Lecture 0: The Sum-of-Squares "proofs to algorithms" paradigm

In this introductory lecture, we will introduce the sum-of-squares (SoS) hierarchy and the "proofs to algorithms" paradigm. We'll cover the topics in this order:

1. Estimation and proofs of identifiability, brute force algorithms (example: robust mean estimation)
2. Polynomial optimization, sum-of-squares relaxations, and sum-of-squares proofs
3. The "proofs to algorithms" paradigm illustrated for robust mean estimation

Some bibliographic remarks will be deferred to the end.

1 Estimation

A focus of this course will be on *estimation* problems. At a high level, estimation is the following task: we have a distribution μ over pairs $(u, y) \in \mathbb{R}^d \times \mathbb{R}^N$. We are given the *observable* y from a sample $(u, y) \sim \mu$, and our goal is to estimate the unknown u, or to return a quantity which is close to $\arg\max_{u^*} \mu(u^*, y)$. Rather than giving a more formal definition, let's see an example:

Example 1.1 (Robust Mean Estimation)**.** Let D be a distribution over $\mathbb{R}^d$ with mean u and covariance $\Sigma \preceq \mathbb{1}$. Let $\varepsilon > 0$ be a real number. Our goal is to estimate the mean u from ε-corrupted samples: we observe $y_1, \dots, y_m \in \mathbb{R}^d$, a $(1-\varepsilon)$-fraction sampled iid from D and the remaining ε-fraction are arbitrary vectors in $\mathbb{R}^d$.

To cast robust mean estimation as an "estimation problem" as described above, the distribution "μ" can be defined by taking any distribution ν over bounded-covariance distributions D, then a sample (u, y) from μ is generated by sampling $D \sim \nu$, setting $u = \mathbf{E}_{a \sim D} a$, then independently sampling $a_1, \dots, a_m \sim D$ and setting $y = y_1, \dots, y_m$ to be any set of vectors in $\mathbb{R}^d$ with $y_i = a_i$ for $(1-\varepsilon)m$ indices $i \in [m]$.[1]

2 Identifiability

When is estimation possible? One concern is that there may be be more than one maximizer of $\max_{u^*} \mu(u^*, y)$ (as is the case, for example, in robust mean estimation when all of the samples are corrupted, $\varepsilon = 1$).

Definition 2.1. For a pair $(u, y) \sim \mu$, we say that *y identifies u up to error δ* if for any u' which occurs with y with probability $\mu(u', y) \approx \max_{u^*} \mu(u^*, y)$,[2] $\|u - u'\| \leqslant \delta$.

A *proof of identifiability* for an estimation problem establishes that (with high probability) for $(u, y) \sim \mu$, y (approximately) identifies u. Let's see an example for robust mean estimation.

Lemma 2.2. *Let D be a distribution over $\mathbb{R}^d$ with mean u and covariance $\preceq \mathbb{1}$. If m is sufficiently large then with high probability, for a set of ε-corrupted samples $y = y_1, \dots, y_m$, y identifies u up to error $O(\sqrt{\varepsilon})$.*

[1] To make this formal, the corruptions can also be chosen according to some distribution.

[2] The use of $\approx$ here is informal; we could replace it with a more formal condition like $\mu(u', y) \geqslant (1-\varepsilon) \arg\max_{u^*} \mu(u, y)$ at the cost of introducing an additional parameter ε.

Proof. We'll show that any set of vectors $z_1, \dots, z_m \in \mathbb{R}^d$ which agree with a $(1-\varepsilon)$-fraction of the observed y_i's and have bounded covariance have empirical mean $\overline{z} = \frac{1}{m}\sum_{i=1}^m z_i$ which is close to u. Such a set always exists (as witnessed by the uncorrupted samples), so this is enough to establish that y identifies u.

