SUM-OF-SQUARES

Amit Rajaraman

Last updated December 20, 2022

Contents

§0. Notation

Given $n \times n$ matrices A, B, denote $\langle A, B \rangle = \mathrm{Tr}(AB) = \sum_{i,j} A_{ij} B_{ij}$.

§1. Fundamentals

1.1. Introduction

The sum-of-squares technique, at its most basic form, is a way of determining whether for some polynomial p over $\mathbb{R}^n$, $p(x) \geq 0$ for x in some base set. For now, suppose that our "base set" is $\{0, 1\}^n$. Elegantly, it manages to convert *disproofs* of such inequalities to *algorithms* to determine a point where $p(x) < 0$.
More concretely, we shall show non-negativity by expressing p as a *sum of squares* of *low degree* polynomials (while low degree is not technically required, it makes the converted algorithm efficient).

Definition 1.1 (Sum-of-squares proof)**.** Given a polynomial f in variables $x_1, \ldots, x_n$, a *degree d sum-of-squares proof* or *degree d sum-of-squares certificate* (abbreviated SoS proof or SoS certificate) of $f \geq 0$ is a set $\{g_1, \ldots, g_m\}$ of polynomials of degree at most $d/2$ such that

$$f(x) = \sum_{i=1}^{m} g_i^2(x) \tag{1.1}$$

for all x. If f has a degree d sum-of-squares certificate, we write that

$$\vdash_d f(x) \geq 0.$$

Let $\mathcal{A}$ be a set of constraints of the form $A_i(x) = 0$ or $B_j(x) \geq 0$ for $i \in [k], j \in [\ell]$. Then, an *degree d SoS proof given $\mathcal{A}$* of $f \geq 0$ is a set $\{g_1, \ldots, g_m\}$ of polynomials of degree at most $d/2$ such that (1.1) holds for all x satisfying the constraints in $\mathcal{A}$. If such a set exists, we write

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

We always assume that d in this context is even.
Note that simple set restrictions can be captured by the set of constraints. In particular, we can check restrict ourselves to the boolean hypercube $\{-1, 1\}^n$ by having $\mathcal{A}$ contain $x_i^2 = 1$ for all i. Note that the set of functions with degree d SoS proofs of non-negativity forms a closed convex cone.

Proposition 1.2. Any non-negative $f : \{-1, 1\}^n \to \mathbb{R}$ has a degree $2n$ sum-of-squares proof.

Proof. Recall that any function $h : \{-1, 1\}^n \to \mathbb{R}$ can be expressed as a polynomial of degree at most n as

$$h(x) = \sum_{S \subseteq [n]} \hat{f}(S) x_S,$$

where $x_S = \prod_{i \in S} x_i$ with the convention $x_\varnothing = 1$. Knowledgeable readers may recognize this as the *Fourier expansion* of h – we omit the details of why such an expansion exists, but refer the reader to the excellent text by O'Donnell [O'D14] for more details. In particular, $\sqrt{f}$ is a polynomial of degree at most n, so squaring both sides we get that f has a degree $2n$ SoS proof. ∎

The above is *not* true in general; not every non-negative polynomial $f : \mathbb{R}^n \to \mathbb{R}$ can be written as a sum of squares.

Definition 1.3. Given a vector $y \in \mathbb{R}^n$, the vector $y^{\otimes k} \in \mathbb{R}^{n^d}$ has entries indexed by elements of $[n]^d$, with the αth entry being $\prod_{j \in d} y_{\alpha_j}$. Also denote $v_k(x)$ to be the size $\binom{n+k}{k}$ vector with entries equal to all the monomials of x of degree at most k.

Note that for $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$, any monomial of degree at most $d/2$ appears in the vector $(1, x)^{\otimes d/2}$, where $(1, x) = (1, x_2, \ldots, x_n) \in \mathbb{R}^{n+1}$. Also recall that a matrix A is said to be positive semidefinite, denoted $A \succeq 0$, if $x^\top A x \geq 0$ for all vectors x, which is equivalent to asserting that all eigenvalues of the matrix are non-negative.

Proposition 1.4. Let f be a polynomial. f has a degree d sum-of-squares proof iff there exists $A \succeq 0$ such that

$$f(x) = \langle v_{d/2}(x), A v_{d/2}(x) \rangle. \tag{1.2}$$

Proof. For the forward direction, suppose that $f = \sum_{i=1}^m g_i^2$, with $g_i(x) = v_i^\top v_{d/2}(x)$ by writing it out in the monomial basis. Then,

$$f(x) = \sum_{i=1}^m v_{d/2}(x)^\top v_i v_i^\top v_{d/2}(x)$$

$$= \left\langle v_{d/2}(x), \underbrace{\sum_{i=1}^m v_i v_i^\top}_{A} v_{d/2}(x) \right\rangle.$$

The backward direction is straightforward by decomposing A as $\sum \lambda_i v_i v_i^\top$, where each $\lambda_i \geq 0$, and observing that each $v_i^\top v_{d/2}(x)$ is a polynomial of degree at most $d/2$. ∎

As a corollary, this implies that if an f has a degree d SoS proof, it has one with at most $\binom{n+d}{d}$ squares. Also note that eq. (1.2) is *linear* in the elements of A.

