Penn State University
CSE 564 Complexity of Combinatorial Problems

Instructor: Pei Wu

Module
13

Degree-2 Sum of Squares algorithms

This lecture introduces the *Lasserre / Sum-of-Squares (SoS) hierarchy* as a unifying view of many semidefinite programming (SDP) relaxations for Boolean optimization. The key conceptual shift is to replace an actual distribution on assignments $x \in \{\pm 1\}^n$ by a *degree-d pseudo-distribution*, i.e., a linear functional $\widetilde{\mathbb{E}}[\cdot]$ that behaves like expectation on all polynomials up to degree d and is required to be nonnegative on squares. In degree 2, the only information retained is the second-moment ("moment matrix") $\widetilde{\mathbb{E}}[xx^\top]$, which is exactly the familiar SDP variable.

We then focus on the *power of degree-2 SoS*, which already captures several classic approximation algorithms for quadratic optimization: Goemans–Williamson for Max-Cut, Nesterov's $\frac{2}{\pi}$-approximation for PSD quadratic forms, and Grothendieck/Krivine-type rounding for bilinear forms. The lecture ends with a useful equivalence between the cut norm and the $\ell_\infty \to \ell_1$ operator norm, which is often used to interpret Grothendieck-type bounds.

13.1 Sum-of-squares and pseudo-distributions

A canonical example is Max-Cut. For a graph $G = (V, E)$ with $|V| = n$, encode a cut by $x \in \{\pm 1\}^n$ and write the cut objective as

$$f(x) \;=\; \sum_{(i,j)\in E} \frac{1 - x_i x_j}{2}, \qquad x \in \{\pm 1\}^n. \tag{13.1}$$

The exact optimum is $\mathrm{OPT} = \max_{x\in\{\pm 1\}^n} f(x)$.

The standard SDP relaxation for Max-Cut is

$$\max \sum_{(i,j)\in E} \frac{1 - Y_{ij}}{2} \qquad \text{s.t.} \qquad Y_{ii} = 1 \;\; \forall i, \quad Y \succeq 0, \tag{13.2}$$

where $Y \succeq 0$ means positive semidefinite. If x is an actual cut assignment then $Y = xx^\top$ is feasible and $Y_{ij} = x_i x_j$, so (13.2) is a relaxation.

The SoS viewpoint interprets Y as a *moment matrix* $Y = \widetilde{\mathbb{E}}[xx^\top]$ of some (putative) distribution on $\{\pm 1\}^n$. When we move to pseudo-distributions, we *do not* insist that $\widetilde{\mathbb{E}}$ comes from a real distribution, only that it is consistent with "$\widetilde{\mathbb{E}}[\cdot]$ is nonnegative on squares".

DEFINITION 13.1 (Degree-d pseudo-distribution / pseudo-expectation). A *degree-d pseudo-expectation* on $\{\pm 1\}^n$ is a linear functional $\widetilde{\mathbb{E}}[\cdot]$ defined on all real polynomials p of degree at most d such that:

1. $\widetilde{\mathbb{E}}[1] = 1$;
2. $\widetilde{\mathbb{E}}[p(x)^2] \geq 0$ for every polynomial p with $\deg(p) \leq d/2$.

We sometimes write $\widetilde{\mathbb{E}} = \widetilde{\mathbb{E}}_\mu$ and refer to μ as a *degree-d pseudo-distribution.*

Given an objective polynomial $P(x)$ (e.g., $P = f$ for Max-Cut), the degree-d SoS relaxation is

$$\mathrm{SDP}_d(P) \;=\; \max_{\widetilde{\mathbb{E}} \text{ degree-}d} \; \widetilde{\mathbb{E}}[P(x)].$$

In degree 2, the pseudo-expectation is fully captured by the matrix

$$M \;=\; \widetilde{\mathbb{E}}[xx^\top] \succeq 0, \qquad M_{ii} = \widetilde{\mathbb{E}}[x_i^2] = 1,$$

so SDP_2 is an SDP of the same form as (13.2).

13.2 Pseudo-Cauchy–Schwarz and duality

The "nonnegative on squares" axiom implies many familiar inequalities at low degree.