Now, suppose that $z_1, \dots, z_m \in \mathbb{R}^d$ are vectors such that $z_i = y_i$ for a $(1-\varepsilon)$ fraction of $i \in [m]$, and further $\Sigma_z = \mathbf{Cov}(z_1, \dots, z_m) \preceq \mathbb{1}$. Let $a_1, \dots, a_m$ be the uncorrupted samples from D, such that $a_i = y_i$ for a $(1-\varepsilon)$ fraction of $i \in [m]$. Let $\overline{a} = \frac{1}{m}\sum_{i=1}^m a_i$, and assume for simplicity that $\overline{a} \approx u$ and that the empirical covariance $\Sigma_a = \mathbf{Cov}(a_1, \dots, a_m) \preceq \mathbb{1}$ (this is the only place where we use the assumption that m is large; we take m is large enough so that the empirical mean and covariance will be close to D's true mean and covariance). Then

$$\begin{aligned}
\|\overline{a} - \overline{z}\|^2 &= \langle \overline{a} - \overline{z}, \overline{a} - \overline{z}\rangle \\
&= \frac{1}{m}\sum_{i=1}^m (1 - \mathbf{1}_{a_i = y_i}\mathbf{1}_{y_i = z_i})\langle a_i - z_i, \overline{a} - \overline{z}\rangle + \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{a_i = y_i}\mathbf{1}_{y_i = z_i}\langle a_i - z_i, \overline{a} - \overline{z}\rangle \\
&= \frac{1}{m}\sum_{i=1}^m (1 - \mathbf{1}_{a_i = y_i}\mathbf{1}_{y_i = z_i})\langle a_i - z_i, \overline{a} - \overline{z}\rangle \\
&\leqslant \sqrt{\left(\frac{1}{m}\sum_{i=1}^m (1 - \mathbf{1}_{a_i = y_i}\mathbf{1}_{y_i = z_i})\right)\left(\frac{1}{m}\sum_{i=1}^m \langle a_i - z_i, \overline{a} - \overline{z}\rangle^2\right)},
\end{aligned}$$

where the inequality was by Cauchy-Schwarz. Now, there are at most $2\varepsilon m$ indices $i \in [m]$ for which $a_i \neq y_i$ or $y_i \neq z_i$, so the first parenthesized term is at most 2ε. We'll bound the second term using our assumption that D has bounded covariance and that the z_i have bounded covariance. Using the shorthand $b = \overline{a} - \overline{z}$, we expand

$$\langle a_i - z_i, b\rangle = \langle a_i - z_i + b - b, b\rangle = \langle a_i - \overline{a}, b\rangle - \langle z_i - \overline{z}, b\rangle + \|b\|^2,$$

and now we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \langle a_i - z_i, \overline{a} - \overline{z}\rangle^2 &= \frac{1}{m}\sum_{i=1}^m \left(\langle a_i - \overline{a}, b\rangle - \langle z_i - \overline{z}, b\rangle + \|b\|^2\right)^2 \\
&\leqslant \frac{1}{m}\frac{10}{3}\sum_{i=1}^m \langle a_i - \overline{a}, b\rangle^2 + \langle z_i - \overline{z}, b\rangle^2 + \|b\|^4 \\
&= \frac{10}{3}\left(b^\top \Sigma_a b + b^\top \Sigma_z b + \|b\|^4\right) \\
&\leqslant \frac{10}{3}\left(2\|b\|^2 + \|b\|^4\right),
\end{aligned}$$

Where in the second step we have used that for real A, B, C, $(A+B)^2 \leqslant 2A^2 + 2B^2$ (since $2A^2 + 2B^2 = (A+B)^2 + (A-B)^2$, and by similar reasoning $(A+B+C)^2 \leqslant 2A^2 + 4B^2 + 4C^2$, which can be improved to a uniform $\frac{10}{3}$ by averaging over permutations of A, B, C), in the third step we have rearranged to obtain quadratic forms with the covariance matrices of the a_i and z_i, and in the final step we have used that $\Sigma_a, \Sigma_z \preceq \mathbb{1}$. So, picking back up where we left off above, we conclude that for $b = \overline{a} - \overline{z}$,

$$\|b\|^4 \leqslant 2\varepsilon \cdot \frac{10}{3}\left(2\|b\|^2 + \|b\|^4\right),$$

and rearranging we have that $\|\overline{a} - \overline{z}\| = \|b\| \leqslant O(\sqrt{\varepsilon})$, as desired. □

Algorithms. The proof above shows us that estimation is information-theoretically possible, but what about algorithmic efficiency? All we have learned is that, if we do an exhaustive search for a set of vectors $z_1, \dots, z_m$ with the properties above, we can estimate u. We would like efficient algorithms for this task.

