Mathematical Definitions

Let us now turn to formally defining the problem of polynomial optimization and the sum-of-squares algorithm. In the first few lectures, we will restrict our attention to the following basic special case, which still captures many interesting examples:

1. Problem (Non-negativity over the hypercube). Given a low-degree polynomial $f\colon \{0,1\}^n \to \mathbb{R}$, decide if $f \geq 0$ over the hypercube or if there exists a point $x \in \{0,1\}^n$ such that $f(x) < 0$.

One interesting computational task captured by this problem is finding a maximum cut in a graph. For an n-vertex graph G, we encode a bipartition of the vertex set of G by a vector $x \in \{0,1\}^n$ and we let $f_G(x)$ be the number of edges cut by the bipartition x. This function is a degree-2 polynomial,

$$f_G(x) = \sum_{\{ij\}\in E(G)} (x_i - x_j)^2 . \tag{1}$$

Therefore, deciding if the polynomial $c - f_G$ takes a negative value over the hypercube is the same as deciding if the maximum cut in G is larger than c.

The traditional definition of the Max-Cut problem is to recover, given a graph G, the cut x maximizing $f_G(x)$. A priori, computing $\max_x f_G(x)$, or deciding whether this maximum is larger than c, is an easier task than recovering the cut. However, in this and many other settings, all known algorithms for solving the decision task (i.e., is $\max f_G(x)$ larger than c?) easily generalize to solving the search problem (i.e., recovering x that exactly or approximately maximizes $f_G(x)$).

The sum-of-squares algorithm, when restricted to the special case Problem 1, gets a polynomial $f\colon \{0,1\}^n \to \mathbb{R}$ as input and outputs

- either a proof that $f(x) \geq 0$ for all $x \in \{0,1\}^n$,
- or an object that "pretends to be" a point $x \in \{0,1\}^n$ with $f(x) < 0$ or, more generally, a *collection* of such points.

What is the form of this proof? What is the meaning of "pretends to be"? And how can we find such an object when finding an actual solution is hard? These are the questions we address next.

Sum-of-squares certificates

How could we efficiently certify for a given polynomial $f\colon \{0,1\}^n \to \mathbb{R}$ that it is nonnegative over the hypercube? Since a *square* is always non-negative, one simple certificate is to show that f agrees with a sum of squares of polynomials over the hypercube. This observation motivates the following definition.

2. Definition (sum-of-squares certificate). A degree-d sum-of-squares certificate (of non-negativity) for a function $f\colon \{0,1\}^n \to \mathbb{R}$ consists of polynomials $g_1, \ldots, g_r\colon \{0,1\}^n \to \mathbb{R}$ of degree at most $d/2$ for some $r \in \mathbb{N}$ such that

$$f(x) = g_1^2(x) + \cdots + g_r^2(x)\,. \tag{2}$$

for every $x \in \{0,1\}^n$.

We will refer to degree-d sos certificate for f also as a *degree-d sum-of-squares proof* of the inequality $f \geq 0$.

Verifying certificates

In what sense is this certificate efficiently verifiable? Since $g_1 \ldots, g_r$ have degree at most $d/2$, we can represent each polynomial g_i by $n^{O(d)}$ coefficients (say in the monomial basis). It also turns out that we can assume r to be at most $n^{O(d)}$.[1] Thus in $n^{O(d)}$ time, we can reduce the task of verifying (2) to the task of checking that an explicit polynomial p (obtained by computing the coefficients of $f - (g_1^2 + \cdots + g_r^2)$) vanishes for every $x \in \{0,1\}^n$. It can be shown that this holds if and only if p becomes the zero polynomial if we reduce it to a *multilinear* polynomial (where every monomial with non-zero coefficient is of the form $\prod_{i\in S} x_i$ for some subset $S \subseteq [n]$) by repeatedly applying the identity $x_i^2 = x_i$ (which holds when $x_i \in \{0,1\}$).[2] Since $f - (g_1^2 + \cdots + g_r^2)$ has degree at most d, we need to consider at most $n^{O(d)}$ coefficients. Finally, some mild assumptions on f allow us to assume that the bit length of the coefficients is bounded by $n^{O(d)}$.[3] It follows that we can verify the certificate in time $n^{O(d)}$.

For large enough degree, every nonnegative functions has a sum-of-squares certificate of non-negativity:

3. Lemma (high-degree sos proofs). *Every nonnegative function $f\colon \{0,1\}^n \to \mathbb{R}$ has a degree-$2n$ sum-of-squares certificate.*

[1] See Exercise 7.