If we bump up a function by enough, we can ensure non-negativity. It turns out that we can do the same to ensure SoS-ness.

Lemma 1.5. Let $f : \{-1, 1\}^n \to \mathbb{R}$ be any function of degree at most d. For sufficiently large L, $L + f$ has a degree d SoS certificate.

Proof. Note that for any S, $1 + x_S \geq 0$ has a degree $\lceil |S|/2 \rceil$ SoS proof. Indeed, setting $S = T_1 \sqcup T_2$ for T_1, T_2 of (almost) equal size, $1 + x_S = \frac{1}{2}(x_{T_1} + x_{T_2})^2$. Similarly, $1 - x_S$ has a degree $|S|$ SoS proof as well. Therefore,

$$\sum_{|S| \leq d} |\hat{f}(S)| + \sum_{|S| \leq d} \hat{f}(S) x_S = \sum_{|S| \leq d} |\hat{f}(S)|(1 + \text{sign}(\hat{f}(S)) x_S)$$

has a degree d SoS certificate, so the statement is true with $L = \sum_{|S| \leq d} \hat{f}(S)$. ∎

1.2. Semidefinite Programming

The reader is likely familiar with *linear programming*, where we are interested in

$$\min_{x \in \mathcal{P}} c^\top x, \text{ where } \mathcal{P} = \{x \geq 0 : Ax = b\}.$$

Although a linear program may in general have inequalities in the constraints, we may merge these into the $x \geq 0$ condition by introducing slack variables (if we have $\sum_i a_i x_i \geq 0$, we may add a non-negative variable y and make it $\sum_i a_i x_i - y = 0$). In *semidefinite programming*, the setting is mostly the same, albeit with the minor change that we represent the variables by a matrix instead of a vector, and we additionally have that this matrix is positive semidefinite. More concretely, denoting

$$\langle C, X \rangle = \sum_{i,j} C_{ij} X_{ij},$$

we are interested in

$$\min_{X \in \mathcal{S}} \langle C, X \rangle, \text{ where } \mathcal{S} = \{X \succeq 0 : \langle A_i, X \rangle = b_i \text{ for } i \in [m]\}.$$

We interchangeably use $\mathcal{S}$ to denote the set of constraints and the corresponding body. Proposition 1.4 suggests a link between SoS proofs and SDPs. A natural question is: can we solve SDPs efficiently?

Note that the set of all PSD matrices X forms a convex cone. In combination with the linear constraints, the intersection of this cone and the affine subspace form a so-called "spectrahedron", which we would like to minimize our quantity over. Note that any linear program is a semidefinite program, by enforcing that all off-diagonal elements of the matrix are zero. To answer our earlier question, it turns out that we cannot solve SDPs exactly.[1] However, if we enforce certain structural restrictions, we can solve them approximately (up to a small additive error).

Definition 1.6 (Separation Oracle). For a convex body $K \subseteq \mathbb{R}^n$, a (*strong*) *separation oracle* for K does the following given as input any $x \in K$.

1. if $x \in K$, it returns yes.

2. if $x \notin K$, it returns no, and in addition a vector a and real b such that $\langle a, y \rangle \geq b$ for all $y \in K$ and $\langle a, x \rangle < b$ – this is a so-called "separating hyperplane" that separates x and K.

More generally, we can efficiently minimize an inner product over a convex (bounded) body up to an additive error of ϵ, given an efficient weak separation oracle.

Theorem 1.7. Let $f : \{-1, 1\}^n \to \mathbb{R}$ have a degree d sum-of-squares proof of non-negativity. Then, for $\epsilon > 0$, there exists an algorithm that finds a sum-of-squares proof of $f + \epsilon$ in $\text{poly}(n^d, \log(1/\epsilon))$.

The high-level idea of the algorithm is as follows.

We first solve the "feasibility problem" of finding a point in a body K, given that $B(c, r) \subseteq K \subseteq B(0, R)$. We begin by setting $\mathcal{E}^{(0)} = B(0, R)$. Given the ellipsoid $\mathcal{E}^{(i)}$, if its center returns yes, we return the point itself. Otherwise, we use the separating hyperplane to get a halfspace H in which K is contained, and set $\mathcal{E}^{(i+1)}$ to be the smallest ellipsoid containing $\mathcal{E}^{(i)} \cap H$. This algorithm runs in $\text{poly}(n, \text{size}(K)) \log(R/r)$ – the proof amounts to showing that the volume of the ellipsoid decreases by a factor of at least $\exp(1/2(n+1))$ at each stage, and we have a lower bound on the volume of K by $\text{vol}(B(0, r))$.

We can slightly modify this algorithm to one that approximately solves the optimization version of maximizing $c^\top x$ as well. Once we get a point α in the body, we begin looking at $K \cap \{x : c^\top x > c^\top \alpha\}$ and repeat the feasibility algorithm. This is repeatedly done until we can guarantee that we are within ϵ of the optimum. The only non-trivial part of this algorithm is showing that we can use the oracle to construct an oracle for the new body $K \cap \{x : c^\top x > c^\top \alpha\}$.