The sum-of-squares algorithmic paradigm gives us a formulaic way to transform the above proof of identifiability into an efficient algorithm. We will write down a polynomial optimization program to search for this set of z_i's, and then we will replace it with a semidefinite programming relaxation (the *sum-of-squares semidefinite program*. Finally, we'll observe that each step in our proof was a *sum-of-squares proof*, from which it will automatically follow that our relaxation to the polynomial optimization problem is exact.

3 Polynomial optimization

We are now in the position of having a system of polynomial equations, to which we would like one solution. We can express this problem (and many others) as a polynomial feasibility problem.

Definition 3.1 (Polynomial feasibility/optimization program)**.** A *polynomial system* in variables $x_1, \dots, x_n$ is a set of polynomial equations $\{f_i(x) = 0\}_{i=1}^m$.[3] A *polynomial feasibility program* asks for an assignment to x in $\mathbb{R}^n$ which satisfies all equations in a polynomial system S (if one exists). A *polynomial optimization program* asks for an assignment x which maximizes the polynomial $h(x)$ subject to satisfying all equations in a polynomial system S.

Following our proof of identifiability, we can write down a polynomial system, a solution to which will solve the robust mean estimation problem:

Problem 3.2 (Polynomial feasibility program for robust mean estimation)**.** We define a polynomial system in the following variables: $Z_1, \dots, Z_m \in \mathbb{R}^d$ represent the vectors $z_1, \dots, z_m$; $W_1, \dots, W_m \in \mathbb{R}$ with W_i representing the indicator that $z_i = y_i$; $B \in \mathbb{R}^{d \times d}$ is a matrix of "slack" variables. We include the following polynomial constraints:

$$W_i^2 = W_i \quad \forall\, i \in [m] \tag{1}$$

$$\sum_{i=1}^m W_i = (1-\varepsilon)m \tag{2}$$

$$W_i(Z_i - y_i) = 0 \quad \forall\, i \in [m] \tag{3}$$

$$\overline{Z} = \frac{1}{m}\sum_{i=1}^m Z_i, \qquad \frac{1}{m}\sum_{i=1}^m (Z_i - \overline{Z})(Z_i - \overline{Z})^\top = \mathbb{1} - BB^\top. \tag{4}$$

The constraints from (1) enforce that the W_i are 0/1 valued. The constraints from (2) and (3) together ensure that $(1-\varepsilon)m$ of the Z_i are equal to the corresponding y_i. Finally, the constraint (4) ensures that the covariance matrix of the Z_i is bounded by $\mathbb{1}$.

In general, we do not have efficient algorithms for solving polynomial feasibility programs (in fact it is NP-hard as problems such as 3SAT can be encoded as a polynomial system). So, we will use a *relaxation*.

[3] In general inequalities may also be allowed, but we ignore them for simplicity.

4 Sum-of-Squares relaxations

We would like to solve for a solution x to a polynomial system S. The Sum-of-Squares relaxations are a family of convex relaxations for polynomial optimization programs. Rather than solving for a solution (or distribution over solutions) to a polynomial system in variables x, we will solve for a linear operator $\tilde{\mathbf{E}} : x^{\leqslant d} \to \mathbb{R}$ where $x^{\leqslant d}$ is the set of monomials of degree at most d in x.

Definition 4.1. A *degree-d pseudoexpectation operator* for a polynomial system $S = \{f_i(x) = 0\}$ is a linear operator $\tilde{\mathbf{E}} : x^{\leqslant d} \to \mathbb{R}$ which enjoys the following properties:

1. "Scaling:" $\tilde{\mathbf{E}}[1] = 1$.
2. "Non-negativity of squares:" $\tilde{\mathbf{E}}[p(x)^2] \geqslant 0$ whenever $\deg(p) \leqslant d/2$.
3. "Feasibility:" for all $i \in [m]$, $\tilde{\mathbf{E}}[f_i(x) \cdot p(x)] = 0$ whenever $\deg(pf_i) \leqslant d$.