FACT 13.2 (Pseudo-Cauchy–Schwarz). *Let $\widetilde{\mathbb{E}}$ be a degree-d pseudo-expectation, and let P, Q be polynomials with $\deg(P), \deg(Q) \leq d/2$. Then*

$$\left(\widetilde{\mathbb{E}}[PQ]\right)^2 \;\leq\; \widetilde{\mathbb{E}}[P^2] \cdot \widetilde{\mathbb{E}}[Q^2].$$

Proof. For any real λ we have $\widetilde{\mathbb{E}}[(P - \lambda Q)^2] \geq 0$, hence

$$0 \;\leq\; \widetilde{\mathbb{E}}[(P - \lambda Q)^2] \;=\; \widetilde{\mathbb{E}}[P^2] - 2\lambda\, \widetilde{\mathbb{E}}[PQ] + \lambda^2 \widetilde{\mathbb{E}}[Q^2].$$

If $\widetilde{\mathbb{E}}[Q^2] = 0$ then the above inequality forces $\widetilde{\mathbb{E}}[PQ] = 0$, and the claim is trivial. Otherwise, choose $\lambda = \widetilde{\mathbb{E}}[PQ]/\widetilde{\mathbb{E}}[Q^2]$ and rearrange to get $\widetilde{\mathbb{E}}[PQ]^2 \leq \widetilde{\mathbb{E}}[P^2]\widetilde{\mathbb{E}}[Q^2]$. □

A complementary viewpoint is that "$\widetilde{\mathbb{E}}$ is nonnegative on squares" is the *dual* of being a *sum-of-squares* polynomial. This duality is closely related to the classical theme of representing nonnegative polynomials as sums of squares (Hilbert's 17th problem).

THEOREM 13.3 (Dual view: SoS $\iff$ nonnegativity under all pseudo-distributions). *Let $P : \{\pm 1\}^n \to \mathbb{R}$ be a polynomial (viewed modulo the Boolean identities $x_i^2 = 1$). Then P has a degree-d sum-of-squares certificate of nonnegativity if and only if*

$$\widetilde{\mathbb{E}}[P] \;\geq\; 0 \qquad \textit{for every degree-}d\textit{ pseudo-expectation } \widetilde{\mathbb{E}}.$$

REMARK 13.4. We will not prove this theorem here; it is an SDP strong-duality / separation statement. Intuitively: if P is not in the cone generated by low-degree squares, then a separating hyperplane provides a linear functional $\widetilde{\mathbb{E}}$ that is nonnegative on all low-degree squares but negative on P.

13.3 SoS proofs as algorithms; the Max-Cut corollary

One reason SoS is algorithmically powerful is that searching for a degree-d certificate reduces to an SDP of size roughly $n^{O(d)}$.

FACT 13.5 (Searching for SoS certificates). *If an inequality of the form $P(x) \le v$ admits a degree-d SoS proof, then such a proof can be found in time $\mathrm{poly}(n^d)$ (equivalently, $n^{O(d)}$ time).*

The following corollary explains how classic SDP rounding translates into SoS certificates.

COROLLARY 13.6 (Degree-2 SoS captures Goemans–Williamson for Max-Cut). *Let $f(x) = \sum_{(i,j)\in E} \frac{1-x_i x_j}{2}$ and let $\mathrm{OPT} = \max_{x\in\{\pm1\}^n} f(x)$. Then*

$$\mathrm{OPT} - \alpha_{\mathrm{GW}} \cdot f(x) \;\ge\; 0$$

admits a degree-2 sum-of-squares proof, where $\alpha_{\mathrm{GW}} \approx 0.878$ is the Goemans–Williamson constant.

Proof sketch (as in the notes). Fix an arbitrary degree-2 pseudo-expectation $\widetilde{\mathbb{E}}$. Let $M = \widetilde{\mathbb{E}}[xx^\top] \succeq 0$ be its moment matrix. Goemans–Williamson Gaussian rounding applied to M produces a random cut $X \in \{\pm1\}^n$ such that

$$\mathbb{E}[f(X)] \;\ge\; \alpha_{\mathrm{GW}} \cdot \widetilde{\mathbb{E}}[f(x)].$$

Since OPT is the maximum cut value, $\mathrm{OPT} \ge \mathbb{E}[f(X)]$, and hence $\mathrm{OPT} - \alpha_{\mathrm{GW}} \cdot \widetilde{\mathbb{E}}[f(x)] \ge 0$, i.e. $\widetilde{\mathbb{E}}[\mathrm{OPT} - \alpha_{\mathrm{GW}} f] \ge 0$. By Theorem 13.3, this implies that $\mathrm{OPT} - \alpha_{\mathrm{GW}} f$ has a degree-2 SoS certificate. ∎

At the opposite extreme, the hierarchy becomes exact once the degree reaches $2n$.