To complete the connection to SDPs, we require that the SDP constraints $\mathcal{S}$ admits an efficient weak separation oracle; we omit the details of this. Next, we require that the body $\mathcal{S}$ contains a ball and is contained in a ball. The former is not true in general because the constraints typically make our body lower-dimensional (a subspace). To get around this, we introduce an additive error in each the contraints, so the new constraints are $|\langle A, X \rangle - b_i| \leq \epsilon$ for each i. In this case, there is a ball of radius $O((\epsilon/\|A\|_F)^n)$ contained in the body, where $\|A\|_F^2 = \sum_{i,j,k} (A_i)_{jk}^2$.

In our context of finding X such that $f(x) = v_{d/2}(x)^\top X v_{d/2}(x)$, we know that $\|A\|_F^2 \leq \text{Tr}(A)^2 \leq \hat{f}(\varnothing)^2$, so the body is bounded as well.

Like how LPs have duals, so do SDPs. If we have the primal

$$\min_{X \in \mathcal{S}} \langle C, X \rangle, \text{ where } \mathcal{S} = \left\{X \succeq 0 : \langle A_i, X \rangle = b_i \text{ for } i \in [m]\right\},$$

---

[1]It is not even known if this is in NP! It is known that it is in PSPACE however.

its dual is

$$\max_{y \in \mathcal{S}^D} b^\top y, \text{ where } \mathcal{S}^D = \left\{ S \succeq 0 : C - \sum_{i=1}^m y_i A_i = S \right\}.$$

The PSDness condition in the dual just says that $C \succeq \sum_{i=1}^m y_i A_i$.

Proposition 1.8 (Weak Duality). Let X and y be solutions to the primal and dual SDPs respectively. Then, $\langle C, X \rangle \geq b^\top y$.

Proof. We have

$$\langle C, X \rangle = \left\langle \sum_{i=1}^m y_i A_i + S, X \right\rangle$$
$$= \sum_{i=1}^m y_i \langle A_i, X \rangle + \langle S, X \rangle$$
$$= b^\top y + \langle S, X \rangle \geq b^\top y.$$

The final inequality requires showing that if $S, X \succeq 0$, then $\langle S, X \rangle \geq 0$ – this is a simple corollary of the Schur Product Theorem we shall see later, using $\mathbf{1}^\top (S \circ X) \mathbf{1} \geq 0$. ∎

In linear programming, we have strong duality which asserts that the two optima are in fact *equal*. However, in SDPs, some mild conditions are required for this to be true.

Theorem 1.9 (Strong duality). Let $\mathcal{S}$ be the set of constraints of a primal SDP and $\mathcal{S}^D$ the set of constraints in its dual, such that the two have optima α^*, β^*. Then, $\langle C, \alpha^* \rangle = \langle b, \beta^* \rangle$ if

1. the spectrahedron $\mathcal{S}$ is non-empty and there exists β such that $\sum_{i \in [m]} \beta_i A_i - C \succ 0$, or

2. the spectrahedron $\mathcal{S}^D$ is non-empty and there exists $\alpha \succ 0$ such that $\langle A, \alpha \rangle = b_i$ for all $i \in [m]$.

As a corollary, one may show that $\langle C, \alpha^* \rangle = \langle b, \beta^* \rangle$ if the set of optimal solutions of either of the two SDPs is non-empty and bounded.

We omit the (rather involved) proof of the above.

1.3. Pseudoexpectations

Let us again restrict ourselves to $\{-1, 1\}^n$ for a while. We have established one link between SoS proofs and SDPs, and now we shall establish another link between them and the following.

Definition 1.10 (Pseudodistribution). A *degree d pseudodistribution* is a function $\mu : \{-1, 1\}^n \to \mathbb{R}$ such that the expectation operator $\widetilde{\mathbb{E}}_\mu$ defined by $\widetilde{\mathbb{E}}_\mu f = \sum_{x \in \{-1,1\}^n} f(x) \mu(x)$ satisfies

(a) $\widetilde{\mathbb{E}}_\mu 1 = 1$, and

(b) for all f of degree at most $d/2$, $\widetilde{\mathbb{E}}_\mu f^2 \geq 0$.

In this case, $\widetilde{\mathbb{E}}_\mu$ is called a *pseudoexpectation*.

Analogous to Proposition 1.2, we get that any degree $\geq 2n$ pseudodistribution is an actual distribution, in the sense that $\mu \geq 0$. Analogous to Proposition 1.4, we get the following.

Proposition 1.11. $\widetilde{\mathbb{E}}$ is a degree d pseudoexpectation iff

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

(b) $\widetilde{\mathbb{E}}v_{d/2}(x)v_{d/2}(x)^\top \succeq 0$.

Proof. Note that for any vector $(\hat{f})$ of Fourier coefficients of a degree $\leq d/2$ function $f : \{-1, 1\}^n \to \mathbb{R}$ (so $f(x) = \hat{f}^\top v_{d/2}(x)$),

$$\widetilde{\mathbb{E}}f^2 = \widetilde{\mathbb{E}}\left(\sum_{|S| \leq d} \hat{f}(S)x_S\right)^2$$
$$= \widetilde{\mathbb{E}}\hat{f}^\top v_{d/2}(x)v_{d/2}(x)^\top \hat{f}$$
$$= \hat{f}^\top \left(\widetilde{\mathbb{E}}v_{d/2}(x)v_{d/2}(x)^\top\right) \hat{f}.$$

To conclude, note that $\widetilde{\mathbb{E}}v_{d/2}(x)v_{d/2}(x) \succeq 0$ iff $\hat{f}^\top \left(\widetilde{\mathbb{E}}v_{d/2}(x)v_{d/2}(x)^\top\right)\hat{f} \geq 0$ for all vectors $\hat{f}$. ∎

Given any function that is not non-negative everywhere, there exists some distribution μ such that $\mathbb{E}_\mu f < 0$. Ideally, we would like a similar result in order to distinguish between functions that have SoS certificates of degree d and those that don't.

