Sums of Squares for Estimation Problems*

Aaron Geelon So

November 13, 2018

1 Review of SOS

First, we review SOS as a proof system for polynomials systems. Before proceeding to estimation problems, let's motivate SOS from a more general perspective to understand the techniques. For now, instead of focusing on polynomials, we'll look at all measurable functions over $\mathbb{R}^n$. Let $\mathcal{A}$ be a collection of constraints on $\mathbb{R}^n$,

$$\mathcal{A} := \{f_1 \geq 0, \ldots, f_m \geq 0\},$$

and let $\mathcal{X} \subset \mathbb{R}^n$ be the feasible set where all constraints are satisfied. We make two remarks:

(i) the collection $\mathcal{C}$ of functions f that are nonnegative on $\mathcal{X}$ is equivalent to the collection of functions of the following form:

$$f = \sum_i g_i^2 f_i + h^2. \tag{1}$$

This collection $\mathcal{C}$ seems somewhat reminiscent of the radical of $\mathcal{A}$ from algebraic geometry.[1]

(ii) let $\mathbb{E}_\mu : \mathcal{F} \to \mathbb{R}$ be a linear functional parametrized by μ, where $\mu : \mathbb{R}^n \to \mathbb{R}$ has finite support,

$$\mathop{\mathbb{E}}_{\mu} f := \sum_{x \in \operatorname{supp}(\mu)} \mu(x) f(x) \qquad \text{and} \qquad \mathop{\mathbb{E}}_{\mu} 1 = 1. \tag{2}$$

The collection of all such μ such that $\mathbb{E}_\mu f \geq 0$ when $f \in \mathcal{C}$ are precisely the finitely-supported probability distributions, where $\operatorname{supp}(\mu) \subset \mathcal{X}$.

From these two facts, we could build a proof system. In particular, if f can be represented as in Equation 1, we can say there is a *sum of squares proof* for $f \geq 0$ when constrained to $\mathcal{A}$, writing:

$$\mathcal{A} \vdash \{f \geq 0\}.$$

Further, we can think of μ satisfying $\mathbb{E}_\mu f \geq 0$ for all $f \in \mathcal{C}$ as a witness to the satisfiability of $\mathcal{A}$, writing:

$$\mu \vDash \mathcal{A}.$$

Notice that if we can prove that there is a unique point $x^* \in \mathbb{R}^n$ that satisfies $\mathcal{A}$, i.e. $\mathcal{A} \vdash \{x = x^*\}$, then $\mu \vDash \mathcal{A}$ implies $\mu = \delta_x$, the Dirac distribution whose mass is concentrated on x. And so, if there were an efficient way to obtain the linear functional $\mathbb{E}_\mu$ when given $\mathcal{A}$, then we could efficiently solve for x^*. Of course, this is asking for too much in this general case, but perhaps not in the case restricted to polynomials.

*Lecture follows [RSS2018] and heavily cites [BS2016].

[1] By Hilbert's Nullstellensatz, the radical of an ideal I generate by a set of functions $\mathcal{A}$ over an algebraically closed field is the collection of functions whose zero set coincide with the zero set of $\mathcal{A}$. And indeed, we could think of $\mathcal{C}$ as the consequence of a Positivstellensatz.

1.1 Formalities

We'll now give the polynomial framework in which SOS operates. All functions we consider will come from $\mathbb{R}[X_1, \ldots, X_n]$, real polynomials with n indeterminants. For brevity, we'll condense $(X_1, \ldots, X_n)$ to just X, and let $\mathbb{R}[X]_{\leq d}$ be the collection of polynomials $p(X_1, \ldots, X_n)$ with degree at most d.

Definition 1. *Let $\mathcal{A} = \{p_1 \geq 0, \ldots, p_m \geq 0\}$ be a collection of polynomial constraints. We say that $\mathcal{A}$ is a collection of* axioms*. We say that there is a* sum-of-squares proof *of $q \geq 0$ from $\mathcal{A}$ if:*

$$\left(\sum_k c_k^2\right) q = \sum_i a_i^2 p_i + \sum_j b_j^2,$$

where $a_i, b_j, c_k \in \mathbb{R}[X]$ and the sums are finite. The degree *of the sum-of-squares proof is the degree d of the polynomial $\sum_i a_i^2 p_i + \sum_j b_j^2$. We write $\mathcal{A} \vdash_d \{q \geq 0\}$.*

