Lecture 11: The sum-of-squares hierarchy and the proofs-to-algorithms paradigm

November 26, 2025

This lecture will serve largely as a self-contained (but very rushed) introduction to the very powerful sum-of-squares (SoS) hierarchy of semi-definite programs, with a particular emphasis on the types of SoS technology we will need to develop fast algorithms for robust statistics (and related tasks) using it. There are whole courses on SoS algorithms, see e.g. the excellent courses by Tselil Schramm and Sam Hopkins linked on the course website, so there will be unfortunately a lot we cannot cover!

The SoS hierarchy was first developed in several independent works by Shor [1] (not the quantum Shor!), Parrilo [2], Nesterov [3], Lasserre [4], and you will sometimes hear the hierarchy referred to by various subsets of these names. The power of this hierarchy for solving statistical estimation problems, as well as the the proofs-to-algorithms paradigm was first popularized in a sequence of works starting with [5], see also [6, 7, 8], and it is this interpretation of the hierarchy that we will be utilizing in this class.

1 Introduction

The goal of the SoS hierarchy is to efficiently find proofs of inequalities between multivariate polynomials, that is, given two degree k polynomials in n variables $p(x), q(x)$, how can I certify that $p(x) \geq q(x)$ for all $x \in \mathbb{R}^n$? In general, this is an NP-hard problem, but it turns out that if the proof has a specific form based on sum-of-squares of polynomials, then we can hope to efficiently find it:

Definition 1.1. Let $p(x), q(x)$ be degree k polynomials in variables $x \in \mathbb{R}^n$, and let k be an even number. We say that a polynomial $r(x)$ is a *sum-of-squares* (SoS) polynomial if it can be written as $r(x) = \sum_{\ell=1}^{L} s_\ell^2(x)$, for polynomials $s_1, \ldots, s_L$. We say that $p \geq q$ has a *degree-k sum-of-squares proof* if $p - q$ is a SoS polynomial of degree at most k.

Notice that crucially, even if p and q may be of degree less than k, the SoS proof is allowed to use degree k polynomials in the proof—this turns out to be a very powerful operation. We will also need a notion of constrained sum-of-squares proofs:

Definition 1.2. Let $\mathcal{A} = \{f_1 \geq 0, \ldots, f_m \geq 0\}$ be a system of polynomial constraints over n variables, i.e. each f_i is a polynomial over $\mathbb{R}^n$. We say that there is a *SoS proof* that $\mathcal{A}$ implies $p \geq 0$ if there exists sum-of-squares polynomials r_S for every $S \subseteq [m]$ so that

$$p = r_\varnothing(x) + \sum_{S \subseteq [m]} r_S \cdot \prod_{i \in S} f_i \ .$$

We say that this proof has degree k if each summand has degree at most k. In this case, we write that

$$\mathcal{A} \vdash_k \{p \geq 0\} \ .$$

When $\mathcal{A}$ is empty, we will also write $\vdash_k \{p \geq 0\}$.

It is easily verified that the existence of an SoS proof does immediately imply that any point satisfying $\mathcal{A}$ also satisfies $p(x) \geq 0$, but it is not clear that such a proof always exists. However, classic results (related to the resolution of Hilbert's 17th problem) show that they are, in fact, complete, but may require arbitrarily high degree:

Theorem 1.1 (Positivstellensatz, [9, 10])**.** *For every system of polynomial constraints $\mathcal{A}$, there either exists a solution to these constraints, or a SoS proof that $\mathcal{A} \vdash_k \{-1 \geq 0\}$ for some $\ell \in \mathbb{N}$.*

SoS proofs have a number of nice, natural properties:

Fact 1.2. *Let $\mathcal{A}, \mathcal{B}, \mathcal{C}$ be systems of polynomial constraints, and let f, g, h be polynomials. Then:*

- **convexity:** $\mathcal{A} \vdash_\ell \{f \geq 0, g \geq 0\}$ *implies that* $\mathcal{A} \vdash_\ell \{f + g \geq 0\}$.
- **multiplicativity:** $\mathcal{A} \vdash_\ell \{f \geq 0, g \geq 0\}$ *implies that* $\mathcal{A} \vdash_{\ell^2} \{fg \geq 0\}$.
- **transitivity:** $\mathcal{A} \vdash_\ell \mathcal{B}$ *and* $\mathcal{B} \vdash_{\ell'} \mathcal{C}$ *implies that* $\mathcal{A} \vdash_{\ell \cdot \ell'} \mathcal{C}$.