A pseudoexpectation operator for a polynomial optimization problem "$\max_x h(x)$ s.t. $S(x)$" is one which maximizes the value $\tilde{\mathbf{E}}[h(x)]$ subject to the above constraints.

Notice that if a polynomial system S (with constraints of degree at most d) is feasible, then there exists a degree-d pseudoexpectation for S as witnessed by taking $\tilde{\mathbf{E}}$ to be the expectation over any distribution over solutions to S. If S is not feasible, then there may be or may not be a degree-d pseudoexpectation for S. In this way, the pseudoexpectation *relaxes* the notion of a distribution over solutions to S. And for well-conditioned polynomial systems [O'D17, RW17] there exists a time $(mn)^{O(d)}$ algorithm, based on semidefinite programming, for finding a pseudoexpectation operator for S, when such an operator exists.

What good is a relaxation? Why is having such a pseudoexpectation operator useful? If we had access to $\mathbf{E}_\mu x$ or $\mathbf{E}_\mu x x^\top$ for μ an *actual* distribution over solutions to our program, we might be able to recover some actual feasible solution x. But when d is small,[4] $\tilde{\mathbf{E}}$ is merely a *pseudo*distribution, and just because we have access to the object $\tilde{\mathbf{E}}[x]$ does not mean that we can find a solution.

In the following section we will show that the pseudoexpectation operator respects a class of proofs called sum-of-squares proofs. So, if our proof of identifiability of the statement "u^* satisfying the polynomial system S must be δ-close to u" is a degree-d sum-of-squares proof, then $\tilde{\mathbf{E}}[u]$ will be close to u.

Solving for pseudoexpectations with semidefinite programming. Here we will see how to solve for a pseudoexpectation operator using a semidefinite program (in the lecture we will likely not have time for this). Recall that a semidefinite program is an optimization problem over symmetric matrices, of the form

$$\begin{aligned} \max_{X \in \mathbb{R}^{N \times N}} \quad & \langle C, X \rangle \\ \text{s.t.} \quad & \langle X, A_i \rangle = b_i \qquad \forall i \in [M] \\ & X \succeq 0, \end{aligned} \tag{5}$$

[4] If the variable x takes values over a discrete domain A^n with $|A| = t$, then a degree-$O(tn)$ pseudodistribution corresponds to the expectation over an actual distribution of solutions x, since the indicator $\mathbf{1}_{x=a}$ can be written as a degree-n polynomial from which one can calculate $\tilde{\mathbf{E}}[\mathbf{1}_{x=a}]$. But we are interested in high-dimensional settings, where a runtime of $n^{O(n)}$ is unacceptable.

For matrices $C, A_1, \dots, A_M \in \mathbb{R}^{N\times N}$ and $b_1, \dots, b_M \in \mathbb{R}$. In words, we are optimizing a linear objective over the cone of positive semidefinite matrices subject to linear constraints. A well-conditioned program of this form can be solved using the Ellipsoid algorithm (see e.g. [LGS88]) in time $\mathrm{poly}(N, M)$, since we can implement a separation oracle for the positive semidefinite cone in time $O(N^3)$ by computing an eigendecomposition.

