encode subsets $S \subset [n]$ which have bounded empirical covariance as solutions to systems of inequalities. We won't solve this directly, but then will look at SoS.

In particular: given $X_1, \dots, X_n \sim_\epsilon D$, we define a system $A_{X_{1:n}}(w, B)$ in variables $w, B$ which is satisfied with $w$ equal to the 0-1 indicator vector of some subset with bounded covariance. We will then have an algorithm which finds $\widetilde{\mathbb{E}}$ satisfying $\widetilde{\mathbb{E}} \vDash A_{X_1,\dots,X_n}$. Instead of trying to extract a

subset encoded by $w$, we just try to find the mean: in particular, we want the empirical mean of samples in $w$, which is:

$$\widetilde{\mathbb{E}}\frac{1}{(1-\epsilon)n}\sum_{i\in[n]} w_i X_i.$$

We never actually round the vector $w$, but just output the value of the above pseudoexpectation.

Given $X_{1:n}$, we said before that we will find some subset that has small empirical covariance. To make things simpler, we actually consider a slight twist: we will find $X'_{1:n}$ so that $\mathrm{TV}(X_{1:n}, X'_{1:n}) \le \epsilon$ (namely, the two sets $X, X'$ differ by at most $\epsilon n$ points), and so that $\mathrm{Cov}(X'_{1:n}) \preceq 2I$. In particular, if $X^\star_{1:n}$ has $\mathrm{Cov}(X^\star_{1:n}) \preceq 2I$, then the lemma we showed above gives that $\|EX' - \mathbb{E}X^\star\| \le O(\sqrt{\epsilon})$ (here the TV statement holds since $X^\star, X$ are close in TV by assumption, and $X', X$ are close in TV by construction).

We now create the following system of polynomial inequalities: $A_{X_{1:n}}(w_{1:n}, X'_{1:n}, \{B_{ij}\}_{i,j\le d})$ :

- $w_i^2 = w_i$, forall $i$.
- $w_i X'_i = w_i X_i$ for all $i$: in particular, for all things we include in the subset, $X'_i = X_i$.
- $\sum_{i=1}^n w_i = (1-\epsilon n)$. This asks for the subset to be large.
- $\frac{1}{n}\sum_i (X'_i - \mathbb{E}_i X'_i)(X'_i - \mathbb{E}_i X'_i)^\top = 2\cdot I - BB^\top$. Here we have an equality of matrices; note that $B$ is a $d \times d$ matrix of variables.

The above system of polynomials describes the problem of finding a collection $X'_1, \dots, X'_n$ which is close in TV to $X_{1:n}$ and has bounded covariance.

Goal: we want to show that

$$A \vdash_{O(1)} \|\mathbb{E}_{i\sim[n]} X'_i - \mathbb{E}_{i\sim[n]} X^\star_i\|^4 \le O(\epsilon^2), \tag{9}$$

assuming that $\mathrm{Cov}(X^\star_i) \preceq 2\cdot I$. Thus, given $\widetilde{\mathbb{E}}$ satisfying $\widetilde{\mathbb{E}} \vDash A$, we can simply output $\widetilde{\mathbb{E}}\mathbb{E}_{i\sim[n]} X'_i$. By the existencee of the SoS proof above, we know that $\widetilde{\mathbb{E}}\|\mathbb{E}_{i\sim[n]} X'_i - \mathbb{E}_{i\sim[n]} X^\star_i\|^4 \le \epsilon^2$. Then we can use pseudoexpectation Cauchy-Schwarz (on HW) to get that

$$\|\widetilde{\mathbb{E}}\left[\mathbb{E}_{i\sim[n]} X'_i - \mathbb{E}_{i\sim[n]} X^\star_i\right]\|^4 \le O(\epsilon^2),$$

which is what we want to show. The statement (9) is exactly the proof of the identifiability lemma from before, explicitly proven in constant-degree SoS proofs.

*Proof.* Define variables $v(w, B, X') = \mathbb{E}_{i\sim[n]} X'_i - \mathbb{E}_{i\sim[n]} X^\star_i$, and $z_i = \mathbb{1}\{X_i = X^\star_i\}$. Here the $z_i$ are numbers, not variables. We now write

$$\mathbb{E}_{i\sim[n]} X'_i - X^\star_i = \mathbb{E}_{i\sim[n]} w_i z_i (X'_i - X^\star_i) + (1 - w_i z_i)(X'_i - X^\star_i).$$

The idea is that $w_i z_i$ is the proxy for the indicator that $X'_i = X^\star_i$. By the SoS axioms, we have that

$$\begin{aligned} A \vdash_{O(1)} \mathbb{E}_{i\sim[n]} w_i z_i (X'_i - X^\star_i) + (1 - w_i z_i)(X'_i - X^\star_i) &= \mathbb{E}_{i\sim[n]} w_i z_i (X_i - X^\star_i) + (1 - w_i z_i)(X'_i - X^\star_i) \\ &= 0 + \mathbb{E}_{i\sim[n]} (1 - w_i z_i)(X'_i - X^\star_i). \end{aligned}$$

where we have used that $w_i X_i' = w_i X_i$ by our axioms. Note that the second step follows since $z_i(X_i - X_i^\star) = 0$. (Proving equality means proving an upper and lower bound.)

Now we have to square everything. Note that by definition of $v$, we have that $\|\mathbb{E}_i X_i' - \mathbb{E}_i X_i^\star\|^4 = \langle \mathbb{E}_i X_i' - \mathbb{E}_i X_i^\star, v\rangle^2$. We now get that

$$\begin{aligned} A \vdash \langle \mathbb{E}_i X_i' - \mathbb{E}_i X_i^\star, v\rangle^2 =& \langle \mathbb{E}_i (1 - w_i z_i)(X_i' - X_i^\star), v\rangle^2 \\ =& \left(\mathbb{E}_{i\sim[n]} (1 - w_i z_i)\langle X_i' - X_i^\star, v\rangle\right)^2 \\ \le& \mathbb{E}_{i\sim[n]}[(1 - w_i z_i)^2] \cdot \mathbb{E}_{i\sim[n]}[\langle X_i' - X_i^\star, v\rangle^2], \end{aligned}$$

where the inequality is by SoS Cauchy-Schwarz.

Now we have to deal with the two terms resultin from Cauchy-Schwarz: we can prove in SoS that:

$$\{(w_i^2 = w_i)\} \vdash_{O(1)} (1 - w_i z_i) = ((1 - w_i)^2 + w_i^2(1 - z_i)^2) \le (1 - w_i) + (1 - z_i).$$

Then, using the proof composition rules to square both sides, we get that

$$A \vdash_{O(1)} \mathbb{E}_i[(1 - w_i z_i)^2] \le 2 \cdot (\mathbb{E}_i[(1 - w_i) + (1 - z_i)]) \cdot \mathbb{E}_i \langle X_i' - X_i^\star, v\rangle^2.$$

by the coupling axiom, we have $A \vdash \mathbb{E}_i(1 - w_i) \le O\epsilon)$, and since the $z_i$ are just numbers, we have that the above isupper bounded (in SoS) by $O(\epsilon) \cdot \mathbb{E}_i \langle X_i' - X_i^\star, v\rangle^2$.

Now we deal with the second term:

$$A \vdash \mathbb{E}\langle X_i' - X_i^\star, v\rangle^2 \le O(1) \cdot \left(\mathbb{E}_i \langle X_i' - \mathbb{E}X_i', v\rangle^2 + \mathbb{E}_i \langle X_i^\star - \mathbb{E}X_i^\star, v\rangle^2 + \langle v, v\rangle^2\right),$$

where we have used Young's inequality (for SoS) above. Now, we have

$$\mathbb{E}_{i\sim[n]} \langle X_i' - \mathbb{E}X_i', v\rangle^2 = v^\top \frac{1}{n} \sum_i (X_i' - \mathbb{E}X_i')(X_i' - \mathbb{E}X_i')^\top v = \|v\|^2 - vBB^\top v^\top,$$

where the final step follows from the SoS axioms. But in SoS we have $A \vdash -v^\top BB^\top v \le 0$, and so the above is bounded above in SoS by $O(1) \cdot (\|v\|^2 + \|v\|^4)$. In particular, we have $A \vdash \|v\|^4 \le O(\epsilon) \cdot (\|v\|^2 + \|v\|^4)$, and rearranging, we get

$$A \vdash \|v\|^4 \le O(\epsilon) \cdot \|v\|^2 \le \frac{1}{2}\|v\|^4 + O(\epsilon^2),$$

where the final inequality follows again by Young's inequality (for SoS). Then rearranging again, we get $A \vdash \|v\|^4 \le O(\epsilon^2)$. □

# 5 October 14, 2022

Today: clustering. General goal is to find some good partiion of $[n]$ into $k$ parts. For instance, we have $X_1, \dots, X_n \in \mathbb{R}^d$, and want to cluster them in some way that "respects geometry". Another example is given a graph and want to cluster into $k$ groups $S_1 \cup \dots \cup S_k = [n]$ so as to minimize $E(S_i, S_j)$.

Today: we cluster points when given that a "good" clustering exists.

## 5.1 Identifiable clustering

A clustering probelm is specified by $\theta = \{((S_1, \dots, X_k), X)\}$, where we receive $X$ as input, and the goal is to find the partition $S_1, \dots, S_k$. Sometimes we can't recover the partition exactly but will consider settings where we can approximately identify the clustering $S_1, \dots, S_k$.

To ensure identifiability, we assume:

**Assumption 5.1** (Identifiability)**.** *First, we assume that $|S_1| = \cdots = |S_k| = n/k$. Further, there exists a mapping $\mathcal{C}(X)$, so that for all $(S_1, \dots, S_k, X)$, letting $\mathcal{C}(X) = (T_1, \dots, T_k)$ gives that $|T_i \cap S_i| \geq (1 - \delta) \cdot \frac{n}{k}$ for all $i$.*

Here we have defined an abstract clustering setting, so it can be specialized to any one specific model.

**Example: Gaussian mixture models.** Given distributions $D_1, \dots, D_k$ on $\mathbb{R}^d$, their *uniform mixture* is $\frac{1}{k}\sum_{i=1}^k D_i$: namely, first choose $i \sim [k]$ uniformly and then output a sample from $D_i$. If $D_i$ are Gaussian, then this is said to be a Gaussian mixture model.

Later in the lecture, we will consider the following problem: given $X_1, \dots, X_n \sim \frac{1}{k}\sum_i D_i$. Here the $S_i$ is the set of $X_i$'s drawn from $D_i$. The goal is to find $S_1, \dots, S_k$ given $X_{1:n}$. This problem only makes sense only when this is information-theoretically possible, so will need some separation assumption on the $D_i$.

## 5.2 Identifiable clustering via SoS

How can we prove identifiability via SoS? This will ultimately lead us to algorithms, via similar ideas to last time when we discussed "proofs to algorithms". Suppose that for all $X$, letting $S_1, \dots, S_k$ be the ground-truth partition for $X$, there exists a system of polynomials $P_X$ of degree $d$, in variables $w_1, \dots, w_n, z$, which identifies the clusters in the following sense: for all $a \neq b \in [k]$,

$$\left\{ w_i^2 = w_i, \quad \sum_{i=1}^n w_i = k \right\} \cup P_X \vdash_d \sum_{i \in S_a,\ j \in S_b} w_i w_j \leq \delta \cdot \left(\frac{n}{k}\right)^2 \tag{10}$$

where $P_X$ is some problem-specific set of axioms that defines "being clustered". We also want to ensure that for all clusters $a$, $1_{S_a}$ solves $P_X$. Here interpret $w_i$ as being the indicator vector for a subset. This inequality is telling us that $w_i$ cannot have too much mass split between multiple clusters: e.g., if it is evenly split between 2 clusters, then $\sum_{i \in S_a,\ j \in S_b} w_i w_j = 1/4 \cdot (n/k)^2$. The idea is that most of the mass of the $w_i$ is on one cluster, and then $\mathrm{poly}(\delta)$ is split between other clusters. So, taken together over all $a, b$, the proofs say that the $w$ vector lies mostly within one cluster.

(Recall that last time we built a polynomial system that identifies the mean of a distribution.)

We will show the below result (roughly):

**Theorem 5.2** (Informal)**.** *If the above holds, then we have: in $\mathrm{poly}(n^d)$ time, given a pseudoexpectation $\widetilde{\mathbb{E}} \vDash \{w_i^2 = w_i, \sum w_i = n/k\} \cup P_X$ (with minimal $\|\widetilde{\mathbb{E}} w\|_2^2$), then we can find a partition $T_1, \dots, T_k$ so that $|S_a \cap T_a| \geq (1 - \delta k^{O(1)}) \cdot \frac{n}{k}$.*

**Example of the sytem $P_X$.** Let's consider $X = X_1, \dots, X_n \in \mathbb{R}$. Suppose each $S_i$ consists of a disjoint interval of length 1: in particular, suppose every two intervals are separated by at least 10.

What is the system of polynomials $P_X(w_1, \dots, w_n)$ here? It is just looking at differences between $X_i, X_j$:

$$P_X(w_1, \dots, w_n) = \{w_i w_j (X_i - X_j)^2 \le 1\}..$$

(In words: whenever $i, j$ are both in the cluster, the distance between $X_i, X_j$ is at most 1.)

What can we prove using $P_X(w)$? For $a \neq b$,

$$P_X(w) \vdash \sum_{i \in S_a, j \in S_b} w_i w_j \le \frac{1}{100} \sum_{i \in S_a, j \in S_b} w_i w_j (X_i - X_j)^2,$$

since in SoS

$$w_i w_j = w_i^2 w_j^2, \quad w_i^2 w_j^2 \cdot \left( \frac{(X_i - X_j)^2}{100} - 1 \right) \ge 0,$$

since the latter expression is a square (where we use that for $i, j$ in different clusters, $(X_i - X_j)^2 \ge 100$).

But in SoS we have

$$P_X(w) \vdash (n/k)^2 \cdot \sum_{i \in S_a, j \in S_b} w_i w_i (X_i - X_j)^2 \le \frac{1}{100} (n/k)^2.$$

We can get a better guarantee by doing repeated squaring, peeling off one copy, repeatedly. This will raise the power of $(X_i - X_j)$, but those are numbers (not variables), so still gets it to work i nconstant degree.

### 5.3 Proving Theorem 5.2 using randomized rounding

Now we extract the clusters from a pseudoexpectation satisfying the above system, using randomized rounding. There are actually lots of different things that work here. This is actually quite standard, so all the work is constructing the polynomials $P_X$ for the particular notino of clustering you want to consider.

What does $\widetilde{\mathbb{E}}$ look like? Can we use the first moments? If we get really lucky and $\widetilde{\mathbb{E}}$ correpsonds to the point mass on the indicator vector of cluster 1, $\widetilde{\mathbb{E}} q(w) = q(1_{S_1})$, then we could output $\widetilde{\mathbb{E}} w_1, \dots, \widetilde{\mathbb{E}} w_n$, and we'd be good. We could of course also hve $\widetilde{\mathbb{E}} q(w) = q(1_{S_2})$. But consider the "bad" pseduexpectation $\widetilde{\mathbb{E}}'$ with $TE[q(w)] = \frac{1}{2} \cdot (q(1_{S_1}) + q(1_{S_2}))$. This is possible since the class of pseudoexpectations is convex. In particular, generalizing, if the pseudoexpectation $\widetilde{\mathbb{E}}$ is the uniform distribution over cluster indicators, then $\widetilde{\mathbb{E}}[w_i] = 1/k$ for all $i$, which is useless. Thus, the first moment is not useful to you!

**Second moments.** Let's consider second moments of pseudoexpectations. Consider $\widetilde{\mathbb{E}} w_i w_j$ when $i, j$ are either from the same or different clusters. By (10), adding up over all $k^2$ clusters, we have

$$P_X \vdash \sum_{i,j \text{ in different clusters}} \widetilde{\mathbb{E}} w_i w_j \le \delta n^2.$$

What about points in the same cluster? If the pseudoexpectation is of degree 2, then we have:

$$\vdash \widetilde{\mathbb{E}} \left( \sum_{i=1}^{n} w_i \right)^2 = \widetilde{\mathbb{E}} \sum_{i,j} w_i w_j = (n/k)^2.$$

Subtracting the two equations, we have

$$\widetilde{\mathbb{E}} \sum_{i,j \text{ in same cluster}} w_i w_j \geq n^2 \cdot \left( \frac{1}{k^2} - \delta \right).$$

### 5.4 A concrete rounding scheme.

Consider the following rounding scheme:

1. Pick $i \sim [n]$, and compute $\{\widetilde{\mathbb{E}}[w_j | w_i = 1]\}_{j=1}^n$. *(Ideally, once we condition on $w_i = 1$, the pseudoexpectation should be 1 on the coordinates which are in the same cluster as point 1.)*
2. Let $T \subset [n]$ be given by including $j$ with probability $\widetilde{\mathbb{E}}[w_j | w_i = 1]$.
3. Remove $T$, and recurse $k$ times. (In particular, condition on different variables being 1 the next time, and so on.)

Note that one way this procedure could go wrong if you're in the situation (which we thought would be good) that the $\widetilde{\mathbb{E}}$ is the point mass on the indicator vector of 1 cluster. In particular, in order to condition on $w_i = 1$, we neet $\widetilde{\mathbb{E}}[w_i] > 0$, since $\widetilde{\mathbb{E}}[p(w) | w_i = 1] = \frac{\widetilde{\mathbb{E}}[w_i \cdot p(w)]}{\widetilde{\mathbb{E}} w_i}$. This is the reason we minimize $\|\widetilde{\mathbb{E}} w\|_2^2$. It turns out that minimality of $\|\widetilde{\mathbb{E}} w\|_2^2$ forces $\widetilde{\mathbb{E}} w_i = 1/k$ for all $i$.

So, the "ideal situation" is that $\widetilde{\mathbb{E}}$ corresponds to a uniform distribution over all $k$ clusters. Then the matrix of $\widetilde{\mathbb{E}}[w_i w_j]$, for $i, j \in [n]$, corresponds to a block diagonal matrix, with block diagonal matrices all 1's and off-diagonal matrices all 0's.