Many natural facts turn out to be provable in SoS. For instance, one can prove versions of Cauchy-Schwarz in SoS:

Fact 1.3 (SoS Cauchy-Schwarz)**.** *Let f, g be polynomials over $\mathbb{R}^d$ with degree at most k. Then, for any $\varepsilon > 0$, we have*

$$\vdash_{2k} \langle f, g \rangle \leq \frac{\varepsilon}{2} \|f\|_2^2 + \frac{1}{2\epsilon} \|g\|_2^2 \, ,$$

and

$$\vdash_{4k} \langle f, g \rangle^2 \leq \|f\|_2^2 \|g\|_2^2 \, .$$

Proof. Indeed, by directly checking, we can write

$$\frac{\varepsilon}{2} \|f\|_2^2 + \frac{1}{2\epsilon} \|g\|_2^2 - \langle f, g \rangle = \frac{1}{2} \left\| \sqrt{\varepsilon} f - \frac{1}{\sqrt{\varepsilon}} g \right\|_2^2 = \frac{1}{2} \sum_{i=1}^{d} \left(\sqrt{\varepsilon} f_i - \frac{1}{\sqrt{\varepsilon}} g_i \right)^2 \; , \; .$$

which is clearly a sum-of-squares polynomial of degree at most $2k$. Similarly, we have that

$$\|f\|_2^2 \|g\|_2^2 - \langle f, g \rangle^2 = \frac{1}{2} \sum_{i,j} (f_i g_j - f_j g_i)^2 \, .$$

□

2 Pseudo-distributions and pseudo-expectations

The basic algorithmic object involved in the SoS hierarchy is a notion of pseudo-distributions and pseudo-expectations. Let $\mathbb{R}[x]_{\leq k}$ denote the set of polynomials over formal variables x with degree at most k.

Definition 2.1. A *degree-k pseudo-expectation* $\widetilde{\mathbb{E}}$ is a linear operator from $\mathbb{R}[x]_{\leq k}$ to $\mathbb{R}$ that satisfies $\widetilde{\mathbb{E}}[1] = 1$, and $\widetilde{\mathbb{E}}[h^2] \geq 0$ for any polynomial $h \in \mathbb{R}[x]_{\leq k/2}$. Given any set of constraints $\mathcal{A} = \{f_1 \geq 0, \ldots, f_m \geq 0\}$, we say that $\widetilde{\mathbb{E}}$ *satisfies* $\mathcal{A}$ at degree k, denoted $\widetilde{\mathbb{E}} \models_\ell \mathcal{A}$, if $\widetilde{\mathbb{E}}[h^2 f_i] \geq 0$ for all $j \in 1, \ldots, m$ and all polynomials h satisfying $h^2 f_i$ has degree at most k.

We note that by standard tricks, this type of constraint also includes equality constraints (to encode $f = 0$, just include the constraints $f \geq 0$ and $-f \geq 0$).

Intuitively, pseudo-expectations are relaxations of proper distributions. Given any real distribution μ over $\mathbb{R}^n$, we can always associate to it an expectation operator $\mathbb{E}$, which takes any (integrable) function f and maps it to its expected value: $f \mapsto \mathbb{E}[f]$. One can easily verify that any expectation operator $\mathbb{E}$ is also a valid pseudo-expectation operator, but this reverse is not true: indeed, a pseudo-expectation operator need to be well-formed for functions which are not low-degree polynomials.

Moreover, suppose that μ was supported only on points in $\mathcal{A}$. Then it is also easily verified that $\mathbb{E}$ is also a pseudo-expectation which satisfies $\mathcal{A}$. Thus, the class of pseudo-expectations that satisfy $\mathcal{A}$ can be thought of as a relaxation of the set of distributions supported on solutions to $\mathcal{A}$. This is the sense in which we will typically use pseudo-expectations: while it is hard to find a distribution over solutions to $\mathcal{A}$, it turns out that it is easy to find a "pseudo-distribution" whose pseudo-expectation acts as if it was a distribution supported on solutions to $\mathcal{A}$.