FACT 13.7 (Trivial degree-$2n$ SoS on the Boolean cube). *If $f : \{\pm1\}^n \to \mathbb{R}_{\ge 0}$ is any nonnegative function on the Boolean cube, then f admits a degree-$2n$ sum-of-squares proof.*

One convenient representation. For each $a \in \{\pm1\}^n$, define the multilinear “indicator” polynomial

$$\delta_a(x) \;=\; \frac{1}{2^n} \prod_{i=1}^{n} (1 + a_i x_i).$$

Over $\{\pm1\}^n$, $\delta_a(x) \in \{0,1\}$ and $\delta_a(x) = 1$ iff $x = a$. Moreover $\deg(\delta_a) = n$ and (modulo $x_i^2 = 1$) we have $\delta_a(x)^2 = \delta_a(x)$. Therefore, for $f \ge 0$,

$$f(x) \;=\; \sum_{a\in\{\pm1\}^n} f(a)\,\delta_a(x) \;\equiv\; \sum_{a\in\{\pm1\}^n} \Big(\sqrt{f(a)}\,\delta_a(x)\Big)^2,$$

which is a sum of squares of degree at most $2n$. ∎

13.4 Quadratic optimization and Nesterov's $\frac{2}{\pi}$ theorem

We now focus on homogeneous quadratic optimization problems of the form

$$\max_{x\in\{\pm 1\}^n} x^\top P x, \tag{13.3}$$

where $P \in \mathbb{R}^{n\times n}$ is a given matrix (assume w.l.o.g. symmetric). Several important approximation results fit this template:

- If P is a graph Laplacian, (13.3) captures Max-Cut and admits a 0.878-approximation (Goemans–Williamson), which is optimal under the Unique Games Conjecture.
- If $P \succeq 0$ is PSD, there is a $\frac{2}{\pi}$-approximation (Nesterov), and this factor is optimal in general.
- For general P, one can obtain a $\frac{1}{O(\log n)}$ approximation (Charikar–Wirth), which is essentially best possible.
- For block off-diagonal matrices $P = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix}$ one encounters Grothendieck-type inequalities and rounding (Alon–Naor).

Theorem 13.8 (Nesterov's $\pi/2$ theorem (PSD case)). *If $P \succeq 0$, then there is a degree-2 SoS (SDP) rounding algorithm that achieves a $\frac{2}{\pi}$-approximation for* (13.3).

Proof sketch. Solve the degree-2 SoS relaxation and obtain the moment matrix $M = \widetilde{\mathbb{E}}[xx^\top] \succeq 0$ with $M_{ii} = 1$. Gaussian rounding samples

$$g \sim \mathcal{N}(0, M) \qquad \text{and outputs} \qquad x_i = \operatorname{sign}(g_i) \in \{\pm 1\}.$$

By Sheppard's lemma (the "arcsine law"),

$$\mathbb{E}[x_i x_j] \;=\; \mathbb{E}[\operatorname{sign}(g_i)\operatorname{sign}(g_j)] \;=\; \frac{2}{\pi}\arcsin(M_{ij}). \tag{13.4}$$

Therefore the expected objective value is

$$\mathbb{E}[x^\top P x] = \left\langle \mathbb{E}[xx^\top], P \right\rangle = \frac{2}{\pi}\left\langle \arcsin(M), P \right\rangle. \tag{13.5}$$