**Analyzing the above rounding scheme.** Suppose that the random $i \sim [n]$ that we picked satisfies $i \in S_a$. Now consider the expected value of the number of elements of $T$ which are not in $S_a$: conditioned on lying in $S_a$, each element of it occurs with equal probability $k/n$:

$$\widetilde{\mathbb{E}} \sum_{b \neq a} |T \cap S_b| = \frac{k}{n} \sum_{i \in S_a} \sum_{j \notin S_a} \frac{\widetilde{\mathbb{E}}[w_j w_i]}{\widetilde{\mathbb{E}} w_i}.$$

**Lemma 5.3.** *Minimality of $\|\widetilde{\mathbb{E}} w\|_2^2$ implies that $\widetilde{\mathbb{E}} w_i = 1/k$ for all $i$.*

*Proof.* Note that the uniform distribution over clusters must be a minimizer by the constraints $\sum_i w_i = n/k$, which implies that $\widetilde{\mathbb{E}} \sum_i w_i = n/k$.

(Alternatively, we can solve this algorithmically by constraining $\widetilde{\mathbb{E}} w_i = 1/k$ for all $i$.) □

Using the above lemma, it follows that

$$\mathbb{E} \sum_{b \neq a} |T \cap S_b| = \frac{k^2}{n} \sum_{i \in S_a} \sum_{j \notin S_a} \widetilde{\mathbb{E}} w_i w_j \leq \frac{k^2}{n} \cdot k \cdot \delta (n/k)^2 \leq kn\delta,$$

where the second-to-last inequality follows by each of the conclusions $\sum_{i\in S_a, j\in S_b} w_i w_j \le \delta\cdot(n/k)^2$ from (10).

We also want to make sure that $T$ is nonempty: $\widetilde{\mathbb{E}}|T| = \sum_j \widetilde{\mathbb{E}}[w_j | w_i = 1] = n/k$, where the inequality here follows from the definition of conditional pseudoexpectation: the LHS is $\sum_j \frac{\widetilde{\mathbb{E}}[w_j w_i]}{\widetilde{\mathbb{E}}[w_i]}$, which is $n/k$. Thus, $\mathbb{E}|T\cap S_a| \ge n/k\cdot(1-k^2\delta)$.

Now we have to put the above together to show that we get all the clusters. One way to do this is to repeatedly choose $i$ uniformly at random from $[n]$ and apply coupon collector. Alternatively, we argue as follows, where we use that we remove $T$ in the algorithm:

**Claim 5.4.** *It holds that* $\mathbb{P}$ *(select all clusters)* $\ge 1-\delta k^{O(1)}$.

*Proof.* At round $t$,

$$\mathbb{E}\,[\text{number of elements from clusters } S_a \text{ chosen already remaining }] \le k\cdot\left(k^2\delta\cdot n/k\right) \le \delta n k^3.$$

Thus, the probability we select a new cluster in round $t$ is at least $\frac{n-\delta n k^3}{n} = 1-\delta k^3$. Thus, the probability that all rounds get a new cluster is at least $1-\delta k^4$. □

Letting $T_a$ be the cluster the algorithm selected when it selected cluster $a$, it follows that if all rounds are "good" in the sense that each gets a new cluster,

$$\mathbb{E}\sum_{b\neq a}|T_a\cap S_b| \le k^2 n\delta,$$

so $|T_a\cap S_b| \le \delta n k^{O(1)}$ for all $a\in[k]$ with probability $1-\delta k^{O(1)}$.

## 5.5 Finding SoS proofs of identifability

Now, for the particular case of GMMs, we will find a system of polynomials $P_X$ and prove that (10) holds. Once we can do this, then we can (in poly$(n^d)$ time) find a pseudoexpectation satisfying the necessary constraints and minimizing $\|\widetilde{\mathbb{E}}w\|_2^2$ (which we have already seen how to do), and then we can apply the randomized rounding procedure from the previous section to find the clusters.

We consider the following case for GMMs: suppose $D_1,\dots,D_k$ on $\mathbb{R}^d$, $D_i = \mathcal{N}(\mu_i,\Sigma_i)$, where $\Sigma_i \preceq I$ (i.e., the clusters themselves are not too far spread out).

We suppose that we have samples $X_1,\dots,X_n \sim \frac{1}{k}\sum_{i=1}^k D_i$, and we assume that we get exactly $n/k$ samples (consisting of the cluster $S_i$) from each $D_i$ (by concentration you get approximately this many, but to make things simple we assume exactly $n/k$ samples per cluster). The goal is to find the clusters $S_1,\dots,S_k$.

**Naive approach.** The naive approach is that "clusterness" property is that every in-cluster pair is closer than every cross-cluster pair. (Where closeness means Euclidean norm.) If this holds, you don't even need SoS: can just do a greedy clustering approach. It turns out that this approach works if there's enough mean separation. Let's write $\Delta_{ab} = \|\mu_a-\mu_b\|_2$ to denote the distance between the means of $a,b$. This approach works if $\Delta_{ab} \gg d^{1/4}$ (DasGupta, 2002).

Can we get things to work with smaller separation? What if there were 2 clusters and we knew the means? Given $X_{1:n}$, we could compute the projections

$$\left\langle X_i, \frac{\mu_1-\mu_2}{\|\mu_1-\mu_2\|}\right\rangle, \quad i\in[n].$$

These will be drawn from a mixture of 2 Gaussians in 1-dimension with unit variance and separation $\|\mu_1 - \mu_2\|_2$. This isn't rigorous, even statistically (since we have assumed we know $\mu_1 - \mu_2$), but it's a hint that we can do better.

**Theorem 5.5** (Regev-Vijayaraghavan, 2017)**.** *If $\Delta_{ab} \gg \sqrt{\log k}$ for all $a, b$, then using $(dk)^{O(1)}$ samples, we can cluster to 0.99-accuracy (and learn the means $\mu_1, \dots, \mu_k$).*

The above result is statistical: the best known algorithm is exponential time (basically discretize over all mean vectors, so you get $\exp(dk)$). We actually don't know how to improve exponential time if we are constrained to use *polynomially* many samples.

Today, it is implicit that $k, d$ are polynomially related.

**Theorem 5.6.** *For all $\epsilon > 0$, given $n \geq d^{(1/\epsilon)^{O(1)}}$ samples, we can cluster to 0.99 accuracy in $d^{(1/\epsilon)^{O(1)}}$ time if $\Delta_{ab} \gg k^\epsilon$.*

**Remark.** If you do the above argument carefully (we don't), you can actually get $d^{\log k}$ samples and time, for $\Delta \gg \sqrt{\log k}$. Also, a very recent paper (Li & Liu, STOC 2022) shows the following: if the covariances are equal to identity, namely the Gaussians $\mathcal{N}(\mu, I)$, then you can get $\Delta \gg \sqrt{\log k}$ in polynomial time and samples.

As we have discussed above, to prove the above theorem, thinking of $\epsilon = O(1)$, it is enough to construct a system of $d^{O(1)}$ polynomials $P_X$ which is satisfied by the indicator vectors $1_{S_a}$, for each cluster $a$, and which satisfies $P_X \vdash_{O(1)} \sum_{i \in S_a, j \in S_b} w_i w_j \leq \delta \cdot (n/k)^2$ for clusters $a \neq b$.

To come up with the polynomials $P_X$, we ask: what does it mean to be a cluster? The insight is that the meaning of clusterness should depend on the projections of the samples to 1-dimensional subspaces. The following fact is good inspiration:

**Fact 5.7.** *For any $t \in \mathbb{N}$, if $Y_1, \dots, Y_n \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succeq I$, then with high probability, for all $v$, $\|v\|^2 = 1$,*

$$\frac{1}{n} \sum_{i=1}^{n} \langle Y_i, v \rangle^t \leq O(t)^{t/2}$$

*if $n \gg d$.*

*Proof.* This follows from the fact that $\langle Y_i, v \rangle$ is a univariate Gaussian with variance at most 1, and then you use the MGF of a Gaussian. □

We therefore construct the following system $P_X(w)$: for all unit vectors $v$,

$$\frac{k}{n} \sum_{i=1}^{n} w_i \left\langle X_i - \frac{k}{n} \sum_{j=1}^{n} w_j X_j, v \right\rangle^t \leq O(t)^{t/2}. \tag{11}$$

If the $w_i$ is the indicator vector of a single Gaussian (i.e., 1 cluster) then the LHS is centering the $X_i$ from that cluster (using that the empirical average should be close to the true average), and is looking at the $t$th moment of that cluster in the direction of $v$. We hope that the following are true:

1. The above system $P_X$ is satisfied by the indicator vectors $1_{S_i}$, for $i \in [k]$. Fact 5.7 verifies this (technically we need concentration as well to ensure that the empirical error is close to expected error).

2. The above is satisfied by "only" $1_{S_1}, \dots, 1_{S_k}$.

3. The previous statement has an SoS proof.

We will prove the second statement using inequalities that we've proven (on the HW) hold in SoS.

*Of second statement above.* Fix $a \neq b$, and $t$ even. Let us define $\Delta_{ab} = \mu_a - \mu_b$, so that we have the equality of polynomials

$$\begin{aligned}\sum_{i \in S_a, j \in S_b} w_i w_j &= \sum_{i \in S_a, j \in S_b} w_i w_j \cdot \frac{\langle \mu_a - \mu_b, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} \\ &= \sum_{i \in S_a, j \in S_b} \frac{\langle \mu_a - \mu(w) + \mu(w) - \mu_b, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}},\end{aligned}$$

where $\mu(w) = \frac{k}{n} \sum_i w_i X_i$. Next, on the HW, we have shown an SoS approximate triangle inequality for $t$-norms, which shows the above is upper bounded (in SoS) by:

$$\sum_{i \in S_a, j \in S_b} w_i w_j \cdot 2^{O(t)} \cdot \left( \frac{\langle \mu_a - \mu(w), \Delta_{ab} \rangle^t + \langle \mu(w) - \mu_b, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} \right),$$

which is upper bounded by

$$\frac{n}{k} \sum_{i \in S_a} w_i \cdot 2^{O(t)} \cdot \frac{\langle \mu_a - \mu(w), \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} + \frac{n}{k} \sum_{i \in S_b} w_i \cdot 2^{O(t)} \cdot \frac{\langle \mu_b - \mu(w), \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}}.$$

Here we have bounded the summation over all $j \in S_b$ for the first term by $n/k$, and similarly for the second term. The above terms are symmetric, so we only bound the first. To do, so, let's add and subtract the $i$th sample for each $i$th term of the summation: the first term is equal to

$$\begin{aligned}&2^{O(t)} \cdot \frac{n}{k} \sum_{i \in S_a} w_i \cdot \frac{\langle \mu_a - X_i + X_i - \mu(w), \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} \\ \leq &2^{O(t)} \cdot \frac{n}{k} \sum_{i \in S_a} w_i \cdot \left( \frac{\langle \mu_a - X_i, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} + \frac{\langle \mu(w) - X_i, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} \right)\end{aligned}$$

where we have again used the approximate triangle inequality. The first term looks like what we know should be bounded, so we do some rearrangements:

$$\leq 2^{O(t)} \cdot \frac{n}{k} \left( \sum_{i \in S_a} \frac{\langle \mu_a - X_i, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} + \sum_i w_i \cdot \frac{\langle \mu(w) - X_i, \Delta_{ab} \rangle^t}{\|\Delta_{ab}\|^{2t}} \right) \tag{12}$$

For the first term, we have used that $w_i \leq 1$ in SoS, and in the second term, we have added a bunch of squares, since $t$ is even (i.e., we are now summing over all $i$, not just those in $S_a$). By Fact 5.7, the first term is at most $\frac{1}{\|\Delta_{ab}\|^{2t}} \cdot O(t)^{t/2} \cdot \|\Delta_{ab}\|^t \cdot \frac{n}{k}$. Furthermore, by the constraints in our polynomial system, we have that the second term is bounded by $\frac{1}{\|\Delta_{ab}\|^{2t}} \cdot O(t)^{t/2} \cdot \|\Delta_{ab}\|^t \cdot \frac{n}{k}$. Thus, the above is bounded above by

$$\frac{1}{\|\Delta_{ab}\|^t} \cdot O(t)^{t/2} \cdot (n/k)^2.$$

If the mean separation (i.e., an upper bound on $\|\Delta_{ab}\|$) is $k^\epsilon$, and $t \gg 1/\epsilon$, then the above is $\leq 1/k^{O(1)}$. (We want the above to be at most $\delta$, and we set $\delta = 1/k^{100}$, since we have a $\delta \cdot \text{poly}(k)$ probability of failure in the rounding algorithm.) $\square$

### 5.6 Sum of squares, squared

One last issue: the system of polynomials $P_X$ has infinitely many constraints, i.e., we need it to hold for all unit vectors $v$. We could try to discretize, but then still have $\exp(d)$ constraints, which is too many for our algorithmic needs (running time is polynomial in number of constraints).

To solve this, we will find a new system, $Q_X$, so that $Q_X$ has $\mathrm{poly}(n, d, k)$ polynomials in it, and $Q_X \vdash_{O(1)} P_X$, and $Q_X$ is satisfied by the true cluster indicators $1_{S_a}$. Since $Q_X$ implies $P_X$ and $P_X$ implies SoS identifability, it follows from compositionthat $Q_X$ implies SoS composability.

Let's consider the special case $t = 2$: we want to prove that for all $v$, $\frac{k}{n}\sum_i w_i \langle X_i - \mu(w), v\rangle^2 \le O(1)$. This looks a lot like saying that $O(1)\cdot I - \frac{k}{n}\sum_i w_i \cdot (X_i - \mu(w))(X_i - \mu(w))^\top \succeq 0$. This doesn't technically mean anything (it's a matrix of polynomials!). But remember the way we encoded this last week: we introduced a slack matrix of *variables* $B = (B_{ij})_{i,j\in[d]}$, and required that the above matrix be equal to $BB^{(t)}$, namely

$$O(1)\cdot I - \frac{k}{n}\sum_i w_i \cdot (X_i - \mu(w))(X_i - \mu(w))^\top = BB^\top.$$

Furthermore, since the variance of a Gaussian is bounded, there is some way to set the slack variables so as to satisfy the above. Also, it is straightforward to see that for any vector $v$, $v^\top BB^\top v$ is a sum-of-squares (in the variables $B$).

We can think of the matrix $BB^\top$ as a "little SoS proof" that a certain SoS statement holds for all $v \in \mathbb{R}^d$, $\|v\| = 1$.

**Generalizing to $t = 4$.** We now generalize to $t = 4$. We want to prove that for all $v$, $\frac{k}{n}\sum_i w_i \langle X_i - \mu(w), v\rangle^4 \le O(1)$. This looks like PSD-ness of

$$O(1)\cdot I_{d^2\times d^2} - \frac{k}{n}\sum_i w_i (X_i - \mu(w))^{\otimes 2}\left((X_i - \mu(w))^{\otimes 2}\right)^\top \succeq 0,$$

since we apply the above as a quadratic form to $v \otimes v$. There's a catch, since we need the system to be satisfied by the ground truth clusters!

In particular, to be concrete, if $Y_{1:m} \sim \mathcal{N}(0, I)$, is it true that $\frac{1}{m}\sum_i (Y_i \otimes Y_i)(Y_i \otimes Y_i)^\top \preceq O(1)\cdot I$. The answer is **no**! In directions of the form $v \otimes v$, this is true; but in other directions, this can be big! For instance, if you consider $u \in \mathbb{R}^{d^2}$ defined by $u_{ii} = \frac{1}{\sqrt{d}}$, $u_{ij} = 0$, then the quadratic form is bigger than $\|u\|_2^2$.

The issue with the above is that we're committing ourselves to a particular type of mini-SoS proof (which doesn't work). All we need is a proof that is satisfied by $(1_{S_a}, B)$ for each cluster $a$. To get around this, we use the following fact:

**Lemma 5.8.** *With high probability, for $m \gg d^{O(1)}$, $\{\|v\|^2 = 1\} \vdash_{O(1)}^{v} \frac{1}{m}\sum_i \langle Y_i - \frac{1}{m}\sum_j Y_j, v\rangle^4 \le O(1)$.*

So, $Q_X$ should be solved by $(w, b)$, where $b$ is a list of coefficients in the "inner" SoS proof of Lemma 5.8. To ensure this, it is simply a system of inequalities and equalities. In more deteail, suppose the SoS proof of Lemma 5.8 is of the following form:

$$O(1)\cdot\|v\|_2^4 - \frac{1}{m}\sum_{i=1}^m \left\langle Y_i - \frac{1}{m}\sum_j Y_j, v\right\rangle^4 = \sum_{\ell=1}^L p_\ell(v)^2.$$

Notice that (11) is satisfied if

$$O(1) \cdot \|v\|_2^4 - \frac{k}{n} \sum_{i=1}^{n} w_i \left\langle X_i - \frac{k}{n} \sum_{j=1}^{n} w_j X_j, v \right\rangle^t = \sum_{\ell=1}^{L} p_\ell(v)^2 \tag{13}$$

for all vectors $v \in \mathbb{R}^d$. Then we "match up coefficients" in (13) as follows: we have variables $p_{\ell,S}$ for each polynomial $p_\ell$ and each subset $S \subset [d]$ of the coordinates corresponding to the term $v^S$, and have constraints of the form $\{\sum_{\ell=1}^{L} p_{\ell,S} = q_{\ell,S}(S, w)\}$ where $q_{\ell,S}(Y)$ is a polynomial of the $X$s and $w$s and is the corresponding coefficient of $v^S$ on the right-hand side. Thus, the system $Q_X$ has variables $(w, \{p_{\ell,S}\}_{\ell,S})$. Furthermore, validity of the coefficient matching constraints $\{\sum_{\ell=1}^{L} p_{\ell,S} = q_{\ell,S}(S, w)\}$ ensures that (13) is an equality *of polynomials* and thus that (11) holds for all $v$. There's a crucial step remaining (which is what failed with our previous approach!): with high probability over $X_{1:n}$, does any of the cluster vectors $w = 1_{S_a}$ satisfy (13)? For such a setting of $w$, (13) reduces to the setting of $m = n/k$ and $Y_j = w_{\phi(j)} X_{\phi_j} = X_{\phi(j)}$, where $\phi(j)$ is the $j$th coordinate where $w$ is nonzero. Then Lemma 5.8 tells us that the equality (13) holds with high probability over the draw of $X$.

Finally, we prove Lemma 5.8.

*Proof.* We prove the version without the empirical mean. We have:

$$\frac{1}{m} \sum_i \langle Y_i, v \rangle^4 = \frac{1}{m} \sum_{p,q,r,s,i} v_p v_q v_r v_s \cdot \sum_i Y_{ip} Y_{iq} Y_{ir} Y_{is}.$$

With at least $\text{poly}(d)$ samples, unless any 2 of $p, q, r, s$ are equal, $\sum_{p,q,r,s} Y_{ip} Y_{iq} Y_{ir} Y_{is}$ has mean 0 and concentrates to $1/\text{poly}(d)$. So, the above is equal to:

$$\sum_{p,q} v_p^2 v_q^2 \cdot O(1) + \frac{1}{\text{poly}(d)} \cdot (\textit{rest of terms}) \leq O(1) \cdot \|v\|_2^4 + O(1) = O(1).$$

□

The idea is that using SoS proofs we can prove more than what we can do using the eigenvalue fact that we tried for the case $t = 2$.