[2] See Exercise 6. The underlying technical reason is the fact that $\{x_i(x_i-1)\}_{i\in[n]}$ is a small Groebner basis for the hypercube $\{0,1\}^n$.

[3] Concretely, we need to assume that already $f - \varepsilon$ has a degree-d sos certificate.

Proof. Let $g\colon \{0,1\}^n \to \mathbb{R}$ be the function that agrees with $\sqrt{f}$ on the hypercube. This function satisfies $f = g^2$ over the hypercube and its multilinear representation of g has degree at most n. Therefore, g is a degree-$2n$ sos certificate for f. □

In the most general setting (when we allow arbitrary polynomial equality and inequality constraints over $\mathbb{R}^n$ instead of just a single polynomial inequality over the hypercube) this result is known as the *Positivstellensatz* and was proven by Krivine in 1964 (and independently but later by Stengle in 1974), extending Artin's 1927 resolution of Hilbert's 17th problem.

Finding certificates

Not only can we check sos certificates efficiently but there is also an efficient algorithm to find them if they exist. This *sum-of-squares algorithm* is based on semidefinite programming and has first been proposed by Naum Shor in 1987, later refined by Pablo Parrilo in 2000, and Jean Lasserre in 2001.

4. Theorem (sum-of-squares algorithm—certificate version). *There exists an algorithm that given a polynomial*[4] $f\colon \{0,1\}^n \to \mathbb{R}$ *(say represented in the monomial basis with polynomial bit complexity) and a number* $k \in \mathbb{N}$*, outputs a degree-k sum-of-squares certificate for* $f + 2^{-n}$ *in time* $n^{O(k)}$ *if* f *has a degree-k sos certificate.*

[4] Unless explicitly specified otherwise, when we give an n-variate degree-d polynomial as an input to an algorithm, we represent it by its coefficients in the monomial basis up to degree d. Furthermore, we assume that the bit length of the coefficients is at most polynomial in the number of coefficients, which is roughly n^d.

This result as well extends far beyond the case of a single polynomial over the hypercube to any set of polynomials equalities and inequalities over $\mathbb{R}^n$.

To get some intuition for the sum-of-squares algorithm, note that the polynomials f with degree-d sos certificates form a convex cone (a set closed under convex combination and nonnegative scaling). See Exercise 8 for some basic properties of this cone. We refer to this cone as the *degree-d sum-of-squares cone* (over the hypercube).

The key insight for Theorem 4 is that the degree-d sos cone admits a small semidefinite programming (SDP) formulation, which turns out to follow from the following characterization of sos certificates in terms of positive semidefinite matrices.

5. Theorem (positive semidefinite matrices and sos certificates). *A polynomial f has a degree-d sos certificate if and only if there exists a positive semidefinite matrix A such that for all $x \in \{0,1\}^n$,*

$$f(x) = \left\langle (1,x)^{\otimes d/2}, A(1,x)^{\otimes d/2} \right\rangle. \tag{3}$$

Proof. Suppose Eq. (3) holds for a positive semidefinite matrix A. Let g_i be the polynomial such that $g_i(x) = \langle e_i, A^{1/2}(1,x)^{\otimes d/2}\rangle$. Then, f has the following degree-d sos certificate,

$$f(x) = \left\|A^{1/2}(1,x)^{\otimes d/2}\right\|^2 = \sum_i g_i(x)^2\,. \tag{4}$$

(Here, we use that positive semidefinite matrices have square roots over the reals.)

On the other hand, suppose that f has a deg-d sos certificate, $f = \sum_{i=1}^r g_i^2$. Let $v_1,\ldots,v_r$ be vectors such that $g_i(x) = \langle v_i, (1,x)^{\otimes d/2}\rangle$ for all $x \in \mathbb{R}^n$ and let $A = \sum_i v_i v_i^\intercal$. Then, for every $x \in \{0,1\}^n$,

$$f(x) = \sum_i g_i(x)^2 = \sum_i \left\langle v_i, (1,x)^{\otimes d/2}\right\rangle^2 = \left\langle (1,x)^{\otimes d/2}, A(1,x)^{\otimes d/2}\right\rangle. \tag{5}$$

□

Exercises I

The following exercises are about basic properties of sos certificates and some examples.

6. Exercise (multilinear representation). Show that every function $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique *multilinear representation* $f(x) = \sum_{S\subseteq[n]} c_S x_S$ where $x_S = \prod_{i\in S} x_i$.

The multilinear representation of a function $f\colon \{0,1\}^n \to \mathbb{R}$ is closely related to its *Fourier transform*, see Ryan O'Donnell's excellent book on this topic.