Fact 2 ([P2000], [L2001])**.** *If a degree-d sum-of-squares proof exists, then semidefinite programming (SDP) can find a proof in $n^{O(d)}$ time. As such an algorithm for each even degree d, we call this family of algorithms the* sum-of-squares SDP hierarchy *or the* Laserre hierarchy*.*

Proposition 3. *If there is a degree-d sum-of-squares proof that $q \geq 0$ given $\mathcal{A}$, i.e. $\mathcal{A} \vdash_d \{q \geq 0\}$, then $q(x) \geq 0$ for all $x \in \mathbb{R}^n$ satisfying $\mathcal{A}$.*

Definition 4. *For a system of polynomial constraints $\mathcal{A}$, a* degree-d pseudo-expectation *is a linear functional $\widetilde{\mathbb{E}} : \mathbb{R}[X]_{\leq d} \to \mathbb{R}$ satisfying the following:*

(i) $\widetilde{\mathbb{E}}\, 1 = 1$

(ii) $\widetilde{\mathbb{E}}\, f \geq 0$ *whenever $f \in \mathbb{R}[X]_{\leq d}$ is of the form:*

$$f = \sum_i a_i^2 p_i + \sum_j b_j^2.$$

Fact 5. *If for a polynomial system $\mathcal{A}$, there exists a degree-d pseudo-expectation, then the dual SDP computes a degree-d pseudo-expectation in time $n^{O(d)}$.*

Remark 6. *Notice that the evaluation functional $\mathrm{ev}_x : \mathbb{R}[X]_{\leq d} \to \mathbb{R}$ is linear, defined so that $\mathrm{ev}_x(f) \equiv f(x)$. Furthermore, since $\mathbb{R}[X]_{\leq d}$ is a finite linear space, its dual is also finite dimensional. It follows that every degree-d pseudo-expectation has representation of the form:*

$$\widetilde{\mathbb{E}} \equiv \sum_{i=1}^{k} \mu_i \mathrm{ev}_{x_i}.$$

In particular, we can think of $\mu : \mathbb{R}^n \to \mathbb{R}$ as a pseudo-distribution *on $\mathbb{R}^n$, supported on the x_i's, and where the mass on the point x_i is μ_i. Such a representation is not unique, but given μ, we can write:*

$$\widetilde{\mathbb{E}}_{\mu} f := \sum_{x \in \mathrm{supp}(\mu)} \mu(x) f(x).$$

Example 7. *In contrast to the motivating discussion over all functions, the support of μ is not restricted to the feasible set of $\mathcal{A}$, nor are the 'pseudo-probability masses' necessarily nonnegative. For example, consider the degree-2 pseudo-distribution $\mu : \mathbb{R}[X]_{\leq 2} \to \mathbb{R}$, where:*

$$\mu(x) = \begin{cases} 1 & x = \pm 1 \\ -1 & x = 0 \\ 0 & \text{o.w.} \end{cases}$$

Indeed, for all $f = a(x+b)$, we see that $\widetilde{\mathbb{E}}_\mu f^2 = a^2(2+b^2) \geq 0$. Note that this is just one representation of the linear functional $\widetilde{\mathbb{E}}$ such that $\widetilde{\mathbb{E}}\,1 = 1$, $\widetilde{\mathbb{E}}\,X = 0$, and $\widetilde{\mathbb{E}}\,X^2 = 2$.

We can define degree-d pseudo-distributions in generally by setting $\mathcal{A} = \varnothing$.

Definition 8. *A function $\mu : \mathbb{R}^n \to \mathbb{R}$ is a degree-d* pseudo-distribution *if for all polynomials f such that $\deg(f^2) \leq d$, the pseudo-expectation with respect to μ is nonnegative, $\widetilde{\mathbb{E}}_\mu f^2 \geq 0$. Let $\mathcal{A}$ be a collection of axioms. We say that μ* satisfies *$\mathcal{A}$ if $\widetilde{\mathbb{E}}_\mu$ is a pseudo-expectation consistent with $\mathcal{A}$. We write $\mu \vDash \mathcal{A}$.*

Lemma 9 (Soundness, [B2016])**.** *Let μ be a pseudo-distribution and let $\mathcal{A}, \mathcal{B}$ be systems of polynomial constraints. If $\mu \vDash \mathcal{A}$ and $\mathcal{A} \vdash \mathcal{B}$, then $\mu \vDash \mathcal{B}$.*

Lemma 10 (Cauchy-Schwarz, [RSS2018])**.** *Let $\widetilde{\mathbb{E}}$ be a degree-d pseudo-expectation. Let $p, q \in \mathbb{R}[X]$ with degree at most $d/2$. Then:*

$$\widetilde{\mathbb{E}}[pq] \leq \left(\widetilde{\mathbb{E}}\,p^2\right)^{1/2} \left(\widetilde{\mathbb{E}}\,q^2\right)^{1/2}.$$

This will be enough to set up the framework for estimation problems.