The connection between pseudo-expectations and SoS proofs can be seen from the following. Let $\widetilde{\mathbb{E}}$ is some degree-k pseudo-expectation satisfying $\mathcal{A}$, and suppose further that $\mathcal{A} \vdash_k f \geq 0$. Then it is readily seen from the definitions that $\widetilde{\mathbb{E}}[f] \geq 0$. Indeed, it turns out that there is a formal sense in which this is an if-and-only-if statement:

Theorem 2.1. *Let $\mathcal{A} = \{f_1 \geq 0, \ldots, \geq f_m \geq 0\}$ be a system of polynomial constraints of degree k that includes the constraint $\|x\|_2^2 \leq C$, for some $C > 0$. Then, for every $\ell \geq k$ even, and every polynomial $f \in \mathbb{R}[x]_{\leq \ell}$, either:*

- *for every $\varepsilon > 0$, we have that $\mathcal{A} \vdash_\ell \{f \geq -\varepsilon\}$, or*
- *there exists a degree-ℓ pseudo-expectation $\widetilde{\mathbb{E}}$ so that $\widetilde{\mathbb{E}} \models_\ell \mathcal{A}$ and $\widetilde{\mathbb{E}} f \leq 0$.*

Proof sketch. Let $C \subseteq \mathbb{R}[x]_{\leq \ell}$ be the set of polynomials g so that $A \vdash_\ell \{g \geq 0\}$. We note that C is a *cone*, that is, if I take any $g_1, g_2 \in C$, and any non-negative coefficients a_1, a_2, then $a_1 g_1 + a_2 g_2 \in C$ (why?). We will show that if f is in the closure of C, then the first item happens, and otherwise, the second one does. Since we are computer scientists whose hearts need not be swayed by abstract nonsense, let us stop worrying about silly topological issues, and just pretend the set C is closed. Then, if $f \in C$, clearly the first case happens. To be formally correct, one actually needs to be a bit careful here: we would actually need to use assumption that $\mathcal{A}$ includes the constraint $\|x\|_2^2 \leq M$.

Thus, it remains to demonstrate that if f is not in the closure of C, then the second case happens. In this case, by the properties of convex cones, there exists a linear functional L on $\mathbb{R}[x]_{\leq \ell}$ so that $L(f) < 0$ and $L(g) \geq 0$ for all $g \in C$. We first claim that $L(1) > 0$. Indeed, since $\vdash_\ell 1 \geq 0$, it is trivially the case that $\mathcal{A} \vdash_\ell 1 \geq 0$, and so $1 \in C$. Similarly, $L(h^2) \geq 0$ for all h of degree at most $\ell/2$. Therefore, if we rescale L so that $L(1) = 1$, we can see that L is a valid pseudo-expectation. Since $\mathcal{A} \vdash_\ell \{h^2 f_i \geq 0\}$ for all f_i, we can also see that $L(h^2 f_i) \geq 0$, which implies that $L \models_\ell \mathcal{A}$. □

3 Background on tensors and tensor optimization

Formally, an order-k tensor over $\mathbb{R}^d$ is a collection of numbers indexed by $i_1, \ldots, i_k \in [d]$. For instance, an order-1 tensor is just a vector $v = (v_1, \ldots, v_d)$ in $\mathbb{R}^d$, and an order-2 tensor is a $d \times d$ matrix, i.e. $M = \{M_{ij}\}_{i,j=1}^d$. So an order-k tensor is just some object T indexed by $T_{i_1, \ldots, t_k}$.

It is often more useful to think of a tensor as a multilinear operator

$$T : \underbrace{\mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{k \text{ times}} \to \mathbb{R} \,,$$

that is, it is a function that takes $x_1, \ldots, x_k \in \mathbb{R}^d$ and maps it to some value $T(x_1, \ldots, x_k)$ so that for any index j, we have that

$$T(x_1, \ldots, a y_j + b z_j, \ldots, x_k) = a \cdot T(x_1, \ldots, y_j, \ldots, x_k) + b \cdot T(x_1, \ldots, z_j, \ldots, x_k) \,.$$

For instance, 1-tensors, or vectors, correspond to linear functions: for every $v \in \mathbb{R}^d$, we can associate it uniquely with the function $\ell_v(x) = \langle x, v \rangle$, and 2-tensors, or matrices, correspond to bilinear functions: for every $M \in \mathbb{R}^d \times \mathbb{R}^d$ we can uniquely associate it with the function

$$f(x, y) = x^\top M y = \sum_{i,j} M_{ij} x_i y_j \ .$$

More generally, given any tensor T with coefficients $T_{i_1,\ldots,i_k}$, this tensor is uniquely associated with the multilinear polynomial

$$p_T(x_1, \ldots, x_k) = \sum_{i_1,\ldots,i_k} T_{i_1,\ldots,i_k} x_{1i_1} \ldots x_{ki_k} \ .$$

An important operation when dealing with tensors (that we have already encountered somewhat) is the notion of tensor products (also known as Kronecker products). Given two tensors $T^{(1)}, T^{(2)}$ of orders k_1 and k_2, respectively, the tensor product of $T^{(1)}$ and $T^{(2)}$, denoted $T^{(1)} \otimes T^{(2)}$, is a tensor of order $k_1 + k_2$ of the form

$$\left(T^{(1)} \otimes T^{(2)}\right)_{i_1,\ldots,i_{k_1},j_1,\ldots,j_{k_2}} = T^{(1)}_{i_1,\ldots,i_{k_1}} T^{(2)}_{j_1,\ldots,j_{k_2}} \ .$$

For instance, for any vector v, we have that $v \otimes v$ is just the matrix $vv^\top$, and $v^{\otimes k}$ is the order-k tensor whose $i_1, \ldots, i_k$-th entry is $v_{i_1} \ldots v_{i_k}$.