7. Exercise (rank of sos representations). Show that every function $f\colon \{0,1\}^n \to \mathbb{R}$ with a degree-d sos certificate has one of rank at most $n^{d/2}$.

8. Exercise (closed convex cone). Show that for every $k \in \mathbb{N}$, the polynomials with degree-k sos certificates of non-negativity form a closed convex cone.

9. Exercise (minimum s-t cut). For an $n+2$-vertex digraph G with a source s and sink t, let $f(x)$ with $x \in \{0,1\}^{V(G)\setminus\{s,t\}}$ be the number of edges going out of $\{s\} \cup \{i \in V(G)\setminus\{s,t\} \mid x_i = 1\}$. Show that f is a degree-2 polynomial and that $f - c$ has a degree-4 sos certificate for all $c \in \mathbb{R}$ such that $f - c \geq 0$.

10. Exercise (spectral bound for max cut). For a graph G, let L_G be the Laplacian matrix

$$L_G = \sum_{(i,j)\in E(G)} (e_i - e_j)(e_i - e_j)^\intercal , \tag{6}$$

where $\{e_i \mid i \in V(G)\}$ is the coordinate basis. Show that every graph G with n vertices the function $\lambda_{\max}(L_G) \cdot n/2 - f_G$ has a degree-2 sos certificate.

11. Exercise (some bound). Show that for every even $d \in \mathbb{N}$ and every function $f\colon \{0,1\}^n \to \mathbb{R}$ of degree at most d, there exists some $M \in \mathbb{R}_{\geq 0}$ such that $M - f$ has a degree-d sos certificate. Also show that M can be chosen at most $n^{O(d)}$ times the largest coefficient of f in the monomial basis.

Pseudo-distributions

What can we say about a function $f\colon \{0,1\}^n \to \mathbb{R}$ if there is no degree-k sos certificate for its non-negativity? Obviously, if the function is not actually non-negative, then there is no certificate for it. Indeed that's the only kind of obstruction for very large values of k (by Lemma 3 it suffices that $k \geq 2n$). However, it turns out that for smaller values of k other kinds of obstructions exist. Since the running time of the sum-of-squares algorithm is exponential in k, understanding these more general obstructions is key.

The most direct description of obstructions for sos certificates is geometric. In the previous section, we saw that functions with degree-k sos certificates form a closed convex cone. By the hyperplane separation theorem for convex cones, for every function $f\colon \{0,1\}^n \to \mathbb{R}$ without degree-k sos certificate there exists a hyperplane through the origin that separates f from the cone of functions with degree-k sos certificates, in the sense that the halfspace H above the hyperplane contains the degree-k sos cone but not f.

How do such halfspaces look like? We can represent a halfspace H by its normal function $\mu\colon \{0,1\}^n \to \mathbb{R}$ so that

$$H = \left\{ g\colon \{0,1\}^n \to \mathbb{R} \;\middle|\; \sum_{x\in\{0,1\}^n} \mu(x) \cdot g(x) \geq 0 \right\}. \tag{7}$$

By scaling we can assume without loss of generality that $\sum_{x\in\{0,1\}^n} \mu(x) = 1$. It's illuminating to consider the special case that μ satisfies $\mu(x) \geq 0$ for all $x \in \{0,1\}^n$. Then, μ corresponds to a probability distribution over the hypercube where every point

$x \in \{0,1\}^n$ has probability $\mu(x)$. In this case, the halfspace H contains all nonnegative functions and therefore also the degree-k sos cone. The condition $f \notin H$ simply says that the expected value of $f(x)$ when x is drawn from the distribution μ is negative. In particular, in this case, if $f \notin H$ then there must exist some $x \in \{0,1\}^n$ such that $f(x) < 0$.

It turns out that even if μ does not satisfy $\mu \geq 0$ it behaves in many ways like a probability distribution. To formalize this idea we introduce the following notation for the formal expectation of a function $f\colon \{0,1\}^n \to \mathbb{R}$ with respect to another function μ (not necessarily corresponding to a probability distribution),

$$\tilde{\mathbb{E}}_{\mu} f = \sum_{x \in \{0,1\}^n} \mu(x) \cdot f(x)\,. \tag{8}$$

In order to emphasize the variable bound by the formal expectation, we use the notation $\tilde{\mathbb{E}}_{\mu(x)} f(x)$. This notation is useful when the expression $f(x)$ also depends on other variables.[5]

[5] This notation is analogous to the notation $\mathbb{E}_{x\sim\mu} f(x)$ for actual probability distributions, where $x \sim \mu$ denotes that x is a sample drawn form μ. We avoid this notation because the process of sampling is not well-defined in the context of formal expectations.