2 Estimation Problems

We can formalize an *estimation problem* as consisting of a set $\mathcal{X} \subset \mathbb{R}^n \times \mathbb{R}^m$ of pairs (x, y). We say that x is the *parameter* and y is the *measurement*. Out of this set of possibilities, one pair (x^*, y^*) is selected, but we only get access the measurement y^*. The goal is to (approximately) recover x^*.

Definition 11. *For a pair $(x, y) \in \mathcal{X}$, we say that y* identifies *x* exactly *when $(x', y) \in \mathcal{X}$ iff $x' = x$. We say that y* identifies *x* up to *ϵ when $(x', y) \in \mathcal{X}$ iff $\|x - x'\| \leq \epsilon$.*

If we have estimation problems that can be described with polynomials, and we measure y^*, then we can write $\mathcal{X}$ as the feasible region to a collection of polynomials

$$\mathcal{A} = \{p_1(X, y^*) \geq 0, \ldots, p_m(X, y^*) \geq 0\}.$$

Let δ_{x^*} be the (pseudo-)distribution with mass concentrated at x^*. It follows that $\widetilde{\mathbb{E}}_{\delta_{x^*}}$ is a pseudo-expectation consistent with $\mathcal{A}$. It would be great if we could retrieve $\widetilde{\mathbb{E}}_{\delta_{x^*}}$, since that would immediately give us x^*, solving the estimation problem. However, running a SDP with constraints $\mathcal{A}$ is not guaranteed to return this particular pseudo-expectation. It may not matter, if we can give a low-degree SOS proof of identifiability:

$$\mathcal{A} \vdash_d \{\|X - x^*\|^2 \leq \epsilon\},$$

where ϵ is possibly 0 if x^* is exactly identifiable. Then, we're guaranteed an efficient algorithm that retrieves x^* within $\epsilon + 2^{-n^d}$. We state this slightly more generally by allowing a third auxilary variable that can often simplify the description of $\mathcal{X}$:

Theorem 12 (Meta-theorem for efficient estimation [RSS2018])**.** *Let $p \in \mathbb{R}[X, Y, Z]$ and let (x^*, y^*, z^*) satisfy $p = 0$. Let $\mathcal{A} = \{p(X, y^*, Z) = 0\}$. Suppose:*

$$\mathcal{A} \vdash_d \{\|X - x^*\|^2 \leq \epsilon\}.$$

Then, every degree-d pseudo-distribution μ consistent with $\mathcal{A}$ satisfies:

$$\left\| x^* - \widetilde{\mathbb{E}}_{\mu} X \right\|^2 \leq \epsilon. \tag{3}$$

Furthermore, for every $d \in \mathbb{N}$, there exists an $n^{O(d)}$-time algorithm with bit-complexity at most n^d that returns and estimate $\hat{x}(y^*)$ such that $\|x^* - \hat{x}(y^*)\|^2 \leq \epsilon + 2^{-n^d}$.

Proof. By assumption, μ is consistent with $\mathcal{A}$, so that $\mu \vDash \mathcal{A}$. When we combine this with $\mathcal{A} \vdash \{\|X - x^*\|^2 \leq \epsilon\}$, soundness implies $\mu \vdash \{\|X - x^*\|^2 \leq \epsilon\}$. It follows that:

$$\widetilde{\mathbb{E}}_{\mu} \|X - x^*\|^2 \leq \epsilon.$$

On the other hand, Cauchy-Schwarz implies:

$$\left\|\widetilde{\mathbb{E}}_{\mu}(X - x^*)\right\|^2 = \widetilde{\mathbb{E}}_{\mu} \left\langle X - x^*, \widetilde{\mathbb{E}}_{\mu}(X - x^*)\right\rangle \leq \left\|\widetilde{\mathbb{E}}_{\mu}(X - x^*)\right\| \cdot \underbrace{\widetilde{\mathbb{E}}_{\mu} \|X - x^*\|^2}_{\epsilon},$$

proving Equation 3.