# 6 October 26, 2022

## 6.1 Tensor decomposition

We first define the tensor decomposition problem. The problem is defined as follows:

**Definition 6.1** (Tensor decomposition)**.** Given $T \in \mathbb{R}^{n^k}$, the goal is to find a low-rank tensor $T'$ so that $T' \approx T$, i.e., $\|T' - T\|$ is small in some norm.

Here, by low-rank, we mean that we can write, e.g., $T' = \sum_{i=1}^{r} a_i \otimes b_i \otimes c_i$, for $a_i, b_i, c_i \in \mathbb{R}^n$ (in the case that $k = 3$).

**Why do we care about tensor decomposition?** To explain the motivation, we begin by reviewing PCA: we're given a dataset $X_1, \dots, X_N \in \mathbb{R}^n$. The goal is to find a direction $v$ such that the data has large variance in the direction of $v$:

$$\max_{\|v\|=1} v^\top \left( \frac{1}{N} \sum_{i=1}^{N} (X_i - \mathbb{E}_j X_j)(X_i - \mathbb{E}_j X_j)^\top \right) v = \max_v \frac{1}{N} \sum_{i=1}^{n} \langle X_i - \mathbb{E}_j X_j, v \rangle^2. \tag{14}$$

Why do we care about PCA?

- Dimension reduction (do the above problem multiple times to find the important dimensions in a dataset).
- Denoising (throw away the noisy directions).
- Inference/model fitting.

Algorithms for PCA: the problem (14) is an eigenvalue problem, i.e., we can solve it by finding the maximum eigenvalue/eigenvector pair for the covariance matrix. To be explicit, if we write that matrix as $\Sigma = \sum_i \lambda_i v_i v_i^\top$, with $\lambda_1 \geq \cdots \geq \lambda_n$, then $v_1$ is the principle component (namely, it solves (14), and it is also the best rank-1 approximation to the covariance. In particular, for lots of norms (Frobenius, spectral, etc), $\|\lambda_1 v_1 v_1^\top - \Sigma\|$ is smallest for $v_1$ as the principal eigenvector, among all rank-1 matrices.

**Mixtures/non-homogeneous datasets** Suppose our dataset is generated as a mixture (i.e., a non-homogeneous dataset). Letting the two components of the mixture have means $\mu_1, \mu_2$, we might hope that $\frac{\mu_1 - \mu_2}{\|\mu_1 - \mu_2\|}$ is approximately the principal component of the dataset (e.g., if the two clusters are sufficiently separated). In particular, interesting structure in a dataset is hiding in the low-rank component.

Today, the focus is on situations where PCA *doesn't* find the structure in data. There are two such problems/situations:

- The *overcompleteness* problem: suppose there are many more interesting directions than there are dimensions. For instance, if there are $d^2$ clusters but only $d$ dimensions. Suppose all the means of the clusters are interesting to us. The covariance only has $d$ eigenvectors, so can't hope to recover all means from it. (Besides, the low-variance directions are probably noise.)
- The rotation problem: perhaps the high-variance directions are unique only up to a subspace in $\mathbb{R}^n$ that they span. To be more precise: suppose we have $n$ dimensions, and $e_1, e_2$ are the first two coordinates: suppose we have two clusters along each of the $e_1, e_2$ axes. To be precise, suppose we have the mixture distribution:

$$\frac{1}{4} \cdot (\mathcal{N}(e_1, I) + \mathcal{N}(-e_1, I) + \mathcal{N}(e_2, I) + \mathcal{N}(-e_2, I)). \tag{15}$$

  We would hope that PCA learns the directions $(e_1, e_2)$. It turns out that PCA in two dimensions will give us back $\text{span}(e_1, e_2)$. But this doesn't tell us the specific directions: in particular, we could have had cluster centers at $\frac{1}{\sqrt{2}} \cdot (\pm e_1 \pm e_2)$, which would be indistinguishable from a PCA perspective from the above problem.

## 6.2 Moving beyond 2nd degree polynomials

So what can we do in light of the limitations of PCA? Perhaps we can hope that higher-degree polynomials give us additional information: e.g., $\frac{1}{N}\sum_{i=1}^N \langle X_i - \mathbb{E}_j X_j, v\rangle^t$, for some power $t \in \{3,4,\ldots\}$. When we do this, we end up with tensors $\frac{1}{N}\sum_{i=1}^N (X_i - \mathbb{E}_j X_j)^{\otimes t}$, as the polynomials in question are $\langle \frac{1}{N}\sum_{i=1}^N (X_i - \mathbb{E}_j X_j)^{\otimes t}, v^{\otimes t}\rangle$.

Why can this solve overcompleteness? The parameter-dimensions counting argument for overcompleteness breaks down: a $k$-tensor has $n^k$ numbers (up a $k$-dependent constant for symmetry). We can hope for a $k$-tensor to specify as many as $n^{k-1}$ directions: since each direction is specified by $n$ numbers.

What about rotation invariance: Imagine the worst possible problem for rotation invariance: suppose we have $n$ directions, $a_1,\ldots,a_n \in \mathbb{R}^n$, which are orthonormal and the covariance is $\Sigma = \sum_i a_i a_i^\top$. But by orthonormality, this covariance is the identity matrix $I_n$. But then for any $v$, $\sum_i \langle v, a_i\rangle^2 = v^\top \Sigma v = \|v\|^2$, which contains no informatino about the directions $a_i$.

What if we could understand the polynomial $\sum_i \langle v, a_i\rangle^3$? This polynomial is maximized at the $a_i$s. Why is that? WLOG, we can take $a_i = e_i$, and then this polynomial is $\sum_i v_i^3$. Next, under the constraint $\|v\| = 1$, this is maximized at the unit vectors. To see this, note that $\sum_i v_i^3 \le \max_i \sum_i v_i^2 = \max_i v_i$. You can make the sum equal to 1 by choosing $v = e_i$ for some $i$. And whenever it's not a unit vector, the objective is thus strictly less than 1.

**A more concrete example.** As a further example, let's remember the distribution $D$ defined in (15), which is a mixture of 4 standard Gaussians with means at $\pm e_1, \pm e_2$. It has mean 0. What is the covariance of $D$?[2] It is:

$$\mathbb{E} XX^\top = \frac{1}{2}(e_1 e_1^\top + I) + \frac{1}{2}(e_2 e_2^\top + I) = I_n + \frac{1}{2}\cdot I_2,$$

which picks out the span of the first two vectors but doesn't tell us anything about the direction there. What about the 4th moment:

$$\mathbb{E}\langle X, v\rangle^4 = \frac{1}{2}\mathbb{E}\langle e_1 + g, v\rangle^4 + \frac{1}{2}\mathbb{E}\langle e_2 + g, v\rangle^4,$$

where we have used that $g \sim \mathcal{N}(0,1)$ above is symmetric (so the moments from the $\pm e_1, \pm e_2$ mixtures are the same). Expanding the above, we get:

$$\begin{aligned}
&\frac{1}{2}\left(\langle e_1, v\rangle^4 + 6\langle e_1, v\rangle^2 \mathbb{E}\langle g, v\rangle^2 + \mathbb{E}\langle g, v\rangle^4\right) + \frac{1}{2}\left(\langle e_2, v\rangle^4 + 6\langle e_2, v\rangle^2 \mathbb{E}\langle g, v\rangle^2 + \mathbb{E}\langle g, v\rangle^4\right) \\
=&\frac{1}{2}\langle e_1, v\rangle^4 + \frac{1}{2}\langle e_2, v\rangle^4 + 3 + 6\cdot\|v\|^2 = \frac{1}{2}\langle e_1, v\rangle^4 + \frac{1}{2}\langle e_2, v\rangle^4 + 9.
\end{aligned}$$

where we have used that $\mathbb{E}\langle g, v\rangle^2 = 1, \mathbb{E}\langle g, v\rangle^4 = 3$. The key point is that the above 4th degree polynomial is maximized as $v \in \{e_1, e_2\}$ (as we saw above for 3rd degree). So, if we could solve this 4th-degree optimization problem (or, find a low-rank approximation to the 4th moment tensor), then we would be able to find $e_1, e_2$.

[2] Note that today we will assume that we draw enough samples so that the empirical covariance/moments approximate the expected ones.

**A real example: independent component analysis ("blind source separation").** Let's move beyond toy problems. The idea here is that we have a bunch of microphones in a room with a bunch of people talking: each microphone is picking up a different (linear) combination of a bunch of people speaking. The problem is to figure out what each person is saying. You don't know where the people are, so you don't know the particular linear combinations (if you knew them, you could invert the linear transformation taking signals to measurements).

Assume that $X$ is a $\mathbb{R}$-valued random variable, with:

1. Coordinates of $X$ are iid, with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$. Here $X$ is the signal vector.
2. There is an unknown "mixing matrix" $A \in \mathbb{R}^{n\times n}$ which is full-rank.

We observe $AX^{(1)}, \dots, AX^{(N)}$, where $X^{(1)}, \dots, X^{(N)}$ are independent relaizations of $X$. The goal is to find $A, X^{(1)}, \dots, X^{(N)}$ (up to sign – note that we can't distinguish between $A$ and $-A$, since we can also negate $X$).

Why is this problem even well-defined (identifiable)? The concern is: how can we tell the difference between observing $AX$ and $AOO^\top X$, where $O$ is an orthogonal matrix? The point is that we assumed the coordinates of $X$ are independent. In order for the coordinates of $O^\top X$ to be independent, we (roughly speaking) need the distribution of $X$ to be rotation-invariant. So, we will need to rule out $X$ being rotation invariant (note that a Gaussian is the canonical rotation-invariant random variable).

So, we assume:

**Assumption 6.1** (Non-gaussian)**.** *We assume that $X$ is "noticeable" non-Gaussian, i.e., for all $i \in [n]$, $\mathbb{E}X_i^4 \neq 3$.*

It turns out that the 4th moment being 3 is more or less forced by rotation invariance. (To do specific sample complexity analyses, we will need that the 4th moment is far from 3, but we won't do finite-sample analysis today.)

Let's let the columns of $A$ be denoted $A_i$, $i \in [n]$. We look at the covariance of $AX$:

$$\mathrm{Cov}(AX) = \mathbb{E}[AXX^\top A^\top] = A\mathbb{E}[XX^\top]A^\top = AA^\top = \sum A_i A_i^\top =: W,$$

where we have used that the covariance of $X$ is the identity (by independence of coordinates and unit variance of each coordinate).

It is straightforward to check that $W^{-1/2}A$ is an orthogonal matrix. So, we can assume that $A$ is an orthogonal matrix WLOG (in particular, we can compute $\mathrm{Cov}(AX)$ since we see samples of $AX$, and thus we can compute $W$, and so we can take the observations $AX^{(i)}$ and hit them with $W^{-1/2}$, i.e., we can take $W^{-1/2}AX^{(i)}$ as our data instead).

Now, we look at moments:

$$\mathbb{E}(AX)^{\otimes 4} = \mathbb{E}\left(\sum_i X_i A_i\right)^{\otimes 4} = \mathbb{E}\sum_{i,j,k,\ell} X_i X_j X_k X_\ell A_i \otimes A_j \otimes A_k \otimes A_\ell. \tag{16}$$

If $i$ appears once in $(i,j,k,\ell)$, then $\mathbb{E}[X_iX_jX_kX_\ell] = \mathbb{E}[X_i]\cdot\mathbb{E}[X_jX_kX_\ell] = 0$ by independence and mean-0 of coordinates. So, all nonzero terms must have some matching. So suppose we have $i = j \neq k = \ell$. Then we have: $\mathbb{E}[X_iX_jX_kX_\ell] = \mathbb{E}[X_i^2]\mathbb{E}[X_k^2] = 1$. Furthermore, by permutations, we have 6 copies of this.

Finally, if $i = j = k = \ell$, then $\mathbb{E}[X_i X_j X_k X_\ell] = \mathbb{E}[X_i^4]$, which is not 3.

Now, we can write (16) as

$$\mathbb{E}(AX)^{\otimes 4} = \sum_{i,j}(A_i \otimes A_i \otimes A_j \otimes A_j + A_i \otimes A_j \otimes A_i \otimes A_j + \cdots) + \sum_i (\mathbb{E}X_i^4 - 3) \cdot A_i \otimes A_i \otimes A_i \otimes A_i.$$

Let's write the inner product of the first term and $v^{\otimes 4}$:

$$\left\langle \sum_{i,j} A_i \otimes A_i \otimes A_j \otimes A_j, v^{\otimes 4} \right\rangle = \sum_{i,j} \langle A_i, v\rangle^2 \langle A_j, v\rangle^4 = \left(\sum_i \langle A_i, v\rangle^2\right)^2 = \|v\|^4 = 1.$$

Thus, up to a constant, we get the polynomial we want, which is $\sum_i \langle A_i, v\rangle^4$ (up to a nonzero constant factor out front).

### 6.3 Algorithms for tensor decomposition

In the worst case, finding the best low-rank approximation of a tensor, and even finding a single vector which maximizes the cubic/quartic form, are NP-hard in the worst case. So, you need to make assumptions: we will rephrase the problem:

**Problem 6.2.** Given a symmetric tensor $T = \sum_{i=1}^r a_i \otimes a_i \otimes a_i + E$, find $a_1, \ldots, a_r \in \mathbb{R}^n$.

We make the following assumptions on $a_1, \ldots, a_r$:

- $a_1, \ldots, a_r$ are orthonormal (this can be replaced by linear independence by using a trick which is similar to the whitening trick);
- $E = 0$;

There is a very nice algorithm called Jenrich's algorithm for the above problem:

- Sample $g \sim \mathcal{N}(0, I)$.
- Let's unfold the tensor $T$, and write it as a $n^2 \times n$ matrix $T_{\{1,2\},\{3\}}$. Then $T_{\{1,2\},\{3\}} g = \sum_{i=1}^r \langle g, a_i\rangle a_i \otimes a_i$, which we can view as the matrix $\sum_{i=1}^r \langle g, a_i\rangle a_i a_i^\top$. With probability 1, the numbers $\langle g, a_i\rangle$ are all distinct. By orthonormality of the vectors $a_i$, the eigendecomposition of this matrix gives us the vectors $a_i$.

To deal with the case where the $a_i$ are only linearly independent (and not orthonormal): we take a different vector $g' \sim \mathcal{N}(0, 1)$, look at the contractions with $g, g'$, and then match up eigenvalues. We don't go into details.

Jennrich's algorithm has a few drawbacks:

- Noise robustness. Note that PCA is very robust to errors: only the smaller eigenvalues should be affected by small noise. Unfortunately, the noise robustness of Jenrich's algorithm isn't that good.

  To be precise, what if $E = \lambda G$, if $\lambda \in \mathbb{R}$ and $G_{ijk} \sim \mathcal{N}(0, 1)$. This is a generic model we could think about with random-looking errors. How big can $\lambda$ be? When we do the random contraction, we get the matrix:

$$\sum_{i=1}^r \langle g, a_i\rangle a_i a_i^\top + \sum_{j=1}^n g_j E_j, \tag{17}$$

where the $E_j$ are the $n \times n$ slices of the tensor $E$. Note that $\langle g, a_i \rangle \sim \mathcal{N}(0,1)$ since the $a_i$ are unit norm. So, the maximum value of $\langle g, a_i \rangle$ is $\Theta(\sqrt{\log n})$, which are approximate as $O(1)$. We might hope that $\sum_j g_j E_j$ has small eigenvalues, which will let us recover the maximum few eigenvalues/eigenvectors of the first (signal) matrix. The random matrix $\sum_j g_j E_j$ is a sum of random matrices, each of whose entries is of size roughly $\lambda\sqrt{n}$. So, by standard matrix concentration inequalities (matrix Bernstein), we get that the max eigenvalue, namely $\| \sum_j g_j E_j \| \approx \lambda n$. So, to have any hope of Jenrich's algorithm working here, we need $\lambda \ll 1/n$.

A natural question: can we beat $\lambda \approx 1/n$? What if we could look for maxima of polynomials of the form $\langle T, v^{\otimes 3} \rangle$? In particular, we have $\sum_i \langle v, a_i \rangle^3 + \langle E, v^{\otimes 3} \rangle$: the first polynomial is maximized at $v \in \{a_i\}_i$, and it has value $\Theta(1)$ there. If the second polynomial is $o(1)$ for $\|v\| = 1$, then we'd be in good shape. So how big can we take $\lambda$ to ensure that $\max_v \langle E, v^{\otimes 3} \rangle \leq o(1)$? We compute:

$$\langle E, v^{\otimes 3} \rangle = \lambda \langle G, v^{\otimes 3} \rangle.$$

If $v$ is fixed, then $\langle G, v^{\otimes 3} \rangle \sim \mathcal{N}(0,1)$ since $\mathbb{E}_G \sum_{ijk} (G_{ijk} v_i v_j v_k)^2 = \sum_{i,j,k} v_i^2 v_j^2 v_k^2 = \|v\|^6 = 1$. To constrain the maximum over $v$, we take a net over the unit sphere with $2^{O(n)}$ points, we note that $\Pr(\langle G, v^{\otimes 3} \rangle \gg n) \leq e^{-100n}$, and so we get $\max_v \langle G, v^{\otimes 3} \rangle \leq O(\sqrt{n})$ wigh high probability. So, we only need $\lambda \ll 1/\sqrt{n}$.

Some more intuition for why Jennrichs' doesn't work: roughly speaking, it corresponds to upper bounding $\max_v \langle G, v^{\otimes 3} \rangle \leq \max_{u \in \mathbb{R}^n, W \in \mathbb{R}^{n^2}} \langle G, v \otimes W \rangle = \sigma_{\max} G_{\{1,2\},\{3\}}$, which will be large.

- Second drawback: Take $a_1, \ldots, a_r$, for $r \gg n$, where $a_1, \ldots, a_n$ are "generic". It turns out that if the $a_i$ are generic (e.g., perturbed slightly), then $a_1 \otimes a_1, \ldots, a_r \otimes a_r$, $r \ll n^2$ are linearly independent. So, if we were willing to look at the 6-tensor, then we could look at: $\sum_i a_i^{\otimes 6} = \sum_{i=1}^r (a_i \otimes a_i)^{\otimes 3}$, and we are back in the setting of Jenrich's algorithm (by "lifting up"). The cost is you have to look at a higher-order tensor.