Theorem 1.12. f has a degree d sum-of-squares proof iff for all degree d pseudoexpectations $\widetilde{\mathbb{E}}$, $\widetilde{\mathbb{E}}f \geq 0$.

Equivalently, f does not have a degree d sum-of-squares proof iff there exists a degree d pseudoexpectation $\widetilde{\mathbb{E}}$ such that $\widetilde{\mathbb{E}}f < 0$.

Proof. The forward direction is straightforward by Definition 1.10(b). For the backward direction, suppose instead that f does not have a degree d SoS proof. Then, there exists a separating hyperplane between f and this set, that is, some degree d pseudoexpectation $\widetilde{\mathbb{E}}$ such that $\widetilde{\mathbb{E}}f < 0$. If we manage to show that $\widetilde{\mathbb{E}}1 > 0$, we are done since we can then rescale μ to make it exactly equal to 1. By Lemma 1.5, we have $L > 0$ such that $\widetilde{\mathbb{E}}(f + L) \geq 0$. Since $\widetilde{\mathbb{E}}f < 0$, this means that $\widetilde{\mathbb{E}}L = L \cdot \widetilde{\mathbb{E}}1 > 0$, completing the proof. ∎

Using our earlier discussion, given a function f without a degree d SoS certificate of positivity, we may find in $\text{poly}(n^d, 1/\epsilon, \text{size}(f))$ time a pseudoexpectation $\widetilde{\mathbb{E}}$ such that $\widetilde{\mathbb{E}}f < \epsilon$.

§2. Quadratic optimization on the hypercube

2.1. Max-cut

In this subsection, let us describe how the content of the previous three subsections interact through an example, and give an approximation algorithm for the max-cut problem.

Question. Given a graph $G = (V, E)$, find $S \subseteq V$ such that the size of the cut $E(S, S^c) = \left\{\{i, j\} \in E : i \in S, j \in S^c\right\}$ is maximized.

Unlike min-cut, which may be solved in polynomial time using flow, the above is NP-complete.

One basic approximation algorithm was proposed by Erdős, which merely returns a random cut. With constant probability, the returned cut is a 1/2-approximation of the max-cut. We shall in this algorithm study an algorithm due to Goemans and Williamson [GW00].

Assume wlog that $V = [n]$, and identify any $S \subseteq V$ with the vector in $\{-1, 1\}^n$ with a 1 at precisely those vertices in S. Note that the function defined by

$$f_G(x) = \frac{1}{4} \sum_{ij \in E} (x_i - x_j)^2 = \frac{1}{2} \sum_{ij \in E} (1 - x_i x_j).$$

on input S returns precisely the size of the cut corresponding to S. Equivalently, considering the *graph Laplacian* $L_G := D_G - A_G$, where D_G is the diagonal matrix of degrees and A_G is the adjacency matrix, we have

$$f_G(x) = \frac{1}{4} x^\top L_G x. \tag{2.1}$$

We are interested in $\max_{x \in \{-1,1\}^n} f_G(x) =: \mathsf{opt}(G)$.

Theorem 2.1. Set $\alpha_{\mathrm{GW}} := \min_{\rho \in [-1,1]} \frac{2 \arccos(\rho)}{\pi(1-\rho)} \approx 0.8786$. Then,

$$\frac{\mathsf{opt}(G)}{\alpha_{\mathrm{GW}}} - f_G(x) \geq 0$$

has a degree 2 sum-of-squares certificate.

Let $\widetilde{\mathbb{E}}_{\mathsf{opt}}$ be a pseudoexpectation that maximizes $\widetilde{\mathbb{E}}_{\mathsf{opt}} f_G$ as $\mathsf{opt}_{\mathrm{SOS}_2}(G)$. Clearly, $\mathsf{opt}_{\mathrm{SOS}_2}(G) \geq \mathsf{opt}(G)$. Furthermore, by the previous theorem,

$$\mathsf{opt}(G) \leq \mathsf{opt}_{\mathrm{SOS}_2}(G) \leq \frac{1}{\alpha_{\mathrm{GW}}} \mathsf{opt}(G).$$

By the discussion at the end of the previous subsection, we can find in $\mathrm{poly}(n, 1/\epsilon)$ a degree 2 pseudodistribution μ such that

$$\widetilde{\mathbb{E}}_\mu f_G \geq \mathsf{opt}_{\mathrm{SOS}_2}(G) - \epsilon.$$

So,

$$\frac{1}{\alpha_{\mathrm{GW}}} \mathsf{opt}(G) \geq \widetilde{\mathbb{E}}_\mu f_G \geq \mathsf{opt}(G) - \epsilon.$$