We define a "pseudo-distribution" to be a function μ such that the formal expectation with respect to μ satisfies some of the properties that expectations of probability distributions satisfy. However unlike actual probability distributions, pseudo-distributions may assign negative probabilities.

12. Definition (pseudo-distribution). A *degree-d pseudo-distribution over* $\{0,1\}^n$ is a function $\mu : \{0,1\}^n \to \mathbb{R}$ such that the formal expectation with respect to μ satisfies $\tilde{\mathbb{E}}_{\mu} 1 = 1$ and for every polynomial f of degree at most $d/2$,

$$\tilde{\mathbb{E}}_{\mu} f^2 \geq 0\,. \tag{9}$$

We refer to formal expectations with respect to degree d pseudo-distributions as degree d pseudo-expectations.

If a pseudo-distribution μ satisfies $\mu(x) \geq 0$ for all x then it corresponds to an actual probability distribution over the hypercube. Lemma 3 implies that every degree-$2n$ pseudo-distribution μ over $\{0,1\}^n$ satisfies $\mu \geq 0$.

Note that a priori a degree-d pseudo-distribution $\mu\colon \{0,1\}^n \to \mathbb{R}$ requires 2^n numbers to specify (i.e., the values of μ on all inputs). However, the following lemma allows us to reduce the number of parameters to $n^{O(d)}$.

13. Lemma (polynomial representation of pseudo-distributions). *Let μ be a degree-ℓ pseudo-distribution over $\{0,1\}^n$, there exists a multi-*

linear polynomial μ' of degree at most ℓ such that

$$\tilde{\mathbb{E}}_{\mu(x)} p = \tilde{\mathbb{E}}_{\mu'(x)} p \,, \tag{10}$$

for every p of degree at most ℓ.

Proof. Let $U_\ell \subseteq \mathbb{R}^{\{0,1\}^n}$ be the linear subspace of multilinear polynomials of degree at most ℓ. By Exercise 6 this subspace contains all polynomials of degree at most ℓ. Decompose the function μ as $\mu = \mu' + \mu''$ such that $\mu' \in U_\ell$ and μ'' is orthogonal to U_ℓ.[6] For every $p \in U_\ell$,

$$\tilde{\mathbb{E}}_{\mu} p = \langle \mu' + \mu'', p \rangle = \langle \mu', p \rangle = \tilde{\mathbb{E}}_{\mu'} p \,, \tag{12}$$

where we used the fact that μ'' is orthogonal to U_ℓ. □

[6] Here, orthgonality is with respect to the following inner product for real-valued functions on $\{0,1\}^n$,

$$\langle f, g \rangle = \sum_{x \in \{0,1\}^n} f(x) g(x) \,. \tag{11}$$

We can extend the notation of $\tilde{\mathbb{E}}_{\mu(x)} f(x)$ to the case that f is a *vector valued* function, in which case this denotes the vector obtained by taking expectation of every coordinate of f. Using this notation we can write the conclusion of Lemma 13 more succinctly as

$$\tilde{\mathbb{E}}_{\mu(x)} (1,x)^{\otimes \ell} = \tilde{\mathbb{E}}_{\mu'(x)} (1,x)^{\otimes \ell} \,, \tag{13}$$

where for an m-dimensional vector v and $d \in \mathbb{N}$, $v^{\otimes d}$ denotes the m^d dimensional vector such that $(v^{\otimes d})_{i_1,\ldots,i_d} = v_{i_1} \cdots v_{i_d}$. Indeed, every coordinate of $(1,x)^\ell$ is a polynomial of degree at most ℓ in x, and these coordinates form a basis for all these polynomials, and so if the expectations of $(1,x)^{\otimes \ell}$ under μ and μ' are equal then the expectation of every degree $\leq \ell$ polynomial p would be equal as well.

Exercises II

The following exercises are about basic properties of pseudo-distributions.

14. Exercise (high-degree pseudo-distributions). Show that every degree-$2n$ pseudo-distribution μ over $\{0,1\}^n$ satisfies $\mu(x) \geq 0$ for every $x \in \{0,1\}^n$. (Therefore, μ corresponds to an actual probability distribution over $\{0,1\}^n$.)