  So, you want to look at the best possible tradeoff between the order of the tensor the number of linearly independent vectors. In general, if $r = n^d$, we need a $3d$-order tensor. What we can hope for is: if $r = n^d$, then we should only need a $(d+1)$-order tensor.

We give an algorithm that addresses the noise robustness issue and also the overcompleteness one, but only in a messy way.

In particular: the best algorithms we know: we can get $r = n^d$ and a $2d$-order tensor, under some assumptions (improving the situation here is somewhat of an open problem). The basic idea is as follows: take a $2d$-tensor and transform it into a $3d$-tensor of the same vectors, and do Jenrich's algorithm on the "dreamed-of" $3d$-tensor. This approach doesn't use the full machinery of SoS (just spectral approach based on eigenvalues/eigenvectors).

### 6.4 Improving the noise robustness

Suppose the error tensor $E$ satisfies the following:

$$\{\|v\|^2 = 1\} \vdash_d \langle E, v^{\otimes 3} \rangle \leq o(1).$$

When does this assumption hold? Note that

$$\max_v \langle E, v^{\otimes 3}\rangle = \lambda \max_v \sum_{ijk} G_{ijk} v_i v_j v_k,$$

which looks like the polynomial we discussed in the contex of refuting random CSPs. It turns out that you can apply the same arguments (Cauchy-Schwarz, get a random matrix, apply matrix Bernstein on that random matrix) as we used before gives that with high probability, we can get:

$$\{\|v\|^2 = 1\} \vdash_d \langle E, v^{\otimes 3}\rangle \leq O(n^{3/4}).$$

Thus, we can get $\lambda \ll 1/n^{3/4}$ (so, we can get halfway from $\lambda = 1/n$ to $\lambda = 1/\sqrt{n}$); there is some evidence that doing better is computationally intractable (i.e., this is state-of-the-art).

The idea is to use SoS to optimize the polynomial, and then do something with the resulting pseudo-expetation:

$$\max_{\widetilde{\mathbb{E}} \vDash \{\|v\|^2=1\}} \widetilde{\mathbb{E}} \langle T, v^{\otimes 3}\rangle.$$

What property do we want that $\widetilde{\mathbb{E}}$ satisfies? We could end up with $\widetilde{\mathbb{E}}$ so that $\widetilde{\mathbb{E}} p(v) = p(a_i)$; this might be nice, so it could help us recover $a_i$, but not the other components (analogously to what we had last week regarding mixture models). What we want, rather, is the uniform distribution on $\{a_1, \ldots, a_n\}$. If we were given the third moments of this distribution, it would be equal to a very nice tensor: $\mathbb{E}_{v \sim \{a_1,\ldots,a_n\}} v^{\otimes 3} = \frac{1}{n}\sum_i a_i^{\otimes 3}$, it is the tensor we want, *with no noise*!

So, we will solve the above maximization problem involving the pseudoexpectation, with some additional constraints, and thene will "round" it in a certain way using Jenrich's algorithm. In particular, we will solve

$$\max_{\widetilde{\mathbb{E}} \vDash \{\|v\|^2=1\}} \widetilde{\mathbb{E}} \langle T, v^{\otimes 3}\rangle, \text{ s.t. } \widetilde{\mathbb{E}} v v^\top = \frac{I}{n}, \|\widetilde{\mathbb{E}} v (v \otimes v)^\top\| \leq \frac{1}{n},$$

where the second term is the $n^2 \times n$ flattening of $\widetilde{\mathbb{E}} v^{\otimes 3}$. The idea behind the constraints is as follows: for the uniform distribution over the $a_i$, we certainly have $\mathbb{E} a_i a_i^\top = \frac{I}{n}$ (by orthonormality), and the final term is $1/n$ for the uniform distribution over the $a_i$s by orthonormality: we have that $a_i$ and $a_i \otimes a_i$ are pairwise orthogonal for $i \neq j$, and in general if we have orthogonal vectors $u_1, \ldots, u_m, v_1, \ldots, v_m$, then the matrix $\sum_{i=1}^m u_i v_i^\top$ has spectral norm bounded by 1. This also shows the above optimization problem is feasible.

Thus the algorithm is as follows:

1. Solve for $\widetilde{\mathbb{E}}$.
2. $n^{O(1)}$ times, sample $g \sim \mathcal{N}(0, I)$, and write $M_g := \widetilde{\mathbb{E}}[\langle g, v\rangle v v^\top]$ (similar to Jenrich's but polynomially-many times).
3. If the span of the top $O(1)$ eigenvectors of $M_g$ contains any $a$ so that $\langle T, a^{\otimes 3}\rangle \geq 1 - o(1)$, keep $a$ (here we take a net over the subspace and do a brute-force search).
4. Discard duplicates.

To analyze this, we need to show that with decent probability, each vector $a_i$ shows up with decent probability in the top $O(1)$ eigenvectors of $M_g$. Note that this algorithm never has a false positive: the analysis we did before shows that $T$ is maximized at the vectors $a_i$ and nowhere else.

## 6.5 Proving noise robustness for tensor decomposition theorem

We now state the guarantee for the algorithm from the previous subsection:

**Theorem 6.2.** *The above algorithm satisfies the following: it finds $b_1, \dots, b_m$ so that $m \le n$, and for $0.99n$ of the $a_i$s, there exists some $b_j$ so that $\langle a_i, b_j \rangle \ge 1 - o(1)$.*

You can actually recover everything: take the ones you recovered and subtract them off from the input tensor. By orthonormality, it suffices to assume that $a_i = e_i$, so that $\langle a_i, v \rangle = v_i$ for each $i$.

**Lemma 6.3.** *For $0.99$ of the $i$'s:*

1. *$\widetilde{\mathbb{E}} v_i^3 \ge \frac{1-o(1)}{n}$. (i.e., the "spreading out" idea worked, since we put mass on many of the $v_i^3$)*
2. *It holds that $\|\widetilde{\mathbb{E}} X_i X X^\top\|_F^2 \le O(1/n^2)$.*

*Proof.* We prove the first statement first. We compute $\widetilde{\mathbb{E}} \left( \sum_i v_i^{\otimes 3} + \langle v_i^{\otimes 3}, E \rangle \right) \ge 1 - o(1)$ since we can take the pseudoexpectation to be the uniform distribution over the $a_i$, which will make this $1 - o(1)$ (assuming appropriate bound on $E$). Furthermore, by the constraint, $|\widetilde{\mathbb{E}} \langle v^{\otimes 3}, E \rangle| \le o(1)$, which implies that $\widetilde{\mathbb{E}} \sum_i v_i^3 \ge 1 - o(1)$.

It remains to check that the coordinates are spread out. Sicne we can prove in SoS that $v_i \le 1, v_i \ge -1$, we have

$$\widetilde{\mathbb{E}} v_i^3 \le \widetilde{\mathbb{E}} v_i^2 = e_i^\top \widetilde{\mathbb{E}} [v v^\top] e_i \le 1/n,$$

since we have that $\widetilde{\mathbb{E}}[v v^\top] = I/n$ (constraint of the pseudoexpectation). Thus, the only way we can have $\widetilde{\mathbb{E}} \sum_i v_i^3 \ge 1 - o(1)$ is to have $\widetilde{\mathbb{E}} v_i^3 \ge (1 - o(1))/n$ for each $i$.

We prove the second statement: we use an averaging argument:

$$\sum_i \|\widetilde{\mathbb{E}} X_i X X^\top\|_F^2 = \sum_i \operatorname{Tr} \left( \widetilde{\mathbb{E}}[X_i X X^\top] \cdot \widetilde{\mathbb{E}}[X_i X X^\top] \right) = \sum_i \operatorname{Tr} \widetilde{\mathbb{E}}[X_i X_i' \langle X, X' \rangle X (X')^\top],$$

which is equal to

$$\sum_i \widetilde{\mathbb{E}} X_i X_i' \langle X, X' \rangle^2 = \widetilde{\mathbb{E}} \langle X, X' \rangle^3 = \widetilde{\mathbb{E}} X^\top (\widetilde{\mathbb{E}} X' (X' \otimes X')^\top)(X \otimes X).$$

By the constraints, the matrix $\widetilde{\mathbb{E}} X' (X' \otimes X')^\top$ has singular values upper bounded by $1/n$, and it follows that the right-hand side above is:

$$\le \|\widetilde{\mathbb{E}} X' (X' \otimes X')^\top\| \le 1/n.$$

By Markov's inequality, it follows that a $0.99$ fraction of the $\|\widetilde{\mathbb{E}} X_i X X^\top\|_F^2$ are at most $O(1/n^2)$. □

**Claim 6.4.** *Define good $i$ as the $i$s satisfying the constraints of the previous lemma. Then god foor $i$, with probability $\ge 1/n^{O(1)}$, the algorithm finds some $a$ so that $a_i \ge 1 - o(1)$.*

*Proof.* We write $M_g$ as follows:

$$\widetilde{\mathbb{E}}\langle X, g\rangle XX^\top = g_1\widetilde{\mathbb{E}}[X_1XX^\top] + \widetilde{\mathbb{E}}\langle g_{rest}, X\rangle XX^\top,$$

where $g_{rest}$ is the rest of the coordinates of $g$. If we are lucky, $g_1 \gg \sqrt{\log n}$, which happens with inverse polynomial probability. We also hope that $\|\mathbb{E}\langle g_{rest}, X\rangle XX^\top\| O(\sqrt{\log n}/n)$.

For the first term, we hope that there are only a few eigenvalues which are $1/n$, and among those eigenvectors are the first coordinate vector (thus the algorithm will find the first coordinate vector in its search over the top $O(1)$ eigenvalues).

Note that the $(1,1)$-entry of $\widetilde{\mathbb{E}}[X_1XX^\top]$ is $\widetilde{\mathbb{E}}[X_1^3]$, which we know is $\geq (1-o(1))/n$. Furthermore, the Frobenius norm of $\widetilde{\mathbb{E}}[X_1XX^\top]$ is $O(1/n^2)$, which means that there are only $O(1)$ eigenvalues which are $\gg \epsilon/n$, meaning that the span of eigenvalues $\geq \epsilon/n$ contains some $a$ so that $a \geq 1-o(1)$.

Now let's discuss the second term: it is $\sum_i g_i\widetilde{\mathbb{E}}[X_iXX^\top]$. We have a bunch of Gaussians multiplying some fixed matrices. We will use Matrix Bernstein: what matters is the variance:

$$\sum_{i>1}(\widetilde{\mathbb{E}}X_iXX^\top)^2 \leq 1/n^2,$$

since we have that $\|\widetilde{\mathbb{E}}[X(X\otimes X)^\top\| \leq 1/n$, which implies that, by Matrix Bernstein inequality, $\mathbb{E}\|\sum_i g_i\widetilde{\mathbb{E}}[X_iXX^\top]\| \leq O(\sqrt{\log n}/n)$. □

# 7 October 28, 2022

Today: lower bounds, Part 1. Overall idea: perhaps there is some sense in whcih SoS is "optimal"? As a warmup, we talk about max-cut. We proved on the probelm set that: given a graph with a $(1-\epsilon)$-cut, we can find a $(1-O(\sqrt{\epsilon}))$-cut. Can be rephased as follows: if the max cut of $G$ is $1-\epsilon$, then:

$$\max_{\widetilde{\mathbb{E}} \text{ of deg } 2} G(x) \leq 1 - \Omega(\epsilon^2).$$

One question is whether we can do better than this? Turns out the answere is no.

## 7.1 SoS lower bounds for simple instances

bounds Before stating, recall that for a $n$-cycle with $n$ even, the max cut is 1, and the with $n$ odd, the max cut is $\leq 1 - 1/n$.

**Theorem 7.1.** *Suppose we take the $n$-vertex cycle. Then there is a degree-2 $\widetilde{\mathbb{E}}$ on the hypercube so that $\widetilde{\mathbb{E}}[G(x)] \geq 1 - O(1/n^2)$.*

This statement is really about the *limitations of our (SoS) tools*; it's certainly not a hard problem, and interpreting about what it means in terms of hardness of max-cut requires different instances.

*Proof.* We construct the pseudoexpectation $\widetilde{\mathbb{E}}$ explicitly: it is enough to specify $\widetilde{\mathbb{E}}[x]$ and $\widetilde{\mathbb{E}}[xx^\top]$. Note that for max-cut, swapping 0 and 1 doesn't reflect the value of $\widetilde{\mathbb{E}}[G(x)]$, so we will take $\widetilde{\mathbb{E}}[x] = \frac{1}{2}\cdot\mathbf{1}$. For the second moments, let $\omega = 2^{2\pi i/n}$ be a primitive $n$th root of unity. Furthermore, write $n = 2k+1$. Next, define

$$u = (1, \omega^k, \omega^{2,}, \ldots, \omega^{(n-1)k}) \in \mathbb{C}^n.$$

Also let $v = \Re(u), w = \Im(w)$, so that $u = v + iw$. Furthermore, set $X = vv^\top + ww^\top$. Note that $X_{ii} = v_i^2 + w_i^2$. Finally, let $\widetilde{\mathbb{E}}[xx^\top] = \frac{1}{4} \cdot \mathbf{1}\mathbf{1}^\top + \frac{1}{4} \cdot X$.

Note that a rank-1 pseudoexpectation would have to be of the form $xx^\top$ for a cut $x$, so we would get a cut; thus, it's a bit surprising that by going up to constant rank, we can construct this counterexample which is tight. We have to check that $\widetilde{\mathbb{E}}$ is a pseudoexpectation:

- First, we check $\widetilde{\mathbb{E}} \vDash \{x_i^2 = x_i\}$. In particular, for all polynomials $p$ so that $p(x)x_i^2$ is of degree at most 2, we need $\widetilde{\mathbb{E}}[p(x)x_i^2 - p(x)x_i] = 0$. So, only need to check for constant polynomial $p$. But $\widetilde{\mathbb{E}}[x_i^2] = 1/4 + 1/4 = 1/2 = \widetilde{\mathbb{E}}[x_i]$, so good.

- Next, we need to check that $\widetilde{\mathbb{E}}[p^2] \geq 0$ for all linear $p$. It is enough to check that $(1, x)(1, x)^\top$ is PSD. This matrix is:

$$\begin{pmatrix} 1 & \mathbf{1/2} \\ \mathbf{1/2} & \frac{1}{4}11^\top + \frac{1}{4}X \end{pmatrix} = \begin{pmatrix} B & C^\top \\ C & A \end{pmatrix}.$$

  A block matrix of this form is PSD Iff it's schur complement is PSD. The Schur complement is $A - CB^{-1}C^\top = \frac{1}{4}11^\top + \frac{1}{4}X - \frac{1}{4}11^\top \succeq 0$. This argument should demystify the $\frac{1}{4}11^\top$: know you will subtract off that from the $-CB^{-1}C^\top$ and the fact that $C$ is a vector of 1/2-entries.

Next, we need to analyze $\widetilde{\mathbb{E}}[G(x)] = \widetilde{\mathbb{E}} \sum_{i=1}^n (x_i - x_{i+1})^2$, which we compute as follows:

$$\begin{aligned}
&\sum_{i=1}^n \widetilde{\mathbb{E}}[x_i^2 + x_{i+1}^2 - 2x_i x_{i+1}] \\
=&\sum_{i=1}^n (1 - 2\widetilde{\mathbb{E}}[x_i x_{i+1}]) \\
=&\sum_{i=1}^n 1 - 2 \cdot \left(\frac{1}{4} + \frac{1}{4}X_{i,i+1}\right) \\
=&\sum_i \frac{1}{2} - \frac{1}{2}X_{i,i+1} \\
=&\sum_i \frac{1}{2} - \frac{1}{2}(v_i v_{i+1} + w_i w_{i+1}) \\
=&\sum_i \frac{1}{4}((v_i - v_{i+1})^2 + (w_i - w_{i+1})^2) \\
=&\sum_i \frac{1}{4} \cdot |u_i - u_{i+1}|^2 \\
=&\sum_i \frac{1}{4}|\omega^{ik} - \omega^{(i+1)k}|^2 \\
=&\frac{1}{4} \cdot n \cdot |1 - \omega^k|^2.
\end{aligned}$$

where we have used that $v_i^2 = w_i^2 = 1$ for all $i$. Note that $|1-\omega^k|$ is the hypotenuse of a right triangle with edges $\Theta(1/n)$ and $2 - \Theta(1/n^2)$. The hypotenuse is $\sqrt{4 - O(1/n^2) + O(1/n^2)} \geq 2 - O(1/n^2)$. Thus we get $\widetilde{\mathbb{E}}[G(x)] \geq n \cdot (1 - O(1/n^2))$, as desired.

Alternatively, we could have noted that $|1-\omega^k|^2 = (1-\omega^k)(1-\omega^{-k}) = (2-\omega^k-\omega^{-k}) = 2\cdot(1-\cos(\omega^k)) = 2\cdot(2-\Theta(1/n^2))$. □

As an example, we can show that degree 6 SoS certifieds the bound $1-1/n$ on the max-cut of an odd cycle. Interestingly, even for degree 4 SoS, we don't have any idea whether this $1-\epsilon$ thing is tight. It turns out that this problem is very tightly related to problems such as the unique games conjecture.

Regarding the 0.878 GW guarantee: people thought it wasn't tight in general, and even for that algorithm. It took almost a decade for people to prove that 0.878 is tight for degree-2 SoS. The graph is not the odd cycle, but is called the F.S. graph (is in the Barak-Steurer notes). This is a very elegant construction: take the unit sphere in $d$ dimensions. The points in the sphere are the vertices, and there is an edge from $v \sim u$ if $\|v-u\| \le \epsilon$. You discretize the sphere, so take $n = 2^d$. (Alternatively, we can even think of having an infinite edge set with infinitely many edges and vertices.) We then use the discretized vectors to construct a PSD matrix and thus build degree-2 pseudomoments. We also need to determine the max-cut in the graph, where we can rely on isoperimetry. Roughly speaking a cut in the graph should have a geometric meaning (i.e., split unit sphere in half), so the crossing edges relate to the surface area of the cut. It turns out that this story extends beyond max-cut, and can be told for any CSP.

Interesting twist: degree-4 SoS solves the F.S. instances! It's not clear if the F.S. instance is hard (depends on geometry of sphere), but turns out that you can solve it using degree 4 SoS. Nobody has any idea what hard instances for max-cut actually look like! (Note that the type of graphs you get from NP-hardness reductions are indeed hard for max-cut, which we will discuss later.)