To implement the computation of $\tilde{\mathbf{E}} : x^{\leqslant d} \to \mathbb{R}$ for the system $S = \{f_i(x) = 0\}_{i=1}^m$ with variables $x \in \mathbb{R}^n$ as a semidefinite program, we take our variable matrix $X \in \mathbb{R}^{N\times N}$ for $N = 1 + [n] + \binom{[n]}{2} + \cdots \binom{[n]}{d/2}$; that is, the rows and columns of X are indexed by subsets of $[n]$ of size at most $d/2$. For $A, B \in [n]^{\leqslant d}$, we will set $\tilde{\mathbf{E}}[x^A x^B] = X_{A,B}$. So, to ensure that $\tilde{\mathbf{E}}$ is well-defined and also to ensure that the scaling property holds, we enforce the following linear constraints on X:

$$X_{\emptyset,\emptyset} = 1, \quad \text{and} \qquad X_{A,B} = X_{U,V} \quad \forall A, B, U, V \subset [n]^{\leqslant d} \; s.t. \; A \cup B = U \cup V,$$

where we are taking the union as multisets. In order to ensure that we meet the feasibility constraints of the system S, we also add linear constraints for each $f_i(x) = \sum_{U\in[n]^{\leqslant d}} (\hat{f}_i)_U x^U$. First, we expand our set S to a system S' which includes $\{x^\alpha f_i(x) = 0\}_{i\in[m],\alpha\subset[n]^{\leqslant d},\deg(f_i x^\alpha)\leqslant d}$; that is, we add to the polynomial system the (redundant) constraints that $f_i(x) \cdot x^\alpha = 0$ for any monomial x^α with $\deg f_i(x)x^\alpha \leqslant d$. Now, for each $g(x) = \sum_{U\in[n]^{\leqslant d}} \hat{g}_U x^U$ in this extended system, we let $M_g \in \mathbb{R}^{N\times N}$ be some matrix for which $\tilde{\mathbf{E}}[g(x)] = \langle X, M_g\rangle$,[5] we add the constraints

$$\langle X, M_g\rangle = 0 \quad \forall g \in S'.$$

Claim 4.2. Taking $\tilde{\mathbf{E}}[x^A x^B] = X_{A,B}$ for X a solution to the program above yields a valid degree-d pseudoexpectation for the system S.

Proof. By construction, linearity and scaling hold for the operator $\tilde{\mathbf{E}}$ defined in this manner. For feasibility, notice that for any polynomial $p(x)$ with $\deg(pf_j) \leqslant d$, we can write $p(x)f_j(x) = \sum_{\alpha\in[n]^{\leqslant d-\deg(f_j)}} \hat{p}_\alpha x^\alpha f_j(x)$, and now by linearity

$$\tilde{\mathbf{E}}[p(x)f_j(x)] = \sum_{\alpha\in[n]^{\leqslant d-\deg(f_j)}} \hat{p}_\alpha \tilde{\mathbf{E}}[x^\alpha f_j(x)] = \sum_{\alpha\in[n]^{\leqslant d-\deg(f_j)}} \hat{p}_\alpha \langle X, M_{x^\alpha f_j}\rangle = 0,$$

since the constraint $x^\alpha f_j(x) = 0$ is in the system S'. To see that non-negativity of squares holds, consider any polynomial $p(x) = \sum_{A\in[n]^{\leqslant d/2}} \hat{p}_A x^A$ of degree at most $d/2$. Let $\hat{p} \in \mathbb{R}^N$ be the vector of p's coefficients. We can verify that

$$\tilde{\mathbf{E}}[p(x)^2] = \hat{p}^\top X \hat{p} \geqslant 0,$$

by the positive semidefiniteness of X. This completes the proof. □

5 Sum-of-Squares proofs

Definition 5.1 (Sum-of-Squares inequality)**.** We will say that a polynomial inequality $f(x) \leqslant g(x)$ is a *degree-d sum-of-squares* inequality if one can write $f(x) + s(x) = g(x)$ for a sum of square polynomials $s(x) = \sum_{i=1}^k h_i^2(x)$ with $\deg(s) \leqslant d$. For a set $S = \{f_i(x) = 0\}$ of polynomial constraints, we say that it is

[5] For example, one can construct M_g by taking for each $U \in [n]^{\leqslant d}$, $(M_g)_{A,B} = (\hat{g})_{A\cup B}$ for the lexicographically first A, B such that $A \cup B = U$, and all other entries of G equal to 0.

a *degree-d sum-of-squares inequality modulo S* if one can write $f(x) + s(x) + \sum_{j=1}^{m} c_j(x)f_j(x) = g(x)$, with $\deg(s), \deg(c_j f_j) \leqslant d$. We write

$$S \vdash_d f(x) \leqslant g(x)$$

to denote that $f(x) \leqslant g(x)$ is a degree-d SoS inequality mod S.