Lemma 2.2. Let μ be a degree 2 pseudodistribution on $\{-1, 1\}^n$. Then, there exists a ("real") distribution μ' on $\{-1, 1\}^n$ such that

$$\mathbb{E}_{\mu'} f_G \geq \alpha_{\mathrm{GW}} \cdot \widetilde{\mathbb{E}}_\mu f_G.$$

Further, it is possible to efficiently sample from μ' given μ. Plugging this back into our previous sequence of equations,

$$\mathbb{E}_{\mu'} f_G \geq \alpha_{\mathrm{GW}}(\mathsf{opt}(G) - \epsilon) \geq (\alpha_{\mathrm{GW}} - \epsilon)\mathsf{opt}(G),$$

and efficient sampling implies that we can in $\mathrm{poly}(n, 1/\epsilon)$ time sample a random cut S such that with good probability, the size of the cut of S is a $(\alpha_{\mathrm{GW}} - \epsilon)$-approximation of the max-cut.

Let us now get to the proofs of the above results.

Proof that Lemma 2.2 implies Theorem 2.1. It suffices to show that for all pseudodistributions $\widetilde{\mathbb{E}}_\mu$,

$$\widetilde{\mathbb{E}}_\mu \left[\frac{\mathsf{opt}(G)}{\alpha_{\mathsf{GW}}} - f_G \right] \geq 0.$$

Equivalently, we would like to show that

$$\widetilde{\mathbb{E}}_\mu f_G \leq \frac{\mathsf{opt}(G)}{\alpha_{\mathsf{GW}}}.$$

Letting μ' be a distribution as in Lemma 2.2,

$$\widetilde{\mathbb{E}}_\mu f_G \leq \frac{1}{\alpha_{\mathsf{GW}}} \mathbb{E}_{\mu'} f_G \leq \frac{1}{\alpha_{\mathsf{GW}}} \mathsf{opt}(G). \qquad \blacksquare$$

Proof of Lemma 2.2. We may assume wlog that $\widetilde{\mathbb{E}}_\mu x = 0$, by changing $\mu(x)$ to $\frac{\mu(x) + \mu(-x)}{2}$ – note that this procedure does not change $\widetilde{\mathbb{E}}_\mu f_G$ because $f_G(x) = f_G(-x)$. Using Proposition 1.11(b) and recalling that any principal submatrix of a PSD matrix is PSD, $\widetilde{\mathbb{E}}_\mu x x^\top \succeq 0$. So, let ν be a normal distribution on $\mathbb{R}^n$ with mean 0 and covariance matrix $\widetilde{\mathbb{E}}_\mu x x^\top$. Finally, define μ' by the process that samples a vector g according to ν and returns $\hat{x} = \mathrm{sign}(g)$, the vector in $\{-1,1\}^n$ whose ith coordinate is just the sign ± 1 of g_i – this is well-defined with probability 1. Note that an $(x_i - x_j)^2$ term in f_G contributes to the cut iff $\mathrm{sign}(g_i) \neq \mathrm{sign}(g_j)$. That is,

$$\mathbb{E}_{\mu'} f_G = \sum_{ij \in E} \Pr[\mathrm{sign}(g_i) \neq \mathrm{sign}(g_j)].$$

For distinct i, j, set $\rho_{ij} = \mathbb{E}_\mu[x_i x_j] = \mathbb{E}[g_i g_j]$. Let $h \sim \mathcal{N}(0, \mathrm{Id}_2)$. Then, to analyze the above probability (for a fixed i, j), we can assume that $g_i = \langle h, v \rangle$ and $g_j = \langle h, w \rangle$ for some v, w such that $\langle v, w \rangle = \rho_{ij}$. So, $\mathrm{sign}(g_i) \neq \mathrm{sign}(g_j)$ iff $\langle h, v \rangle$ and $\langle h, w \rangle$ have opposite signs. Because the "direction" $h/\|h\|$ of h is uniformly distributed on $\mathbb{S}^1$, we get that

$$\Pr[\mathrm{sign}(g_i) \neq \mathrm{sign}(g_j)] = \frac{\arccos(\rho_{ij})}{\pi},$$

as seen in the following diagram, where h must lie in the green region for the signs to be different.
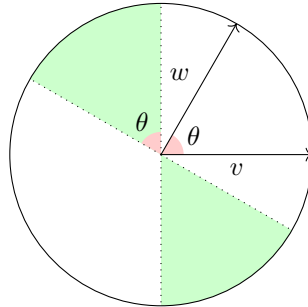
The angle between v, w is $\theta = \arccos(\rho_{ij})$.

Using the facts that $\mathbb{E}[g_i^2] = 1$ and $\mathbb{E}[g_i g_j] = \rho_{ij}$, we have that $\mathbb{E}[(g_i - g_j)^2] = 2(1 - \rho_{ij})$ and so

$$\mathbb{E}_{\mu'} f_G = \sum_{ij \in E} \Pr[\mathrm{sign}(g_i) \neq \mathrm{sign}(g_j)]$$

$$= \sum_{ij \in E} \frac{\arccos(\rho_{ij})}{2\pi(1 - \rho_{ij})} \mathbb{E}[(g_i - g_j)^2]$$

$$\geq \frac{\alpha_{\mathsf{GW}}}{4} \cdot \mathbb{E}\left[\sum_{ij \in E} (g_i - g_j)^2 \right] = \alpha_{\mathsf{GW}} \widetilde{\mathbb{E}}_\mu f_G. \qquad \blacksquare$$

Now, we have managed to get roughly an α_{GW}-approximation using degree 2 SoS. Is it possible to do any better using degree 2 SoS? What about with a higher (but constant) degree? It might even be interesting to see if we can get a better approximation with non-constant degree, say $O(\log n)$.