## 7.2 SoS for NP-hard problems

We begin with the 3-XOR problem. Recall that the 3-XOR problem is defined as follows: the input is a formula $\varphi$ on $n$ variables and $m$ clauses. Each clause has the form $x_i \oplus x_j \oplus x_k = b$ for some $b \in \{0,1\}$. The MAX-3XOR problem is defined as follows: given $\varphi$, find $x$ which maximizes the number of satisfied clauses. We saw several lectures ago that you can phrase 3-XOR as a natural polynomial optimization problem: the constraints can be rephrased as $x_i x_j x_k = a_{ijk} \in \{\pm 1\}$, and thus the max-variant is: $\max \sum_{(i,j,k)\in C_\varphi} x_i x_j x_k \cdot a_{ijk}$.

How hard is MAX-3XOR: Hästad's approximation result says it's NP-hard to distinguish the following two possibilites:

- "Yes" case: $\varphi$ is $(1-\epsilon)$-satisfiable.
- "No" case: $\varphi$ is at most $(1/2+\epsilon)$-satisfiable.

Note that every 3-XOR instance is 1/2-satisfiable (take a random solution). Furthermore, detecting full satisfiable is trivial, since you can solve a linear system over $\mathbb{F}_2$ with Gaussian elimination. Thus, 1-satisfiability versus anything is doable. So, the above NP-hardness result is essentially the best you can hope for.

What should we expect SoS to do if we give it a hard instance that is at most $(1/2+\epsilon)$-satisfiable? To prove a lower bound for SoS, we want to show that SoS thinks it is very satisfiable.

**Theorem 7.2.** *For all $\epsilon < 0$ and large enough $n$, there is a formula $\varphi$ on $n$ variables so that:*

1. *$\varphi$ is at most $(1/2+\epsilon)$-satisfiable.*

2. *There is a $\widetilde{\mathbb{E}}$ of degree $\Omega(n)$ so that:*

$$\widetilde{\mathbb{E}} \vDash x_i^2 = 1, \quad x_i x_j x_k = a_{ijk} \ \forall (i,j,k) \in \varphi.$$

In particular, for the $\widetilde{\mathbb{E}}$ in the above theorem statement, we have $\widetilde{\mathbb{E}} \sum_{ijk\in\varphi} x_i x_j x_k a_{ijk} = \sum a_{ijk}^2 = m$.

Some philosophical points: why are we bothering to prove this lower bound? SoS hardness is *unconditional* in contrast to NP-hardness. The above theorem confirms our belief. It is a bit disappointing that SoS seems to be doing worse than Gaussian elimination, which *can* tell the difference between a perfectly and non-perfectly satisfiable instance. To some extent, this violates the belief that SoS is optimal for CSPs.

To re-establish that belief, we can try to argue that SoS is inherently "noise-robust". If you have a computational problem that undergoes a big qualitative change when you switch between exact and noisy versions, then since SoS's behavior does not change between exact and easy, you should not expected SoS to do well on the exact version of the problem.

Also, note that the above theorem says that it is not possible to sample from a distribution which matches, say, the first 3 moments of a pseudodistribution, the way we did for 2 moments. (Technically, to do this, we need to show that a random 3CSP instance has small value over the *sphere*, as opposed to just over the hypercube.)

To prove the above theorem, we need to introduce the following clause-variable graph: it is bipartite with the two sides representing $m$ clauses and $n$ variables with and edge between a clause and a variable if the variable belongs to the clause. 3XOR instances correspond to graphs which are 3-left regular since each clause (on the left) has 3 neighbors.

**Definition 7.1.** A bipartite graph $(L,R)$ is a $(t,\beta)$-*expander* if for all $S \subset L$, $|S| \le t$, $|\Gamma(S)| \ge \beta \cdot |S|$.

Here $\Gamma(\cdot)$ is the neighborhood operator. Think of $t$ as something like $0.1n$. The best $\beta$ we can hope for is 2. The easy counterexample (to getting something larger, like 3) is to take one clause, which will have 3 neighbors, then take one neighbor and look at another clause containing it, and that shows a set of 2 clauses with 5 neighbors, so definitely can't hope for better than $\beta \ge 5/2$. Can iterate this and get all the way down to 2.

**Proposition 7.3.** *A random bipartite 3-left regular graph with $m \gg n$ is $(\eta \cdot n, 2-\delta)$-expanding with $\delta$ an arbitrary constant and $\eta = \eta_\delta > 0$.*

The proof of Proposition 7.3 is standard Chernoff/union, so we don't prove it. Note that when we were talking about some situations earlier on in the semester (e.g., structured max-cut) when expansion made the problem easier. But here, expansion actually makes the problem harder.

Imagine we specify signs on the left-hand (clause) side uniformly at random:

**Lemma 7.4.** *For any $(L,R)$ with $m \gg n/\epsilon^2$ random signes $a_{ijk}$ lead to a $\le 1/2+\epsilon$-satisfiable instance $\varphi$.*

The proof is again by Chernoff + union: for every fixed setting of the variables, that assignment satisfies at most $1/2+\epsilon$ of the clauses with very high probability, and then do union bound over all $2^n$ settings of the variables.

So, we have shown: *for all $\epsilon, \delta$ there exist $\eta, C$ so that for $m = Cn$, there exists a $(\eta n, 2-\delta)$-expanding $\varphi$ which is $\le 1/2+\epsilon$-satisfiable.*

**Lemma 7.5.** *If $\varphi$ is $(\eta n, 1.7)$-expanding, then for all $a_{ijk}$ there exists a degree $\Omega(\eta n)$ $\widetilde{\mathbb{E}}$ satisfying*

$$\widetilde{\mathbb{E}} \vDash x_i^2 = 1, \quad x_i x_j x_k = a_{ijk}.$$

We first discuss some intuition. We need to describe a procedure that generates the moments. Note that describing the multilinear coefficients sufficient (by the multilinearity constraint $x_i^2 = 1$). Also we only need to focus on monomials (by linearity). The interesting question is what is forced by the constants $x_i x_j x_k = a_{ijk}$. Certainly we need $\widetilde{\mathbb{E}}[x_i x_j x_k] = a_{ijk}$ for $(i,j,k) \in \varphi$. By multiplying together polynomials that form constraints, we need things such as $\widetilde{\mathbb{E}}[x_i x_j x_k x_r x_s x_t] = a_{ijk} a_{rst}$. (Note that this corresponds to adding the linear constraints modulo 2.)

We also get things such as the following: for constraints $ijk, irs$, we have:

$$a_{ijk} a_{irs} = \widetilde{\mathbb{E}}[x_i x_j x_k x_i x_r x_s] = \widetilde{\mathbb{E}}[x_j x_k x_r x_s],$$

which shows a degree-4 constraint. So, constraints can collide in certain ways. The worst nightmare is that you can collide constraints so that everybody cancels, which would yield $\widetilde{\mathbb{E}} 1 = \prod a_{ijk}$, which won't be true with high probability.

*Proof.* Here's an algorithmic description of a procedure generating the pseudomoments:

1. Set $\widetilde{\mathbb{E}} 1 = 1$.
2. For all clauses $ijk \in \varphi$, set $\widetilde{\mathbb{E}} x_i x_j x_k = a_{ijk}$.
3. Repeat until impossible: choose $S, T \subset [n]$ so that $|S \Delta T| = \deg_{\{x_i^2 = 1\}} x^S x^T \leq \eta n/10$, and $\widetilde{\mathbb{E}} x^S, \widetilde{\mathbb{E}} x^T$ were previously defined. If $\widetilde{\mathbb{E}}[x^S x^T]$ has already been set and is not equal to $\widetilde{\mathbb{E}}[x^S] \cdot \widetilde{\mathbb{E}}[x^T]$, then FAIL.

   Otherwise, set $\widetilde{\mathbb{E}}[x^S x^T] \leftarrow \widetilde{\mathbb{E}}[x^S] \cdot \widetilde{\mathbb{E}}[x^T]$.
4. Finally, for all $S$ with $|S| \leq \eta n/10$ with $\widetilde{\mathbb{E}} x^S$ not yet defined, $\widetilde{\mathbb{E}} x^S \leftarrow 0$. *(Why do we do this? The previous ones were forced upon us, but there is no reason to bias this in one direction or another. E.g., for a uniform instance, if we look at the subspace of satisfying assignments, then the uniform distribution over this subspace satisfies that each monomial is fixed or else is uniform over $\{\pm 1\}$.)*

**Lemma 7.6.** *If $\varphi$ is $(\eta n, 1.7)$-expanding, then ALG never FAILs.*

We set $d := \eta n/10$, which will be the degree that we will show our pseudoexpectation $\widetilde{\mathbb{E}}$ has. We now make the following definition:

**Definition 7.2.** For $|S| \leq d$, define a degree-$d$ derivation of $S$ to be a sequence $T_0, \ldots, T_t \subset [m]$ so that $T_0 = \emptyset$ and $x^S = \prod_{ijk \in T_t} x_i x_j x_k$, and for all $r \leq t$, $\deg_{\{x_i^2 = 1\}} \prod_{ijk \in T_r} x_i x_j x_k \leq d$, and for all $r \leq t$, there exists $a, b \leq r$ so that $T_r = T_a \Delta T_b$ or $T_r = T_a \Delta \{i,j,k\}$.

We say that an equation $x^S = a$ is $d$-derivable from $\varphi$ if there exist s a degree-$d$ derivation of $S$ so that $\prod_{\{ijk\} \in T_t} a_{ijk} = a$.

*Proof of Lemma 7.6.* Note that ALG only FAILs if there exists a $d$-derivation of $x^S = a$ and one of $x^S = -a$. If so, then there would be a $2d$-derivation of $1 = -1$ (by doing one derivation and then doing the second derivation starting from $x^S = a$). If we have such a derivation of $1 = -1$, then

let $T_1, \dots, T_t$ be this derivation. Then we have $\prod_{ijk \in T_t} x_i x_j x_k = 1$. Thus, the neighborhood of $T_t$ must tough each variable an even number of times, so each variable which is touched which must be touched at least twice. I.e., $T_t$ double-covers $\Gamma(T_t)$. Note that there are $3|T_t|$ edges leaving $T_t$. The number of vertices in this neighborhood is thus at most $|\Gamma(T_t)| \leq \frac{3}{2} \cdot |T_t|$. But we required that the graph is expanding, which must mean that $|T_t| > \eta n$ (as otherwise the expanding condition would tell us that $|\Gamma(T_t)| \geq 1.7 \cdot |T_t| > 1.5 \cdot |T_t|$.

Our goal is to show that if we have $|T_t| > \eta n$, then in the derivation, there must be some step at which we are looking at too many clauses to stay below the degree bound. In particular, let $r$ be the smallest value so that $|T_r| > 10d = \eta n$. Then $T_r$ came from $T_a, T_b$ with $a, b < r$, meaning that $|T_r| \leq 20d$ since $r$ is chosen as small as possible. But, by definition $T_0 \cdots T_t$ is a $2d$-derivation, so $\deg \prod_{ijk \in T_r} x_i x_j x_k \leq 2d$. So, $T_r$ undergoes a lot of cancellation!

Recall that there are $3|T_r|$ edges leaving $T_r$, and they touch at least $1.7|T_r|$ vertices on the right. Thus, they touch at least $(1.7 - 1.5) \cdot 2 \cdot |T_r| = 0.4|T_r|$ vertices exactly once. Hence $\deg \prod_{ijk \in T_r} x_i x_j x_k \geq 0.4|T_r| \geq 4d$, using that $|T_r| \geq 10d$. This contradicts the upper bound on the degree of $T_r$, which has to have degree at most $2d$ (since it's a $2d$-derivation). □

Finally, we must show that $\widetilde{\mathbb{E}}$ from the above algorithm is a pseudoexpectation of degree $\eta n/10$, which satisfies the necessary constraints:

**Lemma 7.7.** *$\widetilde{\mathbb{E}}$ from the above algorithm is a pseudoexpectation satisfying $\widetilde{\mathbb{E}} \vDash \{x_i^2 = 1, x_i x_j x_k = a_{ijk}\}$ of degree $\eta n/10$.*

*Proof.* $x_i^2 = 1$ are satisfied by construction, since we can just define $\widetilde{\mathbb{E}}$ of polynomials involving squares in the natural way.

To check that $\widetilde{\mathbb{E}}$ satisfies the constraints $x_i x_j x_k = a_{ijk}$, we only need to check that $\widetilde{\mathbb{E}}[x^S x_i x_j x_k] = a_{ijk} \cdot \widetilde{\mathbb{E}}[x^S]$ for all $|S| \leq d$. If the LHS is 0, then $\widetilde{\mathbb{E}}[x^S] = 0$, since otherwise we would have tried to set the value of $\widetilde{\mathbb{E}}[x^S x_i x_j x_k]$. Otherwise, the derivation worked (it didn't fail), and we assigned the value $a_{ijk} \cdot \widetilde{\mathbb{E}}[x^S]$ to $\widetilde{\mathbb{E}}[x_i x_j x_k \cdot x^S]$.

Positivity is the main nontrivial part here. We need to check that $\widetilde{\mathbb{E}}[p(x)^2] \geq 0$ for all polynomials $p$. We define an equivalence relation $\sim$ on $\{S \subset [n], |S| \leq d/2\}$, by $S \sim T$ if $\widetilde{\mathbb{E}}[x^S x^T] \neq 0$. To check that this is an equivalence relation:

- Reflexivity and symmetry are trivial, e.g., since $\widetilde{\mathbb{E}}[x^{2S}] = 1$.
- If $S \sim T, T \sim U$, then $\widetilde{\mathbb{E}}[x^S x^T]$ and $\widetilde{\mathbb{E}}[x^T x^U]$ are defined, and at some point in the procedure, since the products $x^S x^T, x^T x^U$ have degree at most $d$ and their product is $x^S x^U$, which has degree at most $d$, so we would have defined that accordingly (and the algorithm did not fail).

Thus, the equivalence relation $\sim$ partitions $\binom{[n]}{\leq d/2}$ into equivalence classes $C_1, \dots, C_N$. We will use these equivalence classes to construct an explicit diagonalization of the moment matrix. To do so, pick representaitives $S_1, \dots, S_N$ of the equivalence classes.

If $S, T \in C_i$, then $\widetilde{\mathbb{E}}[x^S x^T] = \widetilde{\mathbb{E}}[x^S x^{S_i}] \cdot \widetilde{\mathbb{E}}[x^T x^{S_i}]$ (the proof of this is the same as above; in particular, what we showed above is that the value of $\widetilde{\mathbb{E}}[x^S x^T]$ must be set to the value of $\widetilde{\mathbb{E}}[x^S x^{S_i}] \cdot \widetilde{\mathbb{E}}[x^T x^{S_i}]$). Let $p \in \mathbb{R}[x]_{\leq d/2}$. We can write $p = p_1 + \cdots + p_N$, where we write $p$ in the monomial basis, and define $p_i$ to be the sum of monomials from equivalence class $C_i$. Now, note that $\widetilde{\mathbb{E}}[p_i p_j] = 0$ by the definition of equivalence classes (since $p_i p_j$ is the product of things in different equivalence classes).

Then $\widetilde{\mathbb{E}}p^2 = \sum_i \widetilde{\mathbb{E}}p_i^2$. We now write

$$\widetilde{\mathbb{E}}p_i^2 = \sum_{S,T \in C_i} \hat{p}_i(S)\hat{p}_i(T) \cdot \widetilde{\mathbb{E}}[x^S x^T] = \sum_{S,T \in C_i} \hat{p}_i(S)\hat{p}_i(T)\widetilde{\mathbb{E}}[x^S x^{S_i}]\widetilde{\mathbb{E}}[x^T x^{S_i}] = \left(\sum_{S \in C_i} \hat{p}_i(S) \cdot \widetilde{\mathbb{E}}[x^S x^{S_i}]\right)^2 \geq 0,$$

as desired. □

Note that we have also shown that $\widetilde{\mathbb{E}}[p(x)^2 \cdot q_i(x)] = 0$ for all constraint polynomials $q_i$, since we have in fact shown that $\widetilde{\mathbb{E}}[p(x) \cdot q_i(x)] = 0$ for all polynomials $p$ (not just squares). The key is that we have only equality constraints here. □

What can we do with this? We can show that SoS doesn't solve 3SAT by reducing 3SAT to 3XOR and using the above result, and show that the reduction is low-degree so in particular SoS can't do well on 3SAT.

A few weeks ago, we showed that SoS does (strongly) refute 3XOR with $\gg n^{1.5}$ clauses. Today we were working in the regime with roughly $n$ clauses. Summarizing:

- With $m \asymp n$, we saw today that we need degree $\Omega(n)$.
- With $m \asymp n^{1.5}$, we say a few weeks ago that SoS of degree $O(1)$ refutes random 3XOR.
- There's a very nice tradeoff: e.g., for $m = n^{1.5-\epsilon}$ there are both upper and lower bounds: on the lower bound side, we get a weaker expanding condition (which was the main bottleneck in the argument today) and thus can show a degree $n^{2\epsilon}$ lower bound. There's also an upper bound using more advanced random matrix theory.

Note that randomness is key here: there isn't a short witness to the unsatisfiability! Turns out that a recent paper (by Max Hopkins et al) uses HDX's to construct an explicit instance where there is a short witness to the unsatisfiability.

# 8 November 4, 2022

Today we discuss average-case hardness. The idea here is that if we think SoS is "the" best algorithm, then by proving lower bounds for SoS on these instances, we can hope that we're proving lower bounds on the best algorithm. There are two types of problems for which we have average-case hardness: (1) random CSPs, and (2) planted clique.

Two distributions that matter for us: $G(n, 1/2)$ and $G(n, 1/2) + k$-clique, where the latter is gotten by taking $G(n, 1/2)$, picking a subset of $k$ vertices uniformly at random, and adding all edges between them.

A few different problems we consider:

1. Testing: given $G$ drawn from one of the two distributions above, decide which. Note that since the max clique in $G(n, 1/2)$ is of size $\approx 2 \log n$ with high probability, the two distributions are distinguishable for $k \geq 2.01 \log n$.
2. Search: given $G$ from the planted distribution, find the $k$-clique.

3. Given an arbitrary graph $G$, output CERTIFY or "?" so that if CERTIFY, then $G$ has no $k$-clique. Moreover, $\Pr_{G\sim G(n,1/2)}[\text{CERTIFY given } G] \geq 1 - o(1)$. You can't output CERTIFY for every graph since there are some graphs that have $k$-cliques (where we can't output CERTIFY).