To get an efficient algorithm, we just need to obtain $\widetilde{E}_\mu$ using the SDP, which takes $n^{O(d)}$ time, and with bit complexity n^d, we can ensure that our computation $\hat{x}$ is within 2^{-n^d} of $\widetilde{\mathbb{E}}_\mu X$. ☐

Let's see how this framework plays out in a few examples.

2.1 Matrix completion

In the general problem, let $M \in \mathbb{R}^{n \times n}$ be a rank-r matrix. We're given access to values M_Ω, where $\Omega \subset [n] \times [n]$ is a subset of indices. The goal is then to recover M from M_Ω. It turns out that we can often recover M almost exactly with high probability under certain conditions.

Definition 13. *Let $v_1, \ldots, v_m \in \mathbb{R}^n$ be unit vectors. We say that they are* μ-incoherent *if for all v_i and for all standard basis vectors e_j,*

$$\langle v_i, e_j \rangle^2 \leq \frac{\mu}{n}.$$

For example, random unit vectors from the sphere are μ-incoherent when $\mu \leq O(\log n)$ with high probability. When the right and left singular values of M are μ-incoherent, then as long as M_Ω contains at least $m \geq \mu r n \cdot O(\log n)^2$ entries, we can recover M exactly with high probability. This bound on m is almost optimal, since an $n \times n$ rank-r matrix has $\Omega(rn)$ degrees of freedom, for each singular vector.

Let's consider the simplified case where M is a rank-r projection:

Theorem 14 (Identifiability for matrix completion, [RSS2018])**.** *Let $M = \sum_{i=1}^r a_i a_i^T$ be an r-dimensional projector, where $a_i \in \mathbb{R}^n$ are orthonormal vectors with incoherence $\mu = \max_{i,j} n \cdot \langle a_i, e_j \rangle^2$. Let $\Omega \subset [n] \times [n]$ be a random symmetric subset of size $|\Omega| = m$. Consider the following system of polynomial constraints:*

$$\mathcal{A} = \{(BB^T)_\Omega = M_\Omega, B^T B =_r\}.$$

If $m \geq \mu r n \cdot O(\log n)^2$, then with high probability over choice of Ω,

$$\mathcal{A} \vdash_4 \{\|BB^T - M\|_F = 0\}.$$

The proof will utilize the following lemma, whose proof we'll leave to the literature.

Lemma 15 ([G2011], [R2011], [C2015])**.** *Let M, a_i as before, and in particular, the a_i's are μ-incoherent. Let $\overline{\Omega} \subset [n] \times [n]$ be the complement of Ω. Then, with high probability over the choice of Ω, there exists a symmetric matrix N such that $N_\Omega = 0$, and*

$$-0.9(\mathrm{Id}_n - M) \preceq N - M \preceq 0.9(\mathrm{Id}_n - M).$$

Proof of Theorem 14. As notation, we denote the Frobenius inner product of two matrices $A, B \in \mathbb{R}^{n\times k}$ to be $\langle A, B\rangle = \mathrm{Tr}(AB^T)$.

By the previous lemma, we obtain from the first inequality $-0.9(\mathrm{Id}_n - M) \preceq N - M$:

$$-0.9\langle \mathrm{Id}_n - M, M\rangle \leq \langle N - M, M\rangle,$$

since $M \succeq 0$. And as $\langle \mathrm{Id}_n, M\rangle = \langle M, M\rangle = r$, we obtain $r \leq \langle N, M\rangle$. Because $N = N_\Omega$, this implies that $\langle N, M\rangle = \langle N, M_\Omega\rangle$. Then:

$$\mathcal{A} \vdash r \leq \langle N, BB^T\rangle. \tag{4}$$

Again, by the previous lemma, we obtain from the second inequality $N - M \preceq 0.9(\mathrm{Id}_n - M)$,

$$\langle N - M, BB^T\rangle \leq 0.9\langle \mathrm{Id}_n - M, BB^T\rangle.$$

Simplifying gives us:

$$\mathcal{A} \vdash \langle N, BB^T\rangle \leq 0.9r + \langle M, BB^T\rangle. \tag{5}$$