15. Exercise (pseudo-distributions and psd moments). Show that a function $\mu \colon \{0,1\}^n \to \mathbb{R}$ is a degree-d pseudo-distribution if and only if $\tilde{\mathbb{E}}_\mu 1 = 1$ and the following *pseudo-moment matrix* is positive semidefinite,

$$\tilde{\mathbb{E}}_{\mu(x)} \left((1,x)^{\otimes d/2} \right) \left((1,x)^{\otimes d/2} \right)^\intercal \succeq 0 \,. \tag{14}$$

16. Exercise (boundedness). Show that for every even d and every degree-d pseudo-distribution μ, there exists a degree-d pseudo-distribution μ' with the same pseudo-moments up to degree d such that for every $x \in \{0,1\}^n$,[7]

$$|\mu'(x)| \leq 2^{-n} \cdot \sum_{d'=0}^{d} \binom{n}{d'} . \tag{15}$$

(If μ' was an actual probability distribution, this inequality would mean that μ' has min-entropy at most $\approx \log(n^d)$.)

[7] **Hint:** This exercise might require some Fourier analysis.

17. Exercise (separation algorithm for pseudo-distributions). Show that the set of degree-d pseudo-distributions over $\{0,1\}^n$ admits a separation algorithm with running time $n^{O(d)}$. Concretely, show that there exists an $n^{O(d)}$-time algorithm that given a vector $N \in (\mathbb{R}^n)^{\otimes d}$ outside of the following set $\mathcal{X}_d$ outputs a halfspace that separates N from $\mathcal{X}_d$. Here, $\mathcal{X}_d$ is the set that consists of all coefficient vectors $M \in \left(\mathbb{R}^{n+1}\right)^{\otimes d}$ such that the function $\mu \colon \{0,1\}^n \to \mathbb{R}$ with $\mu(x) = \left\langle M, (1,x)^{\otimes d} \right\rangle$ is a degree-d pseudo-distribution over $\{0,1\}^n$.

18. Exercise (separation algorithm for pseudo-moments). Show that for every even $d \in \mathbb{N}$, the following set of pseudo-moments admits a separation algorithm with running time $n^{O(d)}$,

$$\mathcal{M}_d = \left\{ \tilde{\mathbb{E}}_{\mu(x)} (1,x)^{\otimes d} \;\middle|\; \mu \text{ is deg.-}d \text{ pseudo-distr. over } \{0,1\}^n \right\} . \tag{16}$$

Duality

We now show that pseudo-distributions are indeed dual to sos proofs by demonstrating that their existence certifies the non-existence of a proof and vice versa.

19. Theorem (Duality of sos certificates and pseudodistributions). *For every function $f \colon \{0,1\}^n \to \mathbb{R}$ and every even $d \in \mathbb{N}$, there exists a degree-d sos certificate for the non-negativity of f if and only if every degree-d pseudo-distribution μ over $\{0,1\}^n$ satisfies $\tilde{\mathbb{E}}_\mu f \geq 0$.*

Proof. One direction is immediate. Suppose f has a degree-d sos certificate so that $f = g_1^2 + \cdots + g_r^2$ for some polynomials $g_1, \ldots, g_r$ with $\deg g_i \leq d/2$. Then, every degree-d pseudo-distribution μ over $\{0,1\}^n$ satisfies

$$\tilde{\mathbb{E}}_\mu f = \tilde{\mathbb{E}}_\mu g_1^2 + \cdots + \tilde{\mathbb{E}}_\mu g_r^2 \geq 0 . \tag{17}$$

For the other direction, suppose that f is not contained in the degree-d sum-of-squares cone. By the hyperplane separation theorem, there

exists a halfspace H through the origin that conains the cone but not f. Let $\mu\colon \{0,1\}^n \to \mathbb{R}$ be the "normal" of H so that

$$H = \left\{ g\colon \{0,1\}^n \to \mathbb{R} \;\middle|\; \tilde{\mathbb{E}}_{\mu} g \geq 0 \right\}. \tag{18}$$

Since f is not contained in H, it satisfies $\tilde{\mathbb{E}}_{\mu} f < 0$. Since H contains the degree-d sos cone, every polynomial g of degree at most $d/2$ satisfies $\tilde{\mathbb{E}}_{\mu} g^2 \geq 0$. It remains to argue that $\tilde{\mathbb{E}}_{\mu} 1 > 0$, which means that we can rescale μ by a nonnegative factor to ensure that $\tilde{\mathbb{E}}_{\mu} 1 = 1$. Indeed, by Exercise 11, there exists $M \in \mathbb{R}_{\geq 0}$ such that $M + f$ has a degree-d sos certificate, which means that

$$\tilde{\mathbb{E}}_{\mu} 1 = \tfrac{1}{M} \cdot \left(\tilde{\mathbb{E}}_{\mu} M + f - \tilde{\mathbb{E}}_{\mu} f \right) > 0\,, \tag{19}$$

as desired. □

Sum-of-squares algorithm

Recall that we described the degree d sos algorithm as an algorithm that, given as input a polynomial $f\colon \{0,1\}^n \to \mathbb{R}$, runs in $n^{O(d)}$ and either outputs a certificate that $f(x) \geq 0$ for all x, or outputs an object that "pretends to be" a distribution over vectors $x \in \{0,1\}^n$ such that $f(x) < 0$. We now state this theorem formally.