   (Note that our certification algorithms for random CSPs were of this exact flavor.)

General theme here is: both search and refutation imply distinguishing algorithms. Generally, an algorithm for one wil llead to an algorithm for all 3. But there has been some recent work where you can separate these tasks.

There are many reductions here: given lower bounds for, e.g., the testing variant, we can often deduce lower bounds for the other ones.

Note that the $k$-planted clique problem is easy for $k \approx \sqrt{n}$. We saw one way of showing this in HW 1, but there's an easier way: the typical degree of vertices in $G(n, 1/2)$ is $n/2 \pm O(\sqrt{n})$. The degree of a vertex in the $k$-clique is $\geq n/2 - O(\sqrt{n}) + k$. So, if $k \gg \sqrt{n}$ (say $k \gg \sqrt{n \log n}$, to allow for a union bound), the clique vertices are the largest degree ones (look at vertex degrees). By considering the eigenvalues of the adjacency matrix, you can shave the $\sqrt{\log n}$, i.e., you can solve the distinguishing problem when $k \gg \sqrt{n}$.

Unlike in the CSP case, the brute force search algorithm is not that slow: it just looks for a clique of size $2.01 \cdot \log n$, and so, in time $n^{O(\log n)}$, you can solve the problem for any value of $k$ for which distinguishability holds. So, planted clique only gives evidence of hardness in superpolynomial time.

Given the brute force $n^{O(\log n)}$ time algorithm, we should hope that SoS does something in degree $O(\log n)$. It turns out that degree $O(\log n)$ SoS "solves" planted clique. It turns out that there's a tradeoff here (between $k \sim \sqrt{n}$ and $k \sim \log n$), which we discuss later.

## 8.1 SoS for planted clique

We will discuss SoS algorithms for the retutation algorithm (as for CSPs). Given $G$, we introduce the following system of constraints on $x_1, \ldots, x_n$:

$$\mathcal{C}_k = \left\{ x_i^2 = x_i; \quad x_i x_j = 0, \ i \not\sim j \in G; \quad \sum_i x_i = k \right\}.$$

How can we use the SoS algorithm to refute the presence of a $k$-clique? We can simply look for a degree-$d$ SoS refutation of these. This runs in time $n^{O(d)}$. What low bounds can we hope for:

- For $d = O(1)$, there is no degree $d$ refutation of $\mathcal{C}_k$ with high probability (i.e., an SoS lower bound for constant degree SoS). Here's what's true:

  **Theorem 8.1.** *With high probability over $G \sim G(n, 1/2)$ there exists a degree $d$ $\widetilde{\mathbb{E}} \vDash \mathcal{C}_{n^{1/2-\epsilon}}$ for $d = \Omega\left(\frac{\epsilon^2 \log n}{\log \log n}\right)$.*

  If a pseudoexpectation satisfies $\mathcal{C}_{n^{1/2-\epsilon}}$, then SoS refutations don't exist by duality. Note that a refutation of $\mathcal{C}_k$ will be a proof of the form:

$$-1 = \sum_i p_i(x) \cdot (x_i^2 - x_i) + \sum_{i \not\sim j} p_{ij} x_i x_j + q(x) \cdot (\sum_i x_i - k) + (SoS).$$

So, if the pseudoexpectation exists, when you evaluate it on the RHS, you get something non-negative, whereas the LHS is $-1$. So, we get SoS degree $\approx \log n$ (as high as we can hope for, up to $\log\log n$ factor) and can refute planted $k$-clique with $k \approx n^{0.49}$.

- There are variants of the refutation that we constructed in HW 1 to show that we can refute $k = \sqrt{n}/2^d$ in degree $d$.

## 8.2 Proving the case $d = 2$

We will show the following:

**Theorem 8.2.** *With high probability over $G \sim G(n, 1/2)$ there exists $\widetilde{\mathbb{E}} \vDash \{x_i^2 = x_i, x_i x_j = 0$ for $i \not\sim j\}$, and so that $\widetilde{\mathbb{E}} \sum_{i=1}^n x_i = \Omega(\sqrt{n})$.*

Note that this statement is a bit weaker than the one above since it doesn't have $\widetilde{\mathbb{E}}[p(x) \sum_i x_i] = k \cdot \widetilde{\mathbb{E}}[p(x)]$ for all $p$. It's a bit different from the max-cut lower bound since: (a) we don't have a single graph (we are given a random graph), and (b) we won't be able to analyze PSDness are directly. But it is pretty explicit, unlike Grigoriev's 3-XOR lower bound we saw last time.

We need maps from $G$ to $\widetilde{\mathbb{E}}[x_i], \widetilde{\mathbb{E}}[x_i x_j]$; the only non-multilinear polynomials we need to worry about are $\widetilde{\mathbb{E}}[x_i^2]$, which is forced to be $\widetilde{\mathbb{E}}[x_i]$. There are no other non-multilinear polynomials we have to worry about since we're in degree 2.

Everything in the construction will be forced on us: it will depend on a certain parameter, which is $\widetilde{\mathbb{E}} \sum_i x_i$, which will lead to PSDness when this parameter is $\Omega(\sqrt{n})$.

We are forced to choose the following: $\widetilde{\mathbb{E}}[x_i x_j] = 0$ for $i \not\sim j$. Next, we need to set $\widetilde{\mathbb{E}}[x_i]$ and $\widetilde{\mathbb{E}}[x_i x_j]$ for $i \sim j$.

Note that, by SoS cauchy Schwarz, we must have $\widetilde{\mathbb{E}} \sum_{i,j} x_i x_j = \widetilde{\mathbb{E}} (\sum_i x_i)^2 \geq \left(\widetilde{\mathbb{E}} \sum_i x_i\right)^2 = \Omega(n)$. There are $\Theta(n^2)$ terms in the summation on the LHS which are allowed to be nonzero (those that are edges), and we want those to be on the order of $1/n$ on average.

We have $\widetilde{\mathbb{E}} \succeq 0$ (i.e., $\widetilde{\mathbb{E}}$ satisfies the PSD condition) if and only if $\widetilde{\mathbb{E}}[(1, x)(1, x)^\top] \succeq 0$. As a rule of thumb, increasing off-diagonal elements of a matrix makes the matrix "less PSD". We can try to compensate for this by increasing the elements on the diagonal. But, on the diagonal, we have $\widetilde{\mathbb{E}}[x_i^2] = \widetilde{\mathbb{E}}[x_i]$, which we're constrained from making too large.

Below we give nthe formal proof.

*Proof.* We define $\widetilde{\mathbb{E}}[x_i] = k/n$ for all $i$, and $\widetilde{\mathbb{E}}[x_i x_j] = \lambda$ for $i \sim j$. (By the argument above, we will need $\lambda \geq k^2/n^2$.) We have now commited to the definition (for appropriate choice of $k, \lambda$).

For $k \geq \Omega(\sqrt{n})$, we satisfiy $\widetilde{\mathbb{E}} \sum_i x_i = \Omega(\sqrt{n})$, and have satisfied the constraints $x_i^2 = x_i$ and $x_i x_j = 0$, $i \not\sim j$, by definition. It only remains to check the PSDness condition. What does the matrix $\widetilde{\mathbb{E}}[(1, x)(1, x)^\top]$ look like?

$$\begin{pmatrix} 1 & k/n \cdot \mathbf{1}^\top \\ k/n \cdot \mathbf{1} & (k/n - \lambda) \cdot I + \lambda A_G \end{pmatrix},$$

where we have defined $A_G$ to have 1's on the diagonal, i.e., $(A_G)_{ii} = 1$. In particular, the matrix in lower right-hand corner has $k/n$ on diagonal and $\lambda \cdot A_G$ off-diagonal.

Let $J$ denote the all-1s matrix. Recall that $A_G = \frac{1}{2}J + \bar{A}_G$, where $\|\bar{A}_G\|_\sigma \le O(\sqrt{n})$ whp. We take this as a given for now; you can prove this up to a log factor using matrix Bernstein, and discuss an approach to remove the log later. We now have that

$$\begin{pmatrix} 1 & k/n \cdot \mathbf{1}^\top \\ k/n \cdot \mathbf{1} & (k/n - \lambda) \cdot I + \lambda A_G \end{pmatrix} \succeq \begin{pmatrix} 1 & kn \cdot \mathbf{1}^\top \\ k/n \cdot \mathbf{1} & (k/n - O(\lambda\sqrt{n})I) + \lambda/2 \cdot J \end{pmatrix},$$

since $O(\lambda\sqrt{n}) \cdot I + \lambda\bar{A}_G \succeq 0$, as all eigenvalues of $\lambda\bar{A}_G$ are at least $-\lambda\sqrt{n}$. This is where we use the high-probability statement (namely, the spectral bound).

Recall that a block matrix $\begin{pmatrix} C & B^\top \\ B & A \end{pmatrix} \succeq 0$ iff and only if $A - BC^{-1}B^\top \succeq 0$, which we can compute as follows:

$$(k/n - O(\lambda\sqrt{n})) \cdot I + \lambda/2 \cdot J - (k/n)^2 J.$$

Here is where the constraint $\lambda \ge (k/n)^2$ shows up! In particular, we need $\lambda/2 > (k/n)^2$ to make the coefficient on $J$ positive. Furthermore, to make the coefficient on the identity matrix positive, we need $k/n \gg \lambda\sqrt{n}$.

We want to maximize $k$ so that there exists $\lambda$ satisfying these. In particular, we want $k/(n\sqrt{n}) \gg \lambda \gg (k/n)^2$, i.e., we need $k/(n\sqrt{n}) \gg (k/n)^2$, and this holds as long as $k \ll \sqrt{n}$. So, choosing $k = c \cdot \sqrt{n}$ for sufficiently small $k$ means that the pseudoexpectation will be PSD, as we wanted to show. □

### 8.3 Lower bounds for degree 4

Things get trickier for degree 4. One issue comes from matrix concentration inequalities: we could use matrix Bernstein to analyze the eigenvalues of the adjacency matrix. What happens in degree 4? Let's look at the $\binom{n}{2} \times \binom{n}{2}$ moment matrix:

$$\begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix},$$

where we have shown degrees above. Suppose that $(i, j, k, \ell)$ is not a clique in $G$, e.g., since $(k, \ell) \notin E(G)$. Then $\widetilde{\mathbb{E}}[x_i x_j x_k x_\ell]$ must be 0 by our constraints. So, $\widetilde{\mathbb{E}}[x_i x_j x_k x_\ell]$ is nonzero only if $(i, j, k, \ell)$ is a clique in $G$. So, the indicator matrix of 4-cliques matters; in particular, the spectrum of that matrix matters.

Unfortunately, unlike for the adjacency matrix, that matrix is not a nice sum of independent random matrices (as we had for the adjacency matrix, which is a sum of symmetrix matrices, one for each edge). In particular, the element of the "4-clique" matrix corresponding to $(ij, k\ell)$ looks at, e.g., the edge $(i, j)$. So, we begin with a quick aside, namelt the *trace method* to bound the spectrum of random matrices.

**Trace method.** Let's start with the moment method. Given a scalar r.v. $X$, and we can bound $\mathbb{E}[X^k]$, then we can use these moments for concentration: in particular, $\Pr(X > t) = \Pr(X^k > t^k) \le \mathbb{E}[X^k]/t^k$. All of Bernstein, Chernoff bounds, etc., are just using the moment method here (they are using exponential moments with the right coefficient $\lambda$ in the exponent).

The trace method is a variant of the moment method for random matrices. Suppose $M$ is a random matrix. Note that $\mathbb{E}[\|M\|]$ is already an interesting statement: it's the expectation of a max over eigenvalues, and so it's saying we can do some sort of union bound over the eigenvalues. We need the following inequalities to do the trace moment method: if $M$ is a symmetrix random matrix,

$$\mathbb{E}\|M\| \le \left(\mathbb{E}\lambda_{\max}^{2k}\right)^{1/(2k)} \le \left(\mathbb{E}[\operatorname{Tr}(M^{2k})]\right)^{1/(2k)},$$

where the first inequality is Jensen and the second inequality follows from adding in the other eigenvalues. (This isn't so bad, since at most we lose a factor of $n^{1/(2k)}$, since we lose a factor of $n$ in passing from max eigenvalue fo trace.) In particular, if we take $k = \log n$, we only lose a constant factor. We can actually establish sharp constants by taking $k \propto 1/\epsilon$ for small $\epsilon$. The cost of taking $k$ larger is that $\operatorname{Tr}(M^{2k})$ is a large polynomial, and thus is harder to evaluate. Let's first use this method to reprove the upper bound on the spectrum of a random matrix with independent entries.

**Proposition 8.3.** *Let $M$ be a random matrix with independent random entries $M_{ij} = \{\pm 1\}$, and which is symmetric (i.e., $M_{ij} = M_{ji}$). Then $\mathbb{E}\|M\|_\sigma \le O(\sqrt{n \log n})$.*

To get some intuition, consider the case $k = 1$:

$$\mathbb{E}\operatorname{Tr}(M^2) = \sum_i \sum_j \mathbb{E}[M_{ij}M_{ji}] = \sum_{ij} \mathbb{E}[M_{ij}^2] = n^2,$$

which tells us that $\mathbb{E}\|M\|_\sigma \le n$. Now let's consider $k = 2$, recalling that the trace of a matrix power is the sum over all closed walks on the vertices:

$$\mathbb{E}\operatorname{Tr}(M^4) = \sum_{i,j,k,\ell} \mathbb{E}[M_{ij}M_{jk}M_{k\ell}M_{\ell i}].$$

Any term which has $i, j, k, \ell$ all distinct will be 0. There are $n^4$ terms, but for the remaining terms (i.e., those with a repeated index), there are at most $n^3$ of them. An example of a walk with 3 distinct indices is $i \to j \to k \to j \to i$. Here the monomial is $M_{ij}^2 M_{jk}^2$, which has expectation 1. Thus, $\mathbb{E}[\operatorname{Tr}(M^4)] = O(n^3)$, meaning that $\mathbb{E}\|M\| \le n^{3/4}$. Now we do the general case:

*Proof of Proposition 8.3.* We compute

$$\mathbb{E}[\operatorname{Tr}(M^{2k})] = \sum_{i_1,\dots,i_{2k}} \mathbb{E}\prod_j M_{i_j,i_{j+1}},$$

meaning that we have to consider labelings of the $2k$-cycle. We care about labelings that double-cover all edges. (In particular, if an edge appears an odd number of times, then the expectation of the monomial will be 0.)

If we don't care about the log factor, the following claim will suffice:

**Claim 8.4.** *Any double-covering labeling has at most $k + 1$ distinct labels.*

*Proof.* Take the labeling, and identify vertices with the same label. In the collapsed graph, there are at most $k$ edges (since each edge was double covered), and this graph is connected (since collapsing can't disconnect it). Any connected graph with $k$ edges can have at most $k + 1$ vertices. □

Now let's walk around a cycle with $k+1$ distinct labels and countthe number of ways we can do this: $k+1$ times, we can name a new vertex, for $n^{k+1}$ total possibilities. The remaining times, we have to name a vertex we've already encountered, for $\leq (2k)^{k-1}$ possibilities (this term can actually be improved if you want to get rid of the log). Further, there should be a term which is $2^{2k}$ which tells us whether or not each new vertex is a fresh vertex. Thus,

$$\mathbb{E}\|M\| \leq (2^{2k} \cdot n^{k+1}(2k)^{k-1})^{1/(2k)} \leq \sqrt{n} \cdot n^{1/(2k)} \cdot (2k)^{1/2}.$$

Choosing $k = 2\log n$ gives $\sqrt{n \log n}$. □

Now let $C_4$ be the $n^2 \times n^2$ matrix where $C_{ij,k\ell}$ is 1 if $ijk\ell$ is a clique in $G$, and 0 otherwise. It is useful to center this matrix first: to do so, we note that

$$\mathbb{E}_G C_{ij,k\ell} = 2^{-6},$$

since there are $\binom{4}{2} = 6$ independent edges here. Thus, we define $\bar{C}_4 = C_4 - \mathbb{E}[C_4] = C_4 - 2^{-6} \cdot J$. We now use the trace method to analyze $\mathbb{E}[\|\bar{C}_4\|]$. Choosing $k = 1$ gives a trivial bound, which is the side length of the matrix, namely $n^2$ (analogous to $k = 1$ above, which gave us an upper bound of $n$). So, we choose $k = 2$, which will give us some nontrivial bound (less than $n^2$): so, we consider:

$$\mathbb{E}\,\mathrm{Tr}(\bar{C}_4^4) = \sum_{i_1j_1,i_2j_2,i_3j_3,i_4j_4} \mathbb{E}[(\mathbb{1}\{i_1j_1i_2j_2\} - 2^{-6})(\mathbb{1}\{i_2j_2i_3j_3\} - 2^{-6})(\mathbb{1}\{i_3j_3i_4j_4\} - 2^{-6})(\mathbb{1}\{i_4j_4i_1j_1\} - 2^{-6})],$$

where, e.g., $\mathbb{1}\{i_1j_1, i_2j_2\}$ is the indicator that $i_1j_1i_2j_2$ is a clique. We just have to make sure not all of the terms contributes something, for most terms. In particular, we have that if $i_1, j_1, \ldots, i_4, j_4$ are all distinct, then the expectation is 0 (since the cliques $i_1j_1i_4j_4$ and $i_2j_2i_3j_3$ have disjoint edges: we can thus condition on all other edges, and note that these two cliques are conditionally indepenent, so take the conditional expectation inside the product – need to do this carefully). Thus, if not all distinct, there are at most $O(n^7)$ contributing terms. Thus, $\mathbb{E}[\|\bar{C}_4\|] \leq O(n^{7/4})$, which beats $O(n^2)$ (the trivial one).

Let's go back to planted clique: we choose:

$$\widetilde{\mathbb{E}}1 = 1, \quad \widetilde{\mathbb{E}}[x_i] = k/n, \widetilde{\mathbb{E}}[x_ix_j] = \begin{cases} \lambda_2 & : i \sim j \\ 0 & : \text{else.} \end{cases}$$

$$\widetilde{\mathbb{E}}[x_ix_jx_k] = \begin{cases} \lambda_3 & : ijk \text{ is triangle} \\ 0 & : \text{otherwise} \end{cases}$$

$$\widetilde{\mathbb{E}}[x_ix_jx_kx_\ell] = \begin{cases} \lambda_4 & : ijk\ell \text{ is 4-clique} \\ 0 & : \text{otherwise} \end{cases}.$$

The above definitions will ensure that all constraints are satisfied, so all we have to check is PSD-ness. In particular, we want to check that the matrix $\widetilde{\mathbb{E}}[(1, x, x \otimes x)(1, x, x \otimes x)^\top]$ is PSD.