Notice that if $\tilde{\mathbf{E}} : x^{\leqslant d} \to \mathbb{R}$ is a degree-d pseudoexpectation operator for S, then when $f(x) \leqslant g(x)$ is a degree-d SoS inequality mod S we automatically have that

$$\tilde{\mathbf{E}}[f(x)] = \tilde{\mathbf{E}}[g(x)] - \tilde{\mathbf{E}}\left[\sum_{i=1}^{k} h_i^2(x)\right] - \tilde{\mathbf{E}}\left[\sum_{j=1}^{m} c_j(x)f_j(x)\right] \leqslant \tilde{\mathbf{E}}[g(x)],$$

by linearity, feasibility, and by the non-negativity of squares. So pseudoexpectation operators respect sum-of-squares proofs. We will utilize this to our advantage.

It turns out that many oft-used inequalities are also sum-of-squares inequalities. For example, the Cauchy-Schwarz inequality is sum-of-squares:

Claim 5.2 (SoS Cauchy-Schwarz)**.** Let p, q be vectors with polynomial-valued entries of degree at most d. Then For any $\varepsilon > 0$,

$$\vdash_{2d} \langle p, q\rangle \leqslant \frac{\varepsilon}{2}\|p\|^2 + \frac{1}{2\varepsilon}\|q\|^2, \text{ and } \quad \vdash_{4d} \langle p, q\rangle^2 \leqslant \|p\|^2\|q\|^2.$$

Proof. We can write $\langle p, q\rangle + \frac{1}{2}\|\sqrt{\varepsilon}p - \frac{1}{\sqrt{\varepsilon}}q\|^2 = \frac{\varepsilon}{2}\|p\|^2 + \frac{1}{2\varepsilon}\|q\|^2$, and $\langle p, q\rangle^2 = \|p\|^2\|q\|^2 - \frac{1}{2}\left(\sum_{i,j}(q_i p_j - q_j p_i)^2\right)$. □

Remark 5.3. For any pseudoexpectation $\tilde{\mathbf{E}}$ we can conclude that $\tilde{\mathbf{E}}[\langle p, q\rangle] \leqslant \sqrt{\tilde{\mathbf{E}}[\|p\|^2]\tilde{\mathbf{E}}[\|q\|^2]}$ by applying the claim above and choosing $\varepsilon = \sqrt{\tilde{\mathbf{E}}[\|q\|^2]/\tilde{\mathbf{E}}[\|p\|^2]}$. One can obtain some Hölder's inequalities by applying the claim inductively and following a similar logic.

The fact that a matrix scales a vector by at most its operator norm is also sum-of-squares.

Claim 5.4 (SoS operator norm)**.** Let $x \in \mathbb{R}^n$, $M \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times k}$. Then

$$M = \lambda\mathbb{1} - BB^\top \vdash_d x^\top Mx \leqslant \lambda\|x\|^2,$$

for $d \geqslant \deg(x^\top Mx + x^\top BB^\top x)$.

Proof. Our axioms imply that $x^\top Mx = \lambda x^\top \mathbb{1}x - x^\top BB^\top x = \lambda\|x\|^2 - \|B^\top x\|^2$, which is a sum-of-squares proof that $x^\top Mx \leqslant \lambda\|x\|^2$. □

Armed with these facts, we are now ready to transform our proof of Lemma 2.2 into a sum-of-squares proof, and immediately obtain an algorithm!

6 SoS-izing our proof of identifiability

Claim 6.1. Any degree-6 pseudoexpectation $\tilde{\mathbf{E}}$ over variables $W_1, \dots, W_m$, $Z_1, \dots, Z_m$, and B satisfying the polynomial constraints (1)-(4) also satisfies $\|\tilde{\mathbf{E}}[\overline{Z}] - u\|^2 = O(\sqrt{\varepsilon})$ with high probability so long as m is taken sufficiently large.