20. Theorem (sum-of-squares algorithm). *For every even $d \in \mathbb{N}$, there exists an $n^{O(d)}$-time algorithm that given a polynomial $f\colon \{0,1\}^n \to \mathbb{R}$ of degree at most d (with polynomial bit length) either outputs a degree-d sos certificate for $f + 2^{-n}$ or a degree-d pseudo-distribution μ over $\{0,1\}^n$ such that $\tilde{\mathbb{E}}_{\mu} f < 2^{-n}$.*

Proof sketch. We will show one part of the theorem (about finding pseudo-distributions). The proof of the other part is similar but not needed for most of the algorithmic applications we will discuss. □

Suppose that f does not have a degree-d sos certificate. By the duality between sos certificates and pseudo-distributions, there exists a degree-d pseudo-distribution μ over $\{0,1\}^n$ such that $\tilde{\mathbb{E}}_{\mu} f < 0$. Our goal is to efficiently find a pseudo-distribution μ over $\{0,1\}^n$ such that $\tilde{\mathbb{E}}_{\mu} f < 2^{-n}$. Let v be a vector such that $f(x) = \langle v, (1,x)^{\otimes d} \rangle$. Then, $\tilde{\mathbb{E}}_{\mu} f = \langle v, \tilde{\mathbb{E}}_{\mu} (1,x)^{\otimes d} \rangle$. Therefore, we want to minimize the linear function $y \mapsto \langle v, y \rangle$ over the set $\mathcal{M}_d$ of vectors of the form $\tilde{\mathbb{E}}_{\mu} (1,x)^{\otimes d}$ for a degree-d pseudo-distribution μ over $\{0,1\}^n$. By Exercise 18, this set has a separation algorithm with running time $n^{O(d)}$. Using the ellipsoid algorithm, we can approximately minimize the linear function $y \mapsto \langle v, y \rangle$ over all $y \in \mathcal{M}_d$ also in time $n^{O(d)}$.

The different views of pseudo-distributions

Pseudo-distributions are not very complicated as a mathematical objects- they can be simply represented as positive semidefinite matrices. But they are rather subtle to grasp conceptually. (They are related, though not identical, to *quantum states* which are also modeled by positive semidefinite matrices and not easy to grasp conceptually.) An often useful point of view is to pretend that pseudo-distributions are actual distributions. This viewpoint can help "predict" certain properties of pseudo-distributions. For example, pseudo-distributions satisfy the Cauchy-Schwarz inequality:

21. Theorem (Cauchy-Schwarz for pseudodistributions). *If μ is a degree d pseudo-distribution and P, Q are polynomials of degree at most $d/2$ then*

$$\left(\tilde{\mathbb{E}}_{\mu} PQ\right)^2 \leq \left(\tilde{\mathbb{E}}_{\mu} P^2\right)\left(\tilde{\mathbb{E}}_{\mu} Q^2\right) \tag{20}$$

Proof. We may assume that both $\tilde{\mathbb{E}}_{\mu} P^2$ and $\tilde{\mathbb{E}}_{\mu} Q^2$ are strictly positive. (If at least one is zero, the proof is simpler.) By scaling P and Q by nonnegative scalars, we may further assume without loss of generality that

$$\tilde{\mathbb{E}}_{\mu} P^2 = \tilde{\mathbb{E}}_{\mu} Q^2 = 1\,. \tag{21}$$

It remains to prove $\tilde{\mathbb{E}}_{\mu} PQ \leq 1$. Indeed, $\tilde{\mathbb{E}}_{\mu}(P-Q)^2 \geq 0$ which means by linearity that

$$2\,\tilde{\mathbb{E}}_{\mu} PQ = \tilde{\mathbb{E}}_{\mu} P^2 + \tilde{\mathbb{E}}_{\mu} Q^2 - \tilde{\mathbb{E}}_{\mu}(P-Q)^2 \leq 2\,. \tag{22}$$

□

Do all pseudo-distributions correspond to actual distributions?