For intuition, remember that $(\widetilde{\mathbb{E}} \sum_i x_i)^4 \leq \sum \widetilde{\mathbb{E}} x_ix_jx_kx_\ell$ means that the latter term has to be at least $(k/n)^4$. We won't verify PSD-ness in full, but we will focus on the matrix $\widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top]$. The diagonal of this matrix has: $\widetilde{\mathbb{E}}[x_i^2x_j^2] = \widetilde{\mathbb{E}}[x_ix_j] = \lambda_2$ if $i \sim j$, else 0. This looks suspicious, but we can be saved if the entire row + column corresponding to $(i, j)$ is 0. But, if $ij$ is not an edge,

then for all $k, \ell$, $ijk\ell$ is not a 4-clique, so we're saved. Thus, we can restrict attention to the rows and columns which are nonzero, i.e., those indexed by $(i,j) \in E(G)$. Let's decompose this matrix: it is

$$\lambda_2 \cdot I + \lambda_3 C_3 + \lambda_4 C_4,$$

where $C_3$ is a sparse matrix which indicates that $i,j,k,\ell$ is a 3-clique (so in particular, one of $i,j,k,\ell$ must repeat). We do the centering trick, rewriting the above as:

$$\lambda_2 I + \lambda_3 C_3 + \lambda_4 \cdot (2^{-6} J) + \lambda_4 \bar{C}_4 \succeq (\lambda_2 - O(\lambda_4 n^{7/4})) \cdot I + \lambda_3 C_3,$$

where we have used that $\mathbb{E}\|\bar{C}_4\| \leq O(\lambda_4 n^{7/4})$ and dropped the $\lambda_4 \cdot 2^{-6} \cdot J$. Now what does $C_3$ lok like? It is indexed by indices like $(ij, jk)$; for each repeated index $j$, there are $n \times n$ blocks (modulo overlaps between the blocks). The individual blocks are roughly $n \times n$ random matrices and are close to having independent random entries. Then if we center $C_3$, we get that the individual entries of the centered matrix have trace norm $\leq O(\sqrt{n})$.

Thus, $\lambda_3 C_3 = PSD + \bar{C}_3$, with $\|\bar{C}_3\| \leq O(\sqrt{n})$, meaning that the above display is:

$$\succeq \Big(\lambda_2 - O(\lambda_4 n^{7/4}) - O(\lambda_3 \sqrt{n})\Big) \cdot I.$$

Remember that $\lambda_s \approx (k/n)^s$ is forced on us, by the fact that we will have $\widetilde{\mathbb{E}}[\sum_i x_i] = k/n$ and the PSD constraints involving the degree-1 polynomials. Thus, to ensure thatthe above is PSD, we need:

$$k^2/n^2 \gg k^4/n^4 \cdot n^{7/4}, \qquad k^2/n^2 \gg k^3/n^3 \cdot \sqrt{n}.$$

Claim that if we can choose $k$ to satisfy the above, then can choose $\lambda$ to make the moment matrix block PSD, and then use a Schur complement argument to make the whole thing PSD. To satisfy the above it suffices to choose $k \ll n^{1/8}$. (This can actually be tightened.) This is true even though the largest clique in $G(n, 1/2)$ has size $O(\log n)$.

### 8.4 Improving the bounds

Interestingly, this construction actually breaks for larger $k$: it doesn't work to give lower bounds up to $k = \sqrt{n}$. There was a paper that was put out in 2013 to give a tight $k = \sqrt{n}$ lower bound for SoS for planted clique; people thought the problem was solved (wasn't that surprising since there were some Sherali-Adams lower bounds). That paper was flawed because of a subtle error in the random matrix theory.

Kelner's counterexample shows that the naive construction we presented above is *not* PSD for SoS degree 6 and $k = n^{1/3}$. We only proved the thing for degree 4 and $k \asymp n^{1/8}$ – turns out that the technique can be pushed up to $n^{1/3}$, but we won't go into that.

Let

$$r_i(x) := \sum_{j=1}^{n} (\mathbb{1}\{i \sim j\} - \frac{1}{2}) \cdot x_j.$$

Let $r_{ij} = \mathbb{1}\{i \sim j\} - \frac{1}{2}$. Thinking of the $x_j$ as clique indicators, this is a centered count of the number of clique elements adjacent to $i$. Let $P(x) = \sum_i r_i(x)$. Suppose $\widetilde{\mathbb{E}}$ satisfies all the constraints. Then

$$\widetilde{\mathbb{E}} P(x) = \widetilde{\mathbb{E}} \sum_i r_i(x)^4 \geq \widetilde{\mathbb{E}} \sum_i x_i^2 r_i(x)^4 = \widetilde{\mathbb{E}} \sum_i x_i r_i(x)^4 = \sum_i \sum_{j,k,\ell,s} r_{ij} r_{ik} r_{i\ell} r_{is} \widetilde{\mathbb{E}}[x_i x_j x_k x_\ell x_s].$$

In order for the pseudoexpectation to be nonzero, we need $ijk\ell s$ to be a 4-clique, but then each $r_{ij}, \dots, r_{is} = 1/2$ (i.e., is positive). Thus, the above is equal to:

$$\sum_{i,j,k,\ell,s} 2^{-4} \cdot \widetilde{\mathbb{E}}[x_i x_j x_k x_\ell x_s] = 2^{-4} \cdot \widetilde{\mathbb{E}} \left( \sum_i x_i \right)^5 \geq \Omega(k^5),$$

where the last step uses Cauchy-Schwarz for SoS.

But, the naive pseudoexpectation won't satisfy the above: it assigns $O((k/n)^s)$ to monomials of size $s$ which are cliques. In particular, for a fixed $i$, we have

$$\widetilde{\mathbb{E}} r_i(x)^4 = \sum_{jk\ell s} r_{ij} r_{ik} r_{i\ell} r_{is} \widetilde{\mathbb{E}}[x_j x_k x_\ell x_s].$$

Note that the randomness in $ij, ik, i\ell, is$ is independent of the randomness in the 4-clique $jk\ell s$.

There are $O(n^4)$ terms with 4 distinct indices $j, k, \ell, s$ and $O(n^3)$ terms with 3 distinct indices, all of which contribute 0 in expectation to the above.

If there are only two distinct indices, then we have $n^2$ terms in total, which contribute in total $(k^2/n^2) \cdot n^2 = k^2$. If there's only 1 distinct term, you get contribution on the order of $k$.

Thus, $\mathbb{E}_G \widetilde{\mathbb{E}} \sum_i r_i(x) \leq nk^2$. But we showed before that WHP, $\widetilde{\mathbb{E}} (\sum_i x_i)^5 \geq \Omega(k^5)$, so we need $nk^2 > k^5$, i.e., only if $k < n^{1/3}$.

What went wrong? $\widetilde{\mathbb{E}}[x_j x_k x_\ell x_s]$ didn't know about the randomness in edges connecting these vertices to $i$.

To understand this, it helps to think about what happened: there was some $p(G, x)$ so that

$$\widetilde{\mathbb{E}}_{naive,G}[p(G,x)] \neq \mathbb{E}_{(x,G) \sim planted}[p(G,x)]. \tag{18}$$

In particular, the latter one was roughly $k^5$; there was an SoS proof of $\mathcal{C}_6 \vdash p(G,x) \geq \Omega(\mathbb{E}_{(x,G)\sim planted}[p(G,x)])$.

The main point is to construct a polynomial $p(G, x)$ where (18) can never happen (namely, we have equality), and pray that the pseudoexpectation is PSD. We don't have general techniques for showing this, but it's very technical for cases we do, e.g., planted clique.

To define a PE, we need to defien a mapping from graphs to PEs. In particular, for each $S$, we have $\widetilde{\mathbb{E}}[x^S] : G \to \mathbb{R}$; we will write $\widetilde{\mathbb{E}}_G[x^S]$ to denote this pseudoexpectation. This has a Fourier expansion, and can be written as $\widetilde{\mathbb{E}}_G[x^S] = \sum_{\beta \subset \binom{n}{2}} \widehat{\widetilde{\mathbb{E}} x^S}(\beta) \cdot \chi_\beta(G)$. So we need to figure out how to set the fourier coefficients $\widehat{\widetilde{\mathbb{E}} x^S}(\beta)$.

Fixing $\beta$, if we take $p(G, x) = x^S \cdot \chi_\beta(G)$, then (18) tells us that $\widetilde{\mathbb{E}}_G[p(G,x)] = \widetilde{\mathbb{E}}_G[x^S \cdot \chi_\beta(G)] = \mathbb{E}_{(x,G)\sim planted}[x^S \cdot \chi_\beta(G)]$. Then it follows that

$$\widehat{\widetilde{\mathbb{E}}[x^S]}(\beta) = \mathbb{E}_{G \sim Unif}[\chi_\beta(G) \cdot \widetilde{\mathbb{E}}_G[x^S]] = \mathbb{E}_{G \sim Unif}[\mathbb{E}_{(x,G')\sim planted}[x^S \cdot \chi_\beta(G')]] = \mathbb{E}_{(x,G)\sim planted}[x^S \cdot \chi_\beta(G)].$$

Thus, we set the Fourier coefficients of low degree terms to get what they need to be, and then we define some truncated level. This will ensure that (18) never happens (i.e., we have equality there). You can also check that clique constraints are satisfied. It's very nontrivial to prove it's PSD. There's a heuristic: it only makes sense to chop off a Fourier series if its higher order terms are small; so, the pseudocalibration thing should work if the Fourier series has good tail decay.

For all SoS lower bounds we know (e.g., Grigoriev's one), they can be recovered as an instance of this with exactly the sort of tail decay of the Fourier series that you want.

Pseudocalibration is a general way of taking a planted distribution and constructing PE's; moreover, it indicates that Fourier decay corresponding to a planted distribution corresponds to whether this works.

# 9 November 18, 2022

Idea today and next lecture: can we match the provable guarantees of SoS but avoid black-box appeal to SDP solvers, and in particular get nearly-linear running time? Can't get these results generically, i.e., have to use much more problem-specific analysis.

## 9.1 Planted sparse vector

**Definition 9.1** (Sparse vector in a subspace)**.** Given as input a subset $V \subset \mathbb{R}^n$ of dimension $d < n$, the goal is to find a $k$-sparse vector in $V$, or decide that none exists.

Why do we care about this problem? If you're doing regression, a regression vector with few coordinates means that you're explaining the labels with only a few coordinates of the feature vector. Another problem is the small-set expansion problem in graphs: given a graph, you want to find a set of vertices of small size so that most of the edges touching that set stays inside the set, or to certify that no such set exists. This latter problem can be transformed into the sparse-vector problem since a set that expands poorly translates into a sparse vector that lies in the span of the top eigenvectors of the adjacency matrix of the graph.

Now we describe the planted sparse vector problem: we describe two distributions over instances of the sparse vector problem.

**Definition 9.2** (Planted sparse vector)**.** In one instance, we assume we're given a matrix $P \in \mathbb{R}^{n \times d}$ whose individual entries are iid $\mathcal{N}(0,1)$ (thus its columns give a random $d$-dimensional subspace by rotational invariance).

In the other instance, we take a matrix $Q = \widetilde{Q} \cdot O$ where $O$ is a random orthogonal matrix and all columns of $\widetilde{Q} \in \mathbb{R}^{n \times d}$ are $\mathcal{N}(0,1)$ except the first column, for which every entry is: $\pm\sqrt{n/k}$ with probability $k/n$, and 0 with probability $1 - k/n$. (Note that $\mathbb{E}[\widetilde{Q}_{11}^2] = k/n \cdot n/k = 1$.)

Goals that we associate to an average-case problem:

- Distinguish between $P, Q$.
- Search: given a sample from $Q$, find a sparse vector.
- Refute: given a sample from $P$, find a certificate that it has no sparse vectors.

One thing we have to check that is $P$ has no sparse vectors (e.g., if $d = n$ and $P$ describes all of $\mathbb{R}^n$, then it has sparse vectors). We will think of $k \approx \epsilon n$, where $\epsilon > 0$ is a small constant. Turns out that as long as $d = \Omega(n)$, it holds that $P$ contains no $k$-sparse vector with high probability. (You can verify by using an $\epsilon$-net.)

As $k$ gets larger, this problem gets harder. Further, as $d$ gets larger, the problem gets harder.

We will solve this problem with $k = \epsilon n$, and $d \approx \sqrt{n}$; first with a SoS algorithm, then with a spectral algorithm. There is actually an algorithmic technique (not based on SoS, but rather LLL lattice algorithm) which solves this problem for larger values of $d$. It is believed that SoS

algorithms don't match these guarantees – does this challenge the conjectured supremacy of SoS? The explanation is that those algorithms for larger $d$ are very brittle, whereas the SoS algorithms we will see today are robust: they can handle the case where the subspacee $Q$ only has an approximately sparse vector. Algorithms based on algebraic techniques can't do this.

Before giving a SoS algorithm for this problem, we need to take a detour and talk about sparsity.

## 9.2 Detour: sparsity

Given $\|x\|_0$ to denote the number of nonzero coordinates of $x$: this is not a continuous function of $x$. The traditional way of doing this is to relax 0-sparsity to $\ell_1$-sparsity, namely $\|x\|_1 = \sum_i |x_i|$.

In particular, given a $k$-sparse unit vector $x \in \mathbb{R}^n$, and all its nonzero entries are roughly equal, then $\|x\|_1 = k \cdot 1/\sqrt{k} \approx \sqrt{k}$. On the other hand, if $x$ is dense with roughly equal entries, then $\|x\|_1 \approx \sqrt{n}$. It is classic to use this by minimizing $\ell_1$ norm (which is a convex program, in fact a linear program): classic applications are compressed sensing and sparse regression. Both of these techniques boil down to solving $\min_{x:Ax=b} \|x\|_1$.

What we would want to do is the following: perhaps $\min_{x \in V,\ \|x\|_2=1} \|x\|_1$; the problem here is that $V$ is not affine, so we need the constraint $\|x\|_2 = 1$, which breaks convexity. We will look at $\ell_p$ norms for $p > 2$, e.g., $p = 4$. The key point is that the $\ell_4$ norm is large for sparse vectors. For a dense unit vector, we have $\|x\|_4^4 \approx /n^2 = 1/n$, and for a sparse unit vector $x$, we have $\|x\|_4^4 \approx 1/k$. The idea is to relax sparsity to this analytical notion of sparsity.

The problem we consider is the following one:

$$\max_x \|x\|_4^4 \quad \text{s.t.} \quad \|x\|_2^2 = 1, \quad V^\perp x = 0.$$

The constraint $V^\perp x = 0$ expresses the constraint that $x \in V$. This is not a convex progrma but is a nice polynomial optimization problem.

We now do a quick change of variables: letting $V \in \mathbb{R}^{n \times d}$ denote the matrix, we want to solve:

$$\max \|Vy\|_4^4, \quad \text{s.t.} \quad \|y\|_2^2 = 1.$$

This isn't exactly the same problem since we aren't insisting that $\|Vy\|_2 = 1$ and $V$ isn't necessarily an orthonormal basis (we are insisting $\|y\|_2 = 1$), but since $V$ is random it is close enough to orthonormal that this works.

Our goal is to consider an SoS relaxation and find a certifiable upper bound on the above problem; 2 questions:

- How large is this value for $V \sim Q$?
- What upper bound can SoS certify, if $V \sim P$? (hopefully we can make this small since $P$ does not have a sparse vector).

To answer the first question, let's choose $y$ so that $Oy = e_1 \in \mathbb{R}^d$ and $\|y\|_2 = 1$; let's call it $y^\star$. Then

$$\mathbb{E}\|Vy^\star\|_4^4 = \mathbb{E}\|\widetilde{Q} \cdot e_1\|_4^4 = n \cdot k/n \cdot (n/k)^2 = n^2/k,$$

which is $n/\epsilon$ for $k = \epsilon n$.

Next thing we can hope is that if $V \sim P$ then the upper bound is certifably less than $n/\epsilon$:

**Lemma 9.1.** *If $V \sim P$ and $d \ll \frac{\sqrt{n}}{\log^{O(1)} n}$, then with high probability (over $V$):*

$$\{\|y\|_2^2 = 1\} \sststile{4}{} \|Vy\|_4^4 \leq O(n).$$

To prove the above lemma, we need one more lemma:

**Lemma 9.2.** *If $d \ll \sqrt{n}/\log^{O(1)}(n)$ and $a_1, \ldots, a_n \sim \mathcal{N}(0, I_d)$ are random iid Gaussian vectors in $\mathbb{R}^d$, then with high probability,*

$$0.9\mathbb{E}_a(a \otimes a)(a \otimes a)^\top \preceq \frac{1}{n}\sum_{i=1}^n (a_i \otimes a_i)(a_i \otimes a_i)^\top \preceq 1.1 \cdot \mathbb{E}_a(a \otimes a)(a \otimes a)^\top.$$

We have seen similar matrices when clustering mixture models; we weren't that worried with how large $n$ had to be in order to ensure the eigenvalues of the expected moment matrix approximate those of the empirical matrix. The idea is that if a $d^2$-dimensional (full-rank) matrix is well-approximated by the empirical matrix, then the number of samples certainly has to be at least $n \geq d^2$; the above lemma says that this suffices, up to polylogarithmic factors.