Combining Equations 4 and 5, we obtain $\mathcal{A} \vdash r \leq \langle M, BB^T\rangle$. On the other hand, because both M and BB^T are rank-r, we have $\|M\|_F^2 = \|BB^T\|_F^2 = r$. Cauchy-Schwarz shows that $\mathcal{A} \vdash \langle M, BB^T\rangle \leq r$. It follows that:

$$\mathcal{A} \vdash \|M - BB^T\|^2 = 0,$$

proving identifiability using SOS proofs of degree at most 4. □

2.2 Tensor completion

Tensor completion generalizes matrix completion. For 3-tensors, all known efficient algorithms require $r \cdot \widetilde{O}(n^{1.5})$ observed entries, while the information-theoretic lower bound is $r \cdot O(n)$. It is an open question whether this gap is necessary. Again, we'll consider the more general case, where a tensor $T = \sum_{i=1}^r a_i^{\otimes 3}$ is rank-r with orthonormal components $a_i \in \mathbb{R}^n$ that are μ-incoherent. The following uses $rn^{1.5} \cdot (\mu \log n)^{O(1)}$ random entries of X.

Theorem 16 (Identifiability for tensor completion, [RSS2018])**.** *Let $T = \sum_{i=1}^r a_i^{\otimes 3}$ be a rank r orthogonally decomposable tensor, with incoherence $\mu = \max_{i,j} n \cdot \langle a_i, e_j\rangle^2$. Let $\Omega \subset [n]^3$ be a random symmetric subset of size $|\Omega| = m$. Consider the following system of polynomial constraints:*

$$\mathcal{A} = \left\{ \left(\sum_{i=1}^r b_i^{\otimes 3} \right)_\Omega = T_\Omega, B^T B = \mathrm{Id}_n \right\}.$$

Suppose $m \geq rn^{1.5} \cdot (\mu \log n)^{O(1)}$. Then, with high probability over the choice of Ω,

$$\mathcal{A} \vdash_{O(1)} \left\{ \left\| \sum_{i=1}^r b_i^{\otimes 3} - T \right\|_F^2 = 0 \right\}.$$

The proof technique is similar to the matrix completion case.

2.3 Clustering

Consider a collection of n data points drawn from a mixture of k Gaussians in $\mathbb{R}^d$, $\mathcal{N}(\mu_1, \mathrm{Id}_d), \ldots, \mathcal{N}(\mu_k, \mathrm{Id}_d)$. We can denote by $X^* \in \{0,1\}^{n\times n}$ be the k-clustering matrix where $X^*_{ij} = 1$ iff y_i and y_j were drawn from the same Gaussian. The goal is to recover X^* from seeing the y_i's.

Theorem 17 (Clustering with SOS). *Given the above setting, there exists an algorithm that outputs a k-clustering matrix $X \in \{0,1\}^{n\times n}$ in quasipolynomial time $n + (dk)^{(\log k)^{O(1)}}$ with the guarantees: if $\mu_1, \ldots, \mu_k$ have separation $\min_{i\neq j} \|\mu_i - \mu_j\| \geq O(\sqrt{\log k})$ and $n \geq (dk)^{(\log k)^{O(1)}}$, then with high probability,*

$$\|X - X^*\|_F^2 \leq 0.1 \cdot \|X^*\|_F^2.$$

More generally, the proof technique only requires bounded moments up to ℓ: for each cluster S_κ

$$\left\| \mathop{\mathbb{E}}_{y\in S_\kappa} (1, y-\mu_\kappa)^{\otimes \ell} - \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\mathrm{Id}_d)} (1, g)^{\otimes \ell} \right\|_F^2 \leq \epsilon.$$

This gives rise to a polynomial constraint. Additionally, the constraint that X is a k-clustering matrix with respect to the S_κ's gives a set of constraints $\mathcal{A} = \{p(X, Y, Z) = 0\}$. On observing Y^*, we need derive an SOS proof from $\mathcal{A}(Y^*) = \{p(X, Y^*, Z) = 0\}$ that with high probability for $\ell \leq (\log k)^{O(1)}$,

$$\mathcal{A}(Y^*) \vdash_\ell \left\{ \|X - X^*\|_F^2 \leq 0.1 \cdot \|X^*\|_F^2 \right\}.$$

Details are left to the reference, [HL2018], [KSS2018], [DKS2018].