It turns out that the proofs of many of the inequalities we know and love, including Cauchy-Schwarz, Hölder and more, boil down to a sum-of-squares proof, which means that these statements hold not just for actual distributions but also for pseudo-distributions. In this light, a natural question to ask is whether perhaps every pseudo-distribution is an actual distribution. The answer to this question is negative.

22. Lemma (integrality gap). *There exists a degree-2 polynomial $f\colon \{0,1\}^n \to \mathbb{R}$ that is nonnegative $f \geq 0$ but has no degree-2 sum-of-squares certificate. In particular, there exists a degree-2 pseudo-distribution μ over $\{0,1\}^n$ such that $\tilde{\mathbb{E}}_{\mu} f < 0$.*

Proof. Consider the following nonnegative function on $\{0,1\}^3$,

$$f(x) = 2 - \left((x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_1)^2\right). \tag{23}$$

The fact that this function is nonnegative corresponds to the fact that the maximum cut in a 3-cycle is 2. Consider the degree-2 pseudo-distribution μ over $\{0,1\}^3$ with mean $\tilde{\mathbb{E}}_{\mu(x)} x^\intercal = \frac{1}{2}(1,1,1)$ and covariance,

$$\tilde{\mathop{\mathbb{E}}_{\mu(x)}} xx^\intercal - \left(\tilde{\mathop{\mathbb{E}}_{\mu(x)}} x\right)\left(\tilde{\mathop{\mathbb{E}}_{\mu(x)}} x\right)^\intercal = \frac{1}{8}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \tag{24}$$

Now,

$$\tilde{\mathop{\mathbb{E}}_{\mu(x)}} (x_1 - x_2)^2 = \tilde{\mathop{\mathbb{E}}_{\mu(x)}} (x_2 - x_3)^2 = \tilde{\mathop{\mathbb{E}}_{\mu(x)}} (x_3 - x_1)^2 = 3/4\,. \tag{25}$$

Therefore, f has negative expectation under μ,

$$\tilde{\mathop{\mathbb{E}}_{\mu}} f = 2 - 3 \cdot 3/4 = -1/4\,. \tag{26}$$

□

The Marley paradigm

It took about 80 years from the time Hilbert showed that polynomials that are not SOS exist non-constructively until Motzkin came up with an explicit example, and even that example has a low degree sos proof of positivity. One lesson from that is that if an inequality is non-negative and "natural" (i.e., constructed by methods known to Hilbert—not including probabilistic method), then heuristically there should be a low-degree sos proof for this fact. A corollary of this heuristic in the spirit of Bob Marley[8]:

> "If you analyze the performance of an SOS-based algorithm pretending pseudo-distributions are actual distributions, then unless you used Chernoff+union bound type arguments, then every little thing gonna be alright."

We will use Marley's corollary extensively in analyzing sos algorithms. There is a recurring theme in mathematics of "power from weakness". For example, we can often derandomize certain algorithms by observing that they fall in some restricted complexity classes and hence can be fooled by certain pseudorandom generator. Another example, perhaps closer to ours, is that even though the original way people defined calculus with "infinitesimal" amounts were based on false premises, still much of the results they deduced

[8] Bob Marley and the Wailers, "Three Little Birds" (1980).

were correct. One way to explain this is that they used a weak proof system that cannot prove all true facts about the real numbers, and in particular cannot detect if the real numbers are replaced with an object that does have such an "infinitesimal" quantity added to it. In a similar way, if you analyze an algorithm using a weak proof system (e.g., one that is captured by a small degree sos proof), then the analysis will still hold even if we replaced actual distributions with a pseudo-distribution of sufficiently large degree.

The quadratic sampling lemma

We have seen that not every pseudo-distribution is an actual distribution. However it turns out for every pseudo-distribution μ we can at least match the first two moments of μ by an actual probability distribution—albeit over $\mathbb{R}^n$ instead of $\{0,1\}^n$. The following lemma formalizes this idea which is related to hyperplane rounding in approximation algorithms and Gaussian copula in quantitative finance.

23. Lemma (Quadratic Sampling Lemma). *For every degree-2 pseudo-distribution μ over $\{0,1\}^n$, there exists a probability distribution ρ over $\mathbb{R}^n$ with the same first two moments, that is,*

$$\tilde{\mathbb{E}}_{\mu(x)} (1,x)^{\otimes 2} = \mathop{\mathbb{E}}_{x\sim\rho} (1,x)^{\otimes 2}\,. \tag{27}$$

Moreover, ρ is a multivariate Gaussian distribution.