We now prove the first lemma:

*Proof.* It holds that

$$\begin{aligned}
\|Vy\|_4^4 =& \sum_{i=1}^n \langle y, a_i\rangle^4 = (y \otimes y)^\top \left(\sum_{i=1}^n (a_i \otimes a_i)(a_i \otimes a_i)^\top\right)(y \otimes y) \\
=& 1.1n \cdot (y \otimes y)^\top \mathbb{E}[(a \otimes a)^\top (a \otimes a)](y \otimes y) - (y \otimes y)^\top \cdot M \cdot (y \otimes y) \\
=& 1.1n\mathbb{E}[\langle a, y\rangle^4] - (y \otimes y)^\top \cdot M \cdot (y \otimes y) \\
=& 1.1n \cdot 3 \cdot \|y\|_2^4 - (y \otimes y)^\top \cdot M \cdot (y \otimes y).
\end{aligned}$$

where $M$ is a PSD matrix. The point is that $(y \otimes y)^\top M \cdot (y \otimes y)$ is a SoS, so we get

$$\|Vy\|_4^4 = 3.3n \cdot \|y\|_2^4 - SoS(y) \leq_4 3.3n.$$

□

Now we prove Lemma 9.2. First, let's remind ourselves what Matrix Bernstein says: given $d \times d$ symmetric random matrices $X_1, \ldots, X_n$ with $\mathbb{E}X_i = 0$ and so that $\|X\| \leq R$ and $\|\mathbb{E}\sum_i X_i^2\| \leq \sigma^2$ (where $\|\cdot\|$ denotes spectral norm), then with high probability,

$$\left\|\sum_{i=1}^n X_i\right\| \leq O((\log d) \cdot R + \sqrt{\log d} \cdot \sigma).$$

A few things get us in trouble: the maximum eigenvalue of $(a_i \otimes a_i)(a_i \otimes a_i)^\top$ is pretty big. In particular, if we look at the entries $(jj, kk)$ of this $d^2 \times d^2$ matrix, for any $j, k \in [d]$, it will always be $a_i(j)^2 a_i(k)^2 \geq 0$. We could try subtracting off the mean, but that won't be good enough, the squares:

$$((a_i \otimes a_i)(a_i \otimes a_i)^\top)^2 \approx \|a_i\|^4 \cdot (a_i \otimes a_i)(a_i \otimes a_i)^\top \tag{19}$$

also have some big eigenvalue in the $(jj, kk)$ direction. The good thing for us is that the upper bound has a large component in this "bad" direction as well. In particular, the matrix we want to analyze has a lot of expected value and variance in these $(jj, kk)$ directions.

*Proof of Lemma 9.2.* Let $M = \mathbb{E}[(a \otimes a)(a \otimes a)^\top]$ and $X_i = (M^\dagger)^{1/2} \cdot (a_i \otimes a_i)$. To get $X_i$ we are dampening the big directions of $M$.

Now, we have that:

$$\mathbb{E}[X_i X_i^\top] = \mathbb{E}[M^{-1/2}(a_i \otimes a_i)(a_i \otimes a_i)^\top M^{-1/2}] = M^{-1/2} M M^{-1/2} \preceq I,$$

by definition of pseudoinverse.

Our strategy is to do matrix Bernstein on $\sum_{i=1}^n X_i X_i^\top - \mathbb{E}[X_i X_i^\top]$. We need to compute the parameters $R$ and $\sigma^2$.

We begin by bounding $R$: if we had $\|M^{-1/2}\| \leq O(1)$, then we would be good. To show this, we need to show that all nonzero eigenvalues of $M$ are bounded away from 0. We claim that

$$M = 2 \cdot \Pi_{Sym} + \Phi\Phi^\top$$

where $Sym \subset \mathbb{R}^d$ is defined as $Sym := span\{x \otimes x \ : x \in \mathbb{R}^d\}$, where $\Phi \in \mathbb{R}^{d^2}$ is the flattening of $\Phi_{ij} = 0$ if $i \neq j$ and 1 if $i = j$.

Note that $M$ has a kernel since for any $\phi \in \mathbb{R}^{d^2}$ with $\phi_{ij} = -\phi_{ji}$ is in the kernel of both $\Pi_{Sym}$ and $\Phi\Phi^\top$.

We prove the above claim:

*Proof.* Certainly $Ker(M) \supseteq Sym^\perp$ (since if $v \in Sym^\perp$, then $v^\top M v = \mathbb{E}[\langle v, a \otimes a\rangle^2] = 0$), so it is enough to show that the quadratic form of both sides are equal for any vector $u \in Sym \subset \mathbb{R}^{d^2}$. For any such $u$, we have:

$$a^\top M u = \mathbb{E}\langle a \otimes a, u\rangle^2 = \mathbb{E}[a^\top U a] = \mathbb{E}\left(\sum_{i \leq d} a_i^2 U_{ii}\right)^2 = \sum_{i,j} \mathbb{E}[a_i^2 a_j^2 U_{ii} U_{jj}] = (\sum_i U_{ii})^2 + 2\sum_i U_{ii}^2.$$

where $U$ is $u$ arranged as a $d \times d$ matrix. and the third equality follows since we can rotate $a$ into the basis where $U$ is diagonal, so we can assume that $U$ is diagonal.

We claim the above is the same as the quadratic form as the matrix on the right. Since $u \in Sym$, we have that $\sum_i U_{ii}^2 = \|u\|^2$. Furthermore, $(\sum_i U_{ii}^2) = \langle u, \Phi\rangle^2 = u^\top \Phi\Phi^\top u$. □

Since the norm of $\Phi$ is $\sqrt{d}$, the least eigenvalue of $M$ is 2, meaning that $\|M^{-1/2}\| \leq 1$.

Now we can bound:

$$\|X_i X_i^\top - \mathbb{E}[X_i X_i^\top]\| \leq \|X_i\|_2^2 + 1 = \|M^{-1/2} a_i \otimes a_i\|_2^2 + 1 \leq \|a_i\|_2^4 + 1,$$

which is $O(d^2)$ with high probabiliity by concentration of Gaussians. Thus, with high probability,

$$\sum_i X_i X_i^\top - \mathbb{E}[X_i X_i^\top] = \sum_i (X_i X_i^\top - \mathbb{E} X_i X_i^{(t)}) \cdot \mathbb{1}\{\|X_i X_i^\top - \mathbb{E} X_i X_i^\top\| \leq O(d^2)\},$$

meaning that we can take $R = d^2$.

**Bounding $\sigma^2$.** Now we have to do the key step: bounding $\sigma^2$. We bound the following, writing $X = X_i$:

$$\begin{aligned} & n \cdot \mathbb{E}(X_i X_i^\top - \mathbb{E} X_i X_i^\top)^2 \cdot \mathbb{1}\{\|\cdot\| \le O(d^2)\} \\ \preceq & n \cdot (\mathbb{E}(XX^{(T)})^2 - (\mathbb{E} XX^\top)2) \\ \preceq & n \cdot \mathbb{E}[\|X\|^2 \cdot XX^\top]. \end{aligned}$$

We alsmot have what we want, modulo the 2-norm $\|X\|^2$; but since this is concentrated, we should get what we want.

for any unit vector $u \in \mathbb{R}^{d^2}$, we bound:

$$\begin{aligned} u^\top (n \cdot \mathbb{E}[\|X\|^2 XX^\top]) u = & n \cdot \mathbb{E}[\|X\|^2 \langle X, u\rangle^2] \\ \le & n \cdot (\mathbb{E}\|X\|^4)^{1/2} \cdot (\mathbb{E}\langle X, u\rangle^4)^{1/2} \\ \le & O(n) \cdot \mathbb{E}\|X\|^2 \cdot \mathbb{E}\langle X, u\rangle^2 \\ \le & O(n) \cdot d^2 \cdot 1 = O(nd^2). \end{aligned} \tag{20}$$

where in the last step we have used $\mathbb{E}\langle X, u\rangle^2 \le 1$ since $\mathbb{E}[XX^\top] \preceq I$ and $u$ is a unit vector, and that $\mathbb{E}\|X\|^2 \le O(d^2)$ as we showed above. Above we have also used the following fact: for any polynomial $f$ in $N$ variables, $\mathbb{E}_{g \sim \mathcal{N}(0,I)} f(g)^4 \le 2^{O(\deg f)} \cdot (\mathbb{E}[f(g)^2])^2$. (This is called $(2,4)$-*hypercontracitivty.*)

By Matrix Bernstein, we get that

$$\|\sum_i X_i X_i^\top - \mathbb{E} X_i X_i^\top\| \le \widetilde{O}\left(d^2 + d\sqrt{n}\right),$$

which is bounded above by $0.1n$ if $d^2 \log^{O(1)} d \ll n$.

Finally we just have to renormalize, recalling the definition of $X_i$: from the above we got that:

$$-0.1I \preceq \frac{1}{n} \sum_i (X_i X_i^\top - \mathbb{E} X_i X_i^\top) \preceq 0.1n.$$

We now multiply both side of this matrix by $M^{1/2}$, thus "re-expanding" the "big direction". Using implicitly that $M^{1/2} \cdot M^{-1/2}$ is identity on a subspace containing all symmetric vectors, this gives

$$-0.1 M^{1/2} \cdot I \cdot M^{1/2} \le \frac{1}{n} \sum_i (a_i \otimes a_i)(a_i \otimes a_i)^\top - M^{1/2} \cdot \mathbb{E}[M^{-1/2}(a \otimes a)(a \otimes a)^\top M^{-1/2}] M^{1/2} \preceq 0.1 M^{1/2} \cdot I \cdot M^{1/2}.$$

Adding $M$ to both sides gives

$$0.9M \preceq \frac{1}{n} \sum_i (a_i \otimes a_i)(a_i \otimes a_i)^\top \preceq 1.1M.$$

□

We now have a SoS algorithm to solve the decision problem. Our algorithm also solves the refutation problem: if we can get a certified UB on the 4-norm polynomial, then it certifies there's no sparse vector, and the above shows that the 4-norm quantity is small with high probability over a random subspace.

Finally, what about the search problem: we could try to solve the problem:

$$\max_{\widetilde{\mathbb{E}} \vDash \|y\|^2=1} \widetilde{\mathbb{E}}\|Vy\|_4^4.$$

We don't know how to directly read off a sparse vector from a PE satisfying the above, but instead can do a local search procedure. But why should we expect any such PE to contain information about the sparse coefficients? If $V$ has an $\epsilon n$-sparse vector, then the above program has value much larger than $n/\epsilon$. Letting $y^\star$ be the vector so that $Vy^\star$ is $\epsilon n$-sparse, we can decompose any $y$ as follows:

$$y = \langle y, y^\star\rangle y^\star + (y - \langle y, y^\star\rangle y^\star),$$

and using SoS triangle inequality, we have

$$\begin{aligned}\widetilde{\mathbb{E}}\|Vy\|_4^4 =& \widetilde{\mathbb{E}}\|V\langle y, y^\star\rangle y^\star + V(y - \langle y, y^\star y^\star)\|_4^4 \\ \leq& \widetilde{\mathbb{E}}[\langle y, y^\star\rangle^4] \cdot \|Vy^\star\|_4^4 + \widetilde{\mathbb{E}}\|V(y - \langle y, y^\star\rangle y^\star)\|_4^4.\end{aligned}$$

We know the RHS is large ($\gg n/\epsilon$), and the second term is $\leq O(n)$ since $y - \langle y, y^\star\rangle y^\star$ only picks out the dense vectors in $V$ so the above analysis applies on the $d-1$-dimensional subspace. Sicne $\|Vy^\star\|_4^4$ is the 4-norm of a sparse vector, it follows that $\widetilde{\mathbb{E}}\langle y, y^\star\rangle^4 \geq 1 - o(1)$.

### 9.3 How do we make the above algorithm fast?

How can we avoid solving a degreee-4 SD program?

Given $V$ which is a matrix of $(a_1, \ldots, a_n)$, one thing we can do is compute the eigenvalues of

$$\|M^{-1/2}\frac{1}{n}\sum_i (a_i \otimes a_i)(a_i \otimes a_i)^\top \cdot M^{-1/2}\|.$$

We showed above that the eigenvalues of the above are small. Now let's hit it with $y^\star$, which gives us that

$$(y^\star \otimes y^\star)^\top M^{-1/2} \sum_i (a_i \otimes a_i)(a_i \otimes a_i)^\top M^{-1/2}(y^\star \otimes y^\star) \approx \sum_i \langle y^\star, a_i\rangle^4 \gg 1/\epsilon,$$

where the final inequality follows since $Vy^\star$ is $\epsilon n$-sparse. Note that the first inequality is not completely trivial: the high-level idea is that $M$ is roughly the identity on the symmetric subspace (can formally write down $M^{-1/2}$ explicitly), and thus can push the $y^\star \otimes (y^\star)^\top$ inside the $M^{-1/2}$.

So, one thing we can do is simply compute the maximum eigenvalue of the above matrix.

But how do we actually compute eigenvalues? Perhaps we can just use the power method: given a matrix $A$, draw $x \sim \mathcal{N}(0, I)$ and compute $Ax$, $A^2x, \ldots, A^{\log n}x$, and $A^{\log n}x = \sum_i \lambda_i^{\log n}\langle v_i, x\rangle^2$, and this will pick out the maximum eigenvalue.

This algorithm requires maintaining a vector whose dimensions are the side length of your matrix, which is $d^2$ for us. So, the fundamental limit for us is $d^2$ memory, and we probably incur a factor of $n$ since we have $n$ vectors $a_i$; so, a limit seems to be $nd^2$.

Can we get $nd$ (which is the size of the input)?[3] It turns out that if we consider the $d \times d$ matrix

$$A = \frac{1}{n}\sum_i (\|a_i\|^2 - d)a_i a_i^\top \in \mathbb{R}^{d\times d},$$

[3] "Does anyone else see horses, or are those just dogs?"

then its maximum eigenvalue distinguishes the distributiosn $P, Q$ (which is super easy), and in fact you can recover the sparse vector from its maximum eigenvector.

Heuristically, we hope that:

$$M^{-1/2}\sum_i (a_i \otimes a_i)(a_i \otimes a_i^\top)M^{1/2} \approx \begin{cases} (y^\star \otimes y^\star)(y^\star \otimes y^\star)^\top + noise \text{ if } V \sim Q \\ noise \text{ if } V \sim P. \end{cases} \tag{21}$$

To make this ofrmal, we define:

**Definition 9.3** (Partial trace)**.** The partial trace is an operator $\mathbb{R}^{d\times d} \otimes \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ which simply sums up te matrices on the diagonal.

In particular, an element of $\mathbb{R}^{d\times d} \otimes \mathbb{R}^{d\otimes d}$ is a $d \times d$ block matrix, and if its blocks are $A_{ij}$, then it returns $\sum_i A_{ii}$.

Then the idea is the partial trace of the planted matrix is:

$$\mathrm{Tr}_{\mathbb{R}^d}(y^\star \otimes y^\star)(y^\star \otimes y^\star)^\top = \sum_{i\le d} y^\star(i)^2 \cdot y^\star(y^\star)^\top \approx y^\star(y^\star)^\top.$$

The main thing we have to do is to check that the noise doesn't mess us up.

Note that

$$\mathrm{Tr}_{\mathbb{R}^d} \frac{1}{n}\sum_i (a_i \otimes a_i)(a_i \otimes a_i)^\top = \frac{1}{n}\sum_i \|a_i\|^2 \cdot a_i a_i^\top,$$

which is almost the matrix $A$; except, we have a $-d$ term, which essentially has the purpose of just cancelling out the noise.

Note that we can compute $\lambda_{\max}(A)$ in $\widetilde{O}(nd)$ time. This is because we can compute $Ax = \sum_i (\|a_i\|^2 - d) \cdot a_i \langle a_i, x\rangle$ since each computation $\langle a_i, x\rangle$ is $O(d)$ time, and then we need to add up $n$ things of the form $a_i \cdot w_i$, which takes time $\widetilde{O}(nd)$. Thus, using the power method, we can find $\lambda_{\max}(A)$ in $\widetilde{O}(nd)$ time.

So, we simply need to show that the maximum eigenvalue of $A$ differs in the planted and random cases. Let's first consider the maximum eigenvalue if there's a sparse vector. We want to use the coefficients of the sparse vector as the test vector: by rotational invariance, we can assume $y^\star = e_1$ without loss of genereality: then

$$\mathbb{E}[e_1^\top(\sum_i (\|a_i\|^2 - d)a_i a_i^\top)e_1] = \sum_i \mathbb{E}[(\|a_i\|^2 - d)a_i(1)^2] = \sum_j \left[\sum_{j=1}^d a_i(j)^2 a_i(1)^2 - d \cdot a_i(1)^2\right]$$
$$= n \cdot (d - 1 - d + \mathbb{E}[a_i(1)^4])$$
$$= n \cdot (-1 + k/n \cdot n^2/k^2) = n(n/k - 1) \approx n/\epsilon.$$

Note that the big matrix was solving the refutation problem: if its max eigenvalue is small, that subspace definitely doesn't contain a sparse vector. Here, we've only showed that in expectation (and with high probability), the vector $A$ has a large eigenvalue. But, if the max eigenvalue of this matrix $A$ is small, it won't necessarily tell you that there is no sparse vector. Thus, to do refutation, the only way we can do so is via SoS. (We can do search + decision using fast algorithms).

Finally, we deal with the random case: we assume that $\mathbb{E}[A] = 0$ for simplicity. Then we compute that, with high probability,

$$\sum_i (\|a_i\|^2 - d) a_i a_i^\top = \sum_i (\|a_i\|^2 - d) a_i a_i^\top \cdot \mathbb{1}\{\|(\|a_i\|^2 - d) a_i a_i^\top\| \le \widetilde{O}(d^{1.5})\},$$

which ensures that $R \le \widetilde{O}(d^{1.5})$.

Now we want to bound the spectral norm of the squares:

$$\mathbb{E} \sum_i (\|a_i\|^2 - d)^2 \|a_i\|^2 a_i a_i^\top = n \cdot \mathbb{E}\|a\|^2 (\|a\|^2 - d) a a^\top.$$

Thus, for any test unit vector $u$, we have, using Cauchy-Schwarz and (2,4)-hypercontractivity of the Gaussian,

$$\begin{aligned}
& n \cdot \mathbb{E}\|a\|^2 (\|a\|^2 - d)^2 \langle u, a\rangle^2 \\
\le & n (\mathbb{E}\|a\|^4)^{1/2} (\mathbb{E}(\|a\|^2 - d)^4 \langle u, a\rangle^4)^{1/2} \\
\le & n \cdot (\mathbb{E}\|a\|^4)^{1/2} \cdot (\mathbb{E}(\|a\|^2 - d)^8) \cdot \mathbb{E}\langle u, a\rangle^8 \\
\le & n \cdot (\mathbb{E}\|a\|^2) \cdot (\mathbb{E}(\|a\|^2 - d)^2) \cdot (\mathbb{E}\langle u, a\rangle^2) \cdot O(1) \\
\le & n \cdot d \cdot d \cdot O(1),
\end{aligned}$$

thus giving that $\sigma^2 \le O(nd^2)$. Hence, whp, $\|A\|$, for $V \sim P$, has $\|A\| \le \widetilde{O}(d^{1.5} + \sqrt{n}d) \ll n/\epsilon$ if $d \ll \sqrt{n}/\log^{O(1)} n$.

Why SoS? It gives a principled way of constructing higher-degree moment matrices that gives us inspiration/ways to construct these $d$-dimensional matrices that have the desired properties.