Proof. Recall $\overline{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$, and define $\Sigma_Z = \frac{1}{m}\sum_{i=1}^m (Z_i - \overline{Z})(Z_i - \overline{Z})^\top$. Recall also that W_i is our variable which represents $\mathbf{1}_{Z_i = y_i}$. Let $a_1, \dots, a_m$ be the uncorrupted samples from D, such that $a_i = y_i$ for a $(1-\varepsilon)$ fraction of $i \in [m]$. Let $\overline{a} = \frac{1}{m}\sum_{i=1}^m a_i$, and as before we've assumed m is large enough so that $\overline{a} = u$ and also that $\Sigma_a = \mathbf{Cov}(a_1, \dots, a_m) \preceq \mathbb{1}$. We have that

$$\|\overline{a} - \overline{Z}\|^4 = \langle \overline{a} - \overline{Z}, \overline{a} - \overline{Z}\rangle^2 = \left(\frac{1}{m}\sum_{i=1}^m (1 - W_i \mathbf{1}_{a_i = y_i})\langle a_i - Z_i, \overline{a} - \overline{Z}\rangle + \frac{1}{m}\sum_{i=1}^m W_i \mathbf{1}_{a_i = y_i}\langle a_i - Z_i, \overline{a} - \overline{Z}\rangle\right)^2.$$

Since we have enforced the constraint $W_i(y_i - Z_i) = 0$ in (3), the second term is 0. So we have

$$= \left(\frac{1}{m}\sum_{i=1}^m (1 - W_i \mathbf{1}_{a_i = y_i})\langle a_i - Z_i, \overline{a} - \overline{z}\rangle\right)^2$$

Now, we apply the $\vdash \langle p, q\rangle^2 \leqslant \|p\|^2\|q\|^2$ version of degree-6 SoS Cauchy-Schwarz (Claim 5.2),

$$\leqslant \left(\frac{1}{m}\sum_{i=1}^m (1 - W_i \mathbf{1}_{a_i = y_i})^2\right)\left(\frac{1}{m}\sum_{i=1}^m \langle a_i - Z_i, \overline{a} - \overline{Z}\rangle^2\right)$$

If $A \leqslant B$ is an SoS inequality, then so is $As \leqslant Bs$ for any sum-of-squares s, since $Bs - As = (B - A)s$. So, we can bound the parenthesized terms one at a time. For the first term, notice that (1) $\vdash_2 (1 - W_i \mathbf{1}_{a_i = y_i})^2 = 1 - W_i \mathbf{1}_{a_i = y_i}$.[6] Also, (2), (1) $\vdash_1 \frac{1}{m}\sum_{i=1}^m (1 - W_i \mathbf{1}_{a_i = y_i}) \leqslant 2\varepsilon$.[7] So (1), (2) $\vdash \frac{1}{m}\sum_{i=1}^m (1 - W_i \mathbf{1}_{a_i = y_i})^2 \leqslant 2\varepsilon$.

We bound the second term almost exactly as in the proof of Lemma 2.2. Using the shorthand $b = \overline{a} - \overline{Z}$ and applying the same manipulations as previously,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \langle a_i - Z_i, \overline{a} - \overline{Z}\rangle^2 &= \frac{1}{m}\sum_{i=1}^m \left(\langle a_i - \overline{a}, b\rangle - \langle Z_i - \overline{Z}, b\rangle + \|b\|^2\right)^2 \\ &\leqslant \frac{1}{m}\frac{10}{3}\sum_{i=1}^m \langle a_i - \overline{a}, b\rangle^2 + \langle Z_i - \overline{Z}, b\rangle^2 + \|b\|^4,\end{aligned}$$

where as in the proof of Lemma 2.2 the inequality is a degree-4 SoS inequality. Proceeding,

$$= \frac{10}{3}\left(b^\top \Sigma_a b + b^\top \Sigma_Z b + \|b\|^4\right),$$

And applying Claim 5.4 we have that (4) $\vdash_4 b^\top \Sigma_Z b \leqslant \|b\|^2, b^\top \Sigma_a b \leqslant \|b\|^2$, so we conclude that