Proof. Let $v = \mathbb{E}_{\mu(x)} x$ and $\Sigma = \tilde{\mathbb{E}}_{\mu(x)} (x-v)(x-v)^\intercal$ be the formal mean and covariance of μ. Like for an actual probability distribution, the covariance Σ of a degree-2 pseudo-distribution is positive semidefinite. Indeed, for every $u \in \mathbb{R}^n$,

$$\langle u, \Sigma u\rangle = \tilde{\mathbb{E}}_{\mu(x)} \langle u, x-v\rangle^2 \geq 0\,. \tag{28}$$

The following randomized procedure outputs a random vector y in $\mathbb{R}^n$ with mean v and covariance Σ:

- choose a standard Gaussian vector g, i.e., the coordinates of g are independently identically distributed Gaussian variables with mean 0 and variance 1,
- output the vector $y = v + \Sigma^{1/2} g$.

(In the last step, we use that the matrix Σ has a square root because it is positive semidefinite.) Since $\mathbb{E}\, g = 0$, the mean of this

distribution is $\mathbb{E}\, y = v$. Since $\mathbb{E}\, gg^\intercal = Id$, the distribution has covariance,

$$\mathbb{E}(y-v)(y-v)^\intercal = \Sigma^{1/2}\, \mathbb{E}\, gg^\intercal \Sigma^{1/2} = \Sigma\,. \tag{29}$$

The distribution we described is called the Gaussian distribution with mean v and covariance Σ and is denoted $N(v,\Sigma)$.[9] □

[9] The above sampling procedure shows that such a distribution $N(v,\Sigma)$ exists for every vector $v \in \mathbb{R}^n$ and every positive semidefinite matrix $\Sigma \in \mathbb{R}^{n\times n}$.

Pseudo-distributions as Bayesian probabilities

The problem of maximizing a polynomial over $\{0,1\}^n$ is NP hard (indeed Max-Cut is a special case of it), and so (assuming $P \neq NP$) if we run the sos algorithm with a small (e.g., constant) degree d then the algorithm should sometimes fail to solve it. In other words, on input some function $f\colon \{0,1\}^n \to \mathbb{R}$ the sos algorithm might return a pseudo-distribution μ that will not be an actual distribution over x's with $f(x) < 0$. How do we interpret this pseudo-distribution? One way to think about it is that the pseudo-distribution captures the *uncertainty* of a computationally bounded observer about the unknown x such that $f(x) < 0$. Bayesian probabilities are often used to capture uncertainty even about events that are completely determined, for example, it might make sense for me to say something like "the probability that my great-grandfather had blue eyes is 25%" since even though obviously he either did or didn't have blue eyes, the information I have about this fact can still leave me with some uncertainty.

The fact that we have bounded computational powers means that we sometimes have uncertainty even about facts that are completely determined by the information we are given. For example, while the number $2^{81712357} - 1$ is either prime or composite, the authors (and as far as we know, everyone else) do not know which of the two cases holds. In fact, the information gathered by the Great Internet Mersenne Prime search project only allows us to determine that the probability that this number is prime is roughly $1.46 \cdot 10^{-6}$.

Similarly, even if a function $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique x such that $f(x) < 0$, this value x might be hard to find, and so we could have some *uncertainty* about it. One way to think about the pseudo-distribution is that it captures this uncertainty, and so a statement such as $\tilde{\mathbb{E}}\, x_{17} = 0.7$ can be interpreted as saying that, given the information we have, the probability that $x_{17} = 1$ is 0.7.

What's next?

The type of questions we are interested in regarding the sos algorithm are the following:

- For what families of problems does the sos algorithm give us the best-known guarantees? Are there families of problems for which it is reasonable to conjecture that the sos algorithm is *optimal*, in the sense that for any given d, there is no other algorithm running much faster than n^d time that would do better than the degree d sos algorithm?
- There are some a priori seemingly stronger algorithms and proof systems, such as the "dynamic sos" proof system. Can we show natural classes of problems on which sos matches the guarantees of those seemingly stronger systems?
- Can we use the sos algorithm to solve problems that have eluded us via other means? In particular, there are some average case problems arising in machine learning, statistical physics, and other areas for which the sos algorithm seems promising. There are also some very fascinating worst-case problems for which we do not know the sos algorithm's performance and which resolving could settle important questions such as Khot's unique games conjecture.
- Can we obtain a *systematic understanding* of the sos algorithm's performance? Ideally we would have a "creativity free" analysis, whereby we reduce the question of analyzing the guarantees sos gives on any particular question to some potentially complicated or tedious but ultimately doable and non-creative calculations .

References