To answer the first question, it turns out that what we have done is indeed the best possible. A strong result regarding the constant degree case is due to Khot-Kindler-Mossel-O'Donnell [KKMO07], where it is proved that Khot's *Unique Games Conjecture* (UGC) [Kho02] implies that an $(\alpha_{\text{GW}} + \epsilon)$-approximation is NP-hard for any $\epsilon > 0$. While it is unknown at the time of writing this whether the unique games conjecture is true, we have strong reason to believe that it is due to a recent result [DKK+18] which proves the "2-to-2 Games Conjecture", a close variant of the unique games conjecture. We shall later look at the UGC in more detail.

Let us get back to the Goemans-Williamson algorithm. Instead of looking at the best approximation ratio, can we parametrize the output result in terms of the optimal value?

Proposition 2.3. Let G be a graph with $\text{opt}(G) = (1-\delta)|E|$. For the output distribution μ' of the Goemans-Williamson algorithm, $\mathbb{E}_{\mu'} f_G = (1 - O(\sqrt{\delta}))|E|$.

Proof. *** INCOMPLETE ***   ∎

Is this rounding we have done, called "Gaussian rounding", the best possible? It turns out that it is not, and we can in general do better using the "RPR^2" scheme of roundings. We shall soon study this in more detail.

Let us now return to our earlier statement that it is impossible to do better using degree 2 SoS. That is, for graphs in general, if we can get a degree 2 SoS certificate of non-negativity for

$$\frac{\text{opt}(G)}{c} - f_G(x),$$

how large can c be? We shall show that $c = \alpha_{\text{GW}}$ is optimal, by looking at the cycle C_n for odd n. This serves as a "gap" example. It is easily seen that the max-cut in this graph is $n - 1 = \left(1 - \frac{1}{n}\right)|E|$. We shall show that there exists a degree 2 pseudodistribution μ such that

$$\widetilde{\mathbb{E}}_\mu f_{C_n} \geq \left(1 - O\left(\frac{1}{n^2}\right)\right)|E|.$$

Due to Proposition 2.3, this shows that the Goemans-Williamson algorithm is tight, at least up to constant factors. We can think of the cycle as something of a discretization of the 2-dimensional sphere. If we instead look at the discretization of a high-dimensional sphere, it can be shown that this is tight even up to constant factors. We refer the reader to [FS02] for details. The sketch of the proof for the cycle is as follows.

Recall eq. (2.1), so we are interested in $\max_{x \in \{-1,1\}^n} x^\top L_G x$. This is at most $\max_{x:\|x\|_2 = \sqrt{n}} x^\top L_G x = n\|L_G\|_2$, which can be computed in polynomial time. Now, how do we construct a pseudodistribution $\widetilde{\mathbb{E}}$ for the cycle as mentioned earlier? Note that a given $\widetilde{\mathbb{E}}$ is a well-defined degree 2 pseudodistribution iff $\widetilde{\mathbb{E}}(1,x)(1,x)^\top$ is a PSD matrix with 1s on the diagonal. Now,

$$\begin{aligned}
\widetilde{\mathbb{E}} f_G(x) &= \widetilde{\mathbb{E}} x^\top L_G x \\
&= \widetilde{\mathbb{E}} \langle L_G, xx^\top \rangle \\
&= \left\langle L_G, \widetilde{\mathbb{E}} xx^\top \right\rangle.
\end{aligned}$$

Observe that the top eigenvalue of L_G is indeed $1 - O(1/n^2)$, and this eigenspace is 2-dimensional. It turns out that for an appropriate choice of v_1, v_2 in this eigenspace, we can ensure that $v_1 v_1^\top + v_2 v_2^\top$ does have only 1s on the diagonal (this is essentially a consequence of the fact that $\sin^2 \theta + \cos^2 \theta = 1$).

2.2. The positive semidefinite case

In the previous subsection, we looked at $\max_{x \in \{-1,1\}^n} x^\top L_G x$, where L_G is the (positive semidefinite) Laplacian of a graph. This is an example of quadratic optimization, where we are more generally interested in

$$\mathsf{opt}(B) \coloneqq \max_{x \in \{-1,1\}^n} x^\top B x$$

for some $n \times n$ matrix B.

In the case where $B \succeq 0$, it turns out that we can do something similar to what we had done in the max-cut algorithm.

Theorem 2.4 (Nesterov). Let B be a positive semidefinite $n \times n$ matrix. Then,

$$\frac{\mathsf{opt}(B)}{c} - x^\top B x$$

has a degree 2 sum-of-squares certificate for $c = 2/\pi \approx 0.63$.

By the discussion in the previous section, this means as a corollary that we have a $\mathrm{poly}(n, 1/\epsilon)$-time $(2/\pi - \epsilon)$-approximation algorithm for any $\epsilon > 0$.