$$\leqslant \frac{10}{3}(2\|b\|^2 + \|b\|^4).$$

So, putting everything together, we conclude that $\tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^4] \leqslant O(\varepsilon) \cdot \tilde{\mathbf{E}}[2\|\overline{a} - \overline{Z}\|^2 + \|\overline{a} - \overline{Z}\|^4]$. Rearranging, we have $\tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^4] \leqslant O(\varepsilon) \cdot \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2]$, and because

$$0 \leqslant \tilde{\mathbf{E}}\left[\left(\|\overline{a} - \overline{Z}\|^2 - \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2]\right)^2\right] = \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^4] - \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2]^2 \leqslant \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2](O(\varepsilon) - \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2]),$$

we have that $O(\varepsilon) \geqslant \tilde{\mathbf{E}}[\|\overline{a} - \overline{Z}\|^2] \geqslant \|\overline{a} - \tilde{\mathbf{E}}[\overline{Z}]\|^2$ (one can verify that the last inequality is sum-of-squares). □

[6] Since $(1 - W_i \mathbf{1}_{a_i = y_i})^2 = 1 - 2W_i \mathbf{1}_{a_i = y_i} + W_i^2 \mathbf{1}_{a_i = y_i}^2 = 1 - W_i \mathbf{1}_{a_i = y_i}$.

[7] Since $1 - W_i \mathbf{1}_{a_i = y_i} = 1 - W_i + W_i \mathbf{1}_{a_i \neq y_i}$, (2) $\vdash_1 \sum_i W_i = m(1 - \varepsilon)$ and (1) $\vdash_2 W_i \mathbf{1}_{a_i \neq y_i} \leqslant \mathbf{1}_{a_i \neq y_i}$.

7 Conclusion

We have seen how a proof of identifiability which is captured by low-degree sum-of-squares proofs can automatically yield a polynomial time algorithm via sum-of-squares relaxations. This is the "sum-of-squares algorithmic paradigm" after which the course is named. The theme of proofs-to-algorithms will show up again and again throughout the course.

Many issues remain to be discussed. A semidefinite program in a number of variables which is linear in the input size is often prohibitively slow in practice; could there be more efficient implementations which build on these algorithms? What can be done in settings where there is no proof of identifiability? How does the power of sum-of-squares algorithms vary as a function of the degree d? These questions are all part of an active area of research; stay tuned for the best answers that science has mustered so far.

Bibliographic remarks. The problem of estimating the mean under adversarial corruptions goes back as far as the 1960's (e.g. [Ans60, Tuk60]). The first polynomial-time algorithm with dimension-independent error was given by Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart [DKK+19] (see also [LRV16]); their convex programming approach bears some similarity to the SoS program that we use here, but the analysis is more complicated. Since then there have been numerous works on this topic, including the time- and sample-efficient algorithms [DL19, DHL19, CDGS20]. See e.g. [Li18] for a more complete survey. The presentation in this lecture was based on the works of Hopkins-Li [HL18] and Kothari-Steinhardt-Steurer [KSS18], with invaluable advice from Sam B. Hopkins. Thanks also to Sam for suggesting robust mean estimation as a topic for the introductory lecture.

Sum-of-squares programming originated in several independent works by Lasserre [Las01], Nesterov [Nes00], Parrilo [Par00], and Shor [Sho87] near the end of the 20th century. The proofs-to-algorithms paradigm was popularized in the algorithms community starting with the work of Barak, Brandao, Harrow, Kelner, Steurer and Zhou [BBH+12] (see also [OZ13, BKS14, BKS15]). The proof of the SoS Cauchy-Schwarz inequality $\langle a, b\rangle^2 \leqslant \|a\|^2\|b\|^2$ is taken from Ma-Shi-Steurer [MSS16], Lemma A.1.

Thanks to Jay Mardia for helpful suggestions in improving the presentation of these notes.

Contact. Comments are welcome at tselil@stanford.edu.