4 Optimizing over pseudo-expectations and SoS proofs

As alluded to previously, the key appeal of pseudo-expectations is that the set of degree-k pseudo-expectations satisfying any set of constraints $\mathcal{A} = \{f_1 \geq 0, \ldots, f_m \geq 0\}$ forms a convex set. Moreover, by linearity, a degree-k pseudo-expectation over $\mathbb{R}^d$ is parameterized by its value on every monomial. Let us be slightly more formal about this. Let $\mathcal{P}_{d,k} \subset \mathbb{N}^d$ denote the set of d-length vectors consisting of non-negative integers $(\alpha_1, \ldots, \alpha_d)$ so that $\sum_{i=1}^d \alpha_i \leq k$. Each vector $\alpha \in \mathcal{P}_{d,k}$ corresponds to a unique monomial $x^\alpha = \prod_{i=1}^d x_i^{\alpha_i}$, and these monomials form a complete basis of all polynomials of degree at most k. In other words, a degree-k pseudo-expectation $\widetilde{\mathbb{E}}$ is characterized by an associated *moment tensor* $M : \mathcal{P}_{d,k} \to \mathbb{R}$ whose values are given by $M_\alpha = \widetilde{\mathbb{E}}[x^\alpha]$. It will actually be slightly more convenient to think of M as a matrix over $\mathcal{P}_{d,k/2} \times \mathcal{P}_{d,k/2}$, where we define

$$M_{\alpha,\beta} = \widetilde{\mathbb{E}}\left[x^{\alpha+\beta}\right] \ .$$

Similarly, given a degree-k polynomial $f = \sum_\alpha c_\alpha x^\alpha$, we can associate it with a matrix T_f over $\mathcal{P}_{d,k/2} \times \mathcal{P}_{d,k/2}$:

$$(T_f)_{\alpha,\beta} = \frac{1}{n_{\alpha+\beta}} c_{\alpha+\beta} \ ,$$

where n_α is the number of ways $\alpha \in \mathcal{P}_{d,k}$ can be written as a sum of two elements in $\mathcal{P}_{d,k/2}$, so that

$$\widetilde{\mathbb{E}}[f] = \sum_\alpha c_\alpha \widetilde{\mathbb{E}}[x^\alpha] = \langle T, M \rangle \ .$$

Here's the punchline: suppose that $f \in \mathcal{P}_{d,k/2}$ has the form $f = \sum_{|\alpha| \leq k/2} c_\alpha x^\alpha$. Then,

$$\widetilde{\mathbb{E}}[f^2] = \sum_{\alpha,\beta} c_\alpha c_\beta \widetilde{\mathbb{E}}\left[x^{\alpha+\beta}\right] = c^\top M c \ .$$

So in other words, pseudo-expectations of squares of functions correspond to evaluating quadratic forms against this moment matrix M. In particular, this implies that $\widetilde{\mathbb{E}}[f^2] \geq 0$ if and only if M is positive semi-definite! Thus, we can rephase the condition that pE is a valid pseudo-expectation to the condition that $M_{(0,\ldots,0)} = 1$, and M is positive semi-definite.

Moreover, suppose we want to constrain that $\widetilde{\mathbb{E}} \models_\ell \{f \geq 0\}$, i.e. that $\widetilde{\mathbb{E}}[hf] \geq 0$ for all sum-of-squares polynomials h. First, suppose that f was already of degree ℓ, so that we can't add any non-trivial polynomials h. Then, this condition is equivalent to the condition that $\widetilde{\mathbb{E}}[f] \geq 0$, which is just a linear constraint on M, as is just the constraint that $\langle M, T_f \rangle \geq 0$. For f which are of degree less than ℓ, the constraint that $\widetilde{\mathbb{E}}[hf] \geq 0$ for all h can be encoded as follows: we can define a new linear operator L defined by $L[g] = \widetilde{\mathbb{E}}[gf]$. Note that this may itself be a pseudo-expectation, we can still write its coefficients in a moment tensor, and when done so, these coefficients are just a linear transformation of the coefficients of those in M. Then, the condition that $\widetilde{\mathbb{E}}[hf] \geq 0$ is equivalent to the condition that $L(h) \geq 0$ for all sum-of-squares polynomials, which is equivalent to the condition that L's moment tensor is PSD. In particular, this is again a convex condition on M. By combining these observations with the classic theory of convex optimization, we obtain:

Theorem 4.1 ((informal))**.** *Let k be even. There is an algorithm which, given a set of polynomial constraints $\{f_1 \geq 0, \ldots, f_m \geq 0\}$ of degree at most k over n variables which has a non-trivial solution, finds a pseudo-expectation which approximately satisfies these constraints, in time* $\mathrm{poly}((nm)^k)$.

One needs to be a bit careful here, actually, in terms of what we mean by "approximately". This is because one issue is that pseudo-expectations need not be bounded necessarily. But as long as one has reasonable (e.g. even exponential) upper bounds on how large the coefficients of the moment tensors can be, classical algorithms can solve these problems efficiently, and to extremely high precision. So for the rest of these lectures, we'll ignore issues of precision and approximation when we solve these SoS programs. See [11] for a more in-depth discussion of these issues.

Finding SoS proofs via convex optimization There is a similar sense in which one can hope to efficiently encode and find low-degree SoS proofs directly. In fact, there is a formal sense in which this operation is the convex dual of the problem of finding a pseudo-expectation described above.

As a demonstration, let's see how to automatically find a degree-k SoS proof that $f \geq 0$ for some f, that is, we wish to find $s_1, \ldots, s_m$ of degree at most $k/2$ so that $f(x) = \sum_{i=1}^m s_i(x)^2$. Once again, the matrix interpretation will be useful: note that if we let $s_i(x) = \sum_{|\alpha| \leq k/2} (c_i)_\alpha x^\alpha$, then $s_i^2(x) = \sum_{\alpha,\beta} (c_i)_\alpha (c_i)_\beta x^{\alpha+\beta}$, so once again we have that

$$(T_{s_i^2})_{\alpha,\beta} = c_\alpha c_\beta \ ,$$

and so this condition is equivalent to the condition that

$$M_f = \sum_{i=1}^m c_i c_i^\top \ ,$$

which is equivalent to the condition that M_f is PSD! So to certify that $f \geq 0$ has a sum-of-squares proof, it suffices to check that $M_f \succeq 0$. Moreover, to find a proof, if it exists, it's also straightforward: simply take the eigenvectors of M_f, and these will correspond to the c vectors before.

5 Using SoS

We will have examples of doing so in the next lecture, but just to give a high-level overview of how SoS algorithms typically work, these algorithms typically follow a structure like the following. Suppose I want to solve some optimization problem:

1. Write down the SoS relaxation of this problem, i.e. write down some system of polynomial constraints $\mathcal{C} = \{f_1 \geq 0, \ldots, f_m \geq 0\}$ which models this optimization problem, and find a pseudo-expectation $\widetilde{\mathbb{E}}$ that solves it.

2. Prove that $\mathcal{C} \vdash \mathcal{A}$, where $\mathcal{A}$ is some useful property. Note that this implies that $\widetilde{\mathbb{E}} \models \mathcal{A}$.

3. Use the fact that $\mathcal{A}$ is useful to somehow "round" the solution $\widetilde{\mathbb{E}}$ into something which gives genuinely useful information.

It is these last two steps that require some bespoke knowledge about the problem instance and structure. The typical way this is done is to pretend that we have a genuine distribution over solutions. Then, derive inferences based on the structure of $\mathcal{C}$ that are "not too complicated" which yield useful information about what these solutions must've been which allow us to recover individual solutions from the distribution. The key insight is that so long as these proofs were SoS, then they are captured by $\widetilde{\mathbb{E}}$, and thus we can hope this information to round! This is the so-called "Marley's principle": just don't worry too much, and hope that every little thing will be alright. We'll see a more concrete instantiation of this algorithmic philosophy next lecture.

References

[1] Naum Zuselevich Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics*, 23(5):695–700, 1987.

[2] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* California Institute of Technology, 2000.

[3] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

[4] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[5] Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326. ACM, 2012.

[6] Ryan O'Donnell and Yuan Zhou. Approximability and proof complexity. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1537–1556. SIAM, 2013.

[7] Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40, 2014.

[8] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015.

[9] Jean-Louis Krivine. Anneaux préordonnés. *Journal d'analyse mathématique*, 12:p–307, 1964.

[10] Gilbert Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974.

[11] Ryan O'Donnell. Sos is not obviously automatizable, even approximately. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, pages 59–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017.