Definition 2.5. Let $M \in \mathbb{R}^{n \times n}$. Given $f : \mathbb{R} \to \mathbb{R}$, define $f[M] \in \mathbb{R}^{n \times n}$ by $f[M]_{ij} = f(M_{ij})$ for all i, j.

Proposition 2.6. Suppose M is a positive semidefinite matrix and f a function whose Taylor series has all positive Taylor coefficients and is uniformly convergent on $[-1, 1]$. Then, $f[M]$ is positive semidefinite.

The above is a corollary of the following simple observation.

Proposition 2.7 (Schur Product Theorem). Let M, M' be positive semidefinite matrices. Denote by $M \circ M'$ the *Hadamard product* of M, M' defined by $(M \circ M')_{ij} = M_{ij} M'_{ij}$. Then, $M \circ M'$ is positive semidefinite.

Proof. Let $M = \sum_i \lambda_i v_i v_i^\top$ and $M' = \sum_j \lambda'_j v'_j v'^\top_j$. Using linearity of the Hadamard product,

$$M \circ M' = \sum_{i,j} \lambda_i \lambda'_j (v_i v_i^\top) \circ (v_j v'^\top_j) = \sum_{i,j} \lambda_i \lambda'_j (v_i \circ v'_j)(v_i \circ v'_j)^\top \succeq 0. \qquad \blacksquare$$

Proof of Proposition 2.6. Denote $[M]^2 = M \circ M$, and $[M]^i = [M]^{i-1} \circ M$ more generally. By the Schur Product Theorem, $[M]^i \succeq 0$ for all i. Therefore, $\sum c_i [M]^i \succeq 0$, that is, $(\sum c_i x^i)[M] \succeq 0$. $\qquad \blacksquare$

Proof of Nesterov. As in the previous subsection, let μ be a zero mean degree 2 pseudodistribution on $\{-1,1\}^n$, g a normal random variable with zero mean and covariance matrix $\widetilde{\mathbb{E}}_\mu x x^\top$, and $\hat{x} \coloneqq \mathrm{sign}(g)$ distributed as μ'. A straightforward byproduct of the final part of the proof of Lemma 2.2 is that

$$\mathbb{E}_{\mu'}[\hat{x}_i \hat{x}_j] = \frac{2}{\pi} \arcsin \mathbb{E}[g_i g_j].$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\mu'} \hat{x}^\top B \hat{x} &= \sum_{i,j} B_{ij} \mathbb{E}[\hat{x}_i \hat{x}_j] \\
&= \sum_{i,j} B_{ij} \frac{2}{\pi} \arcsin[g_i g_j] \\
&= \frac{2}{\pi} \left\langle B, \arcsin\left[\mathbb{E} g g^\top\right] \right\rangle.
\end{aligned}$$

Recall that if $B, C \succeq 0$, then $\langle B, C \rangle \geq 0$. In particular,

$$\left\langle B, \arcsin \left[\mathbb{E} gg^\top \right] - \mathbb{E} gg^\top \right\rangle \geq 0,$$

so

$$\mathbb{E}_{\mu'} \hat{x}^\top B \hat{x} \geq \frac{2}{\pi} \langle B, \mathbb{E} gg^\top \rangle = \frac{2}{\pi} \widetilde{\mathbb{E}}_\mu x^\top B x.$$

The remainder of the proof is identical to that in the previous subsection. ∎

2.3. The most general case

Let us next look at the case where B is any arbitrary matrix. First of all, we may assume that B is symmetric by looking at its symmetrization $(B + B^\top)/2$ instead. We may also assume that all diagonal entries of B are 0, since if we set $B = D + N$ where D is diagonal and N has all diagonal entries zero,

$$\max_{y \in \{-1,1\}^n} y^\top B y = \mathrm{Tr}(B) + \max_{y \in \{-1,1\}^n} y^\top N y.$$

We assume so for the remainder of this subsection.
We shall give an $O(\log n)$-approximation algorithm. First of all, is $\mathrm{opt}(B)$ even non-negative?

Proposition 2.8. Let $y \in [-1, 1]^n$. Then, there exists $\hat{y} \in \{-1, 1\}^n$ such that $\hat{y}^\top B y \geq y^\top B y$.

In particular, setting $y = 0$ implies that the desired value is non-negative.

Proof. Consider the random variable $\hat{y}$ on $\{-1, 1\}^n$ which has $\hat{y}_i = 1$ with probability $(1 + y_i)/2$ and 0 with probability $(1 - y_i)/2$, independently for different coordinates i. Note that $\mathbb{E} \hat{y}_i \hat{y}_j = y_i y_j$ for distinct i, j. Consequently,

$$\mathbb{E}_\mu \hat{y}^\top B \hat{y} = y^\top B y,$$

so the desideratum follows. ∎

The above result is also true in the more general case where $\mathrm{Tr}(B) \geq 0$, but we do not require it.

We can in fact get a stronger lower bound than just the 0 in the above proposition.

Proposition 2.9.

$$\max_{y \in \{-1,1\}^n} y^\top B y \geq \frac{1}{n} \sum_{i,j} |B_{ij}|.$$

Proof. *** INCOMPLETE *** ∎

The sum-of-squares proof we shall give is due to [Meg01, CW04].