3 Additional Techniques

3.1 Tensor decomposition

Many estimation problems can be reduced to tensor decomposition. For example, see [A+2014] for use of tensor decomposition to estimate problems like latent Dirichlet allocation, mixtures of Gaussians, etc. Concretely, given an order-k tensor, $T = \sum_{i=1}^r a_i^{\otimes k}$ with $a_i \in \mathbb{R}^n$, find $u \in \mathbb{R}^n$ that is close to some a_i in the sense of cosine similarity (up to signs):

$$\max_{i\in[r]} \frac{|\langle a_i, u\rangle|}{\|a_i\| \cdot \|u\|} \geq 0.9.$$

We can use sums of squares to help solve this problem in polynomial time and $\widetilde{\Omega}(n^{1.5})$ components drawn uniformly at random from the unit sphere.

The technique will reduce to the robust Jennrich's algorithm for tensor decomposition:

Theorem 18 (Robust Jennrich's algorithm, [MSS2016] [SS2017]). *Let $T \in (\mathbb{R}^n)^{\otimes 3}$ be an order-3 tensor, and let $a_1, \ldots, a_r \in \mathbb{R}^n$ be unit vectors with orthogonality defect $\|\mathrm{Id}_r - A^T A\| \leq \epsilon$. Suppose:*

$$\left\| T - \sum_{i=1}^r a_i^{\otimes 3} \right\|_F^2 \leq \epsilon \cdot r,$$

and that $\max\{\|T\|_{\{1,3\},\{2\}}, \|T\|_{\{1\},\{2,3\}}\} \leq 10$. Then, there exists a randomized polynomial-time algorithm that outputs a unit vector $u \in \mathbb{R}^n$ such that:

$$\max_{i\in[r]} \langle a_i, u\rangle \geq 0.9,$$

with at least inverse polynomial probability.

Notice that Jennrich's algorithm requires a small orthogonality defect. The strategy for using SOS is to estimate a noisy version of $\sum_{i=1}^r a_i^{\otimes 6}$ from seeing $\sum_{i=1}^r a_i^{\otimes 3}$. Then, viewing the order-6 tensor as a 3-tensor over $\mathbb{R}^{n^2}$, with squared components $a_i \otimes a_i$, these vectors may become linearly indepedent and nearly orthogonal when $r \ll n$ under certain conditions. Thus, our SOS problem is to estimate X from Y over the feasible domain:

$$\mathcal{X} = \left\{ (X, Y) : X = \sum_{i=1}^r a_i^{\otimes 6}, Y = \sum_{i=1}^r a_i^{\otimes 3} \right\}.$$

References

[A+2014] Anandkumar, A., et al. *"Tensor decompositions for learning latent variable models."* The Journal of Machine Learning Research 15.1 (2014): 2773-2832.

[BS2016] Barak, B. Steurer, D. *Proof, beliefs, and algorithms through the lens of sum-of-squares* (2016).

[C2015] Chen, Y. *"Incoherence-Optimal Matrix Completion."* IEEE Trans. Information Theory 61.5 (2015): 2909-2923.

[DKS2018] Diakonikolas, I., Kane, D., Stewart, S.. *"List-decodable robust mean estimation and learning mixtures of spherical gaussians."* Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 2018.

[G2011] Gross, David. *"Recovering low-rank matrices from few coefficients in any basis."* IEEE Transactions on Information Theory 57.3 (2011): 1548-1566.

[HL2018] Hopkins, S., and Li, J. *"Mixture models, robustness, and sum of squares proofs."* Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 2018.

[KSS2018] Kothari, P., Steinhardt, J., Steurer, D. *"Robust moment estimation and improved clustering via sum of squares."* Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 2018.

[MSS2016] Ma, T., Shi, J., Steurer, D. *"Polynomial-time tensor decompositions with sum-of-squares."* Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on. IEEE, 2016.

[L2001] Lasserre, J. *"Global optimization with polynomials and the problem of moments."* SIAM Journal on optimization 11.3 (2001): 796-817.

[P2000] Parrilo, P. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* Diss. California Institute of Technology (2000).

[R2011] Recht, B. *"A simpler approach to matrix completion."* Journal of Machine Learning Research 12.Dec (2011): 3413-3430.

[RSS2018] Raghavendra, P., Schramm, T., Steurer, D. *"High-dimensional estimation via sum-of-squares proofs."* arXiv preprint arXiv:1807.11419 (2018).

[SS2017] Schramm, T., Steurer, D. *"Fast and robust tensor decomposition with applications to dictionary learning."* arXiv preprint arXiv:1706.08672 (2017).