Theorem 2.10. For sufficiently large n and $c = O(\log n)$,

$$\frac{\mathrm{opt}(B)}{c} - x^\top B x$$

has a degree 2 sum-of-squares certificate.

While we do not compute the exact constants exactly, the above is true for roughly $n > 60$ and $c = 4 \log n$.

Proof. As in the max-cut and PSD cases, we prove that given any pseudodistribution μ, there exists an (efficiently sampleable) distribution μ' on $\{-1, 1\}^n$ such that $\mathbb{E}_{\mu'}\hat{x}^\top B\hat{x} \geq \frac{1}{O(\log n)}\widetilde{\mathbb{E}}_\mu x^\top Bx$. By Proposition 2.8, it suffices to show this for a distribution on the continuous hypercube $[-1, 1]^n$ instead of $\{-1, 1\}^n$.

As before, choose $g \sim \mathcal{N}(0, \mathbb{E}_\mu xx^\top)$, so

$$\mathbb{E}[g^\top Bg] = \widetilde{\mathbb{E}}_\mu[x^\top Bx].$$

Make the mild assumption that $\mathbb{E}[g^\top Bg] \geq 0$; the analysis of the general case is nearly identical. For a suitable constant C, we have that

$$\Pr[\|g\|_\infty > C\log n] \cdot \mathbb{E}[g^\top Bg \mid \|g\|_\infty > C\log n] \leq \frac{1}{n^2}\mathbb{E}[g^\top Bg], \text{ so}$$

$$\Pr[\|g\|_\infty \leq C\log n] \cdot \mathbb{E}[g^\top Bg \mid \|g\|_\infty \leq C\log n] \geq \left(1 - \frac{1}{n^2}\right)\mathbb{E}[g^\top Bg] \tag{2.2}$$

Our assumption that $\mathbb{E}[g^\top Bg] \geq 0$ also implies that all the quantities above are non-negative. The final random variable $\hat{x}$ on the solid hypercube is defined by

$$\hat{x}_i = \begin{cases} \frac{g_i}{C\sqrt{\log n}}, & |g_i| \leq C\sqrt{\log n}, \\ \frac{g_i}{|g_i|}, & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} \mathbb{E}[\hat{x}^\top Bx] &\geq \Pr[\|g\|_\infty \leq C\sqrt{\log n}] \cdot \mathbb{E}[\hat{x}^\top B\hat{x} \mid \|g\|_\infty \leq C\sqrt{\log n}] \\ &\geq \frac{1}{2C^2\log n}\mathbb{E}[g^\top Bg \mid \|g\|_\infty \leq C\sqrt{\log n}] \\ &\overset{(2.2)}{\geq} \frac{1}{O(\log n)}\mathbb{E}[g^\top Bg] = \frac{1}{O(\log n)}\widetilde{\mathbb{E}}_\mu[x^\top Bx], \end{aligned}$$

completing the proof. ∎

This above rounding is a specific case of the more general RPR^2 scheme of roundings that we mentioned earlier. In this, we "modify" $\widetilde{\mathbb{E}}_\mu xx^\top$ in some way (in the above method of Nesterov, we scale it down), sample a Gaussian with this modified covariance matrix, then do randomized rounding. In the setting of max-cut, we search over all RPR^2 roundings and output whichever returns the maximum cut value.

2.4. The bipartite support case

Alon and Naor [AN04] give a constant factor approximation in the case where the support $\text{supp}(B) \coloneqq \{\{i, j\} : B_{ij} \neq 0\}$ of B defines a bipartite graph (on vertex set $[n]$).

Interestingly, the exact constant of approximation is an open problem, and the interested reader may look up *Grothendieck's inequality* for more details. The approximation constant is known to be somewhere between $\pi/2 \approx 1.57$ and $\pi/(2\ln(1+\sqrt{2})) \approx 1.78$.

References

[AN04] Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, page 72–80, New York, NY, USA, 2004. Association for Computing Machinery.

[CW04] M. Charikar and A. Wirth. Maximizing quadratic programs: extending grothendieck's inequality. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 54–60, 2004.

[DKK+18] Irit Dinur, Subhash Khot, Guy Kindler, Dor Minzer, and Muli Safra. Towards a proof of the 2-to-1 games conjecture? STOC 2018, page 376–389, New York, NY, USA, 2018. Association for Computing Machinery.

[FS02]   Uriel Feige and Gideon Schechtman. On the optimality of the random hyperplane rounding technique for max cut. *Random Structures & Algorithms*, 20, 2002.

[GW00]   Michel Goemans and David Williamson. 0.878 approximation algorithms for max cut and max 2-sat. *Journal of the ACM*, 42, 07 2000.

[Kho02]   Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 767–775, New York, NY, USA, 2002. Association for Computing Machinery.

[KKMO07]   Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? *SIAM Journal on Computing*, 37(1):319–357, 2007.

[Meg01]   Alexandre Megretski. Relaxations of quadratic programs in operator theory and system analysis. In Alexander A. Borichev and Nikolai K. Nikolski, editors, *Systems, Approximation, Singular Integral Operators, and Related Topics*. Birkhäuser Basel, 2001.

[O'D14]   Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.