To compare $\langle \arcsin(M), P\rangle$ to $\langle M, P\rangle$, use the Taylor series

$$\arcsin(t) = t + \frac{1}{2}\cdot\frac{t^3}{3} + \frac{1\cdot 3}{2\cdot 4}\cdot\frac{t^5}{5} + \cdots, \qquad t \in [-1, 1],$$

whose coefficients are all nonnegative. Interpreting this *entrywise* on matrices gives

$$\arcsin(M) = M + \sum_{k\geq 1} a_k\, M^{\circ(2k+1)}, \qquad a_k > 0, \tag{13.6}$$

where $M^{\circ r}$ denotes the r-fold Hadamard (entrywise) power. Since $M \succeq 0$, every Hadamard power $M^{\circ r}$ is PSD as well (by repeated use of the Schur product theorem: $A, B \succeq 0 \Rightarrow A \circ B \succeq 0$). Thus (13.6) implies $\arcsin(M) \succeq M$.

Because $P \succeq 0$, we get $\langle \arcsin(M), P\rangle \ge \langle M, P\rangle$. Finally,

$$\langle M, P\rangle = \widetilde{\mathbb{E}}[x^\top P x] \;\ge\; \max_{x\in\{\pm1\}^n} x^\top P x,$$

since the SDP is a relaxation. Putting everything together,

$$\mathbb{E}[x^\top P x] \;\ge\; \frac{2}{\pi}\langle M, P\rangle \;\ge\; \frac{2}{\pi}\mathrm{OPT},$$

which is the desired approximation ratio. ∎

Remark 13.9 (Useful Taylor series). For reference (as in the notes),

$$\sin(t) = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots, \qquad \sinh(t) = t + \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots.$$

These series (together with Schur products) reappear in Krivine rounding below.

13.5 Grothendieck inequality via degree-2 SoS and Krivine rounding

Consider the bilinear optimization problem

$$\max_{x\in\{\pm1\}^m,\; y\in\{\pm1\}^n} x^\top B y \;=\; \max_{x,y} \sum_{i,j} B_{ij} x_i y_j. \tag{13.7}$$

This can be written as a quadratic form by introducing the block matrix $P = \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix}$ and the concatenated vector (x, y).

The degree-2 SDP relaxation corresponds to replacing the signs x_i, y_j by unit vectors u_i, v_j and maximizing $\sum_{i,j} B_{ij}\langle u_i, v_j\rangle$. Grothendieck's inequality states that this relaxation is tight up to a universal constant.

Theorem 13.10 (Grothendieck/Krivine (algorithmic form)). *There exists a universal constant K_{G} such that for every matrix B,*

$$\underbrace{\max_{\substack{\|u_i\|\le 1\\ \|v_j\|\le 1}} \sum_{i,j} B_{ij}\langle u_i, v_j\rangle}_{\|B\|_{\gamma*}} \;\le\; K_{\mathrm{G}} \cdot \underbrace{\max_{x\in\{\pm1\}^m, y\in\{\pm1\}^n} \sum_{i,j} B_{ij} x_i y_j}_{\|B\|_{\infty\to1}}.$$

Equivalently, given vectors $(u_i), (v_j)$ one can efficiently round to signs (x, y) achieving at least a $\frac{1}{K_{\mathrm{G}}}$ fraction of the vector value. With Krivine's bound, one may take

$$K_{\mathrm{G}} \;=\; \frac{\pi}{2\ln(1+\sqrt{2})}.$$

In the SoS language, this implies that if $\mathrm{OPT} = \|B\|_{\infty\to1}$, then the polynomial

$$\mathrm{OPT} - \frac{1}{K_{\mathrm{G}}}\sum_{i,j} B_{ij} x_i y_j$$

has a degree-2 SoS certificate (by the same "rounding ⇒ dual certificate" logic as in Corollary 13.6).

Proof sketch (Krivine rounding as in the notes). Let $\widetilde{\mathbb{E}}$ be an optimal degree-2 pseudo-expectation for (13.7). Define the block moment matrix

$$M = \widetilde{\mathbb{E}}\left[\begin{pmatrix} x \\ y \end{pmatrix} \begin{pmatrix} x^\top & y^\top \end{pmatrix}\right] = \begin{pmatrix} \widetilde{\mathbb{E}}[xx^\top] & \widetilde{\mathbb{E}}[xy^\top] \\ \widetilde{\mathbb{E}}[yx^\top] & \widetilde{\mathbb{E}}[yy^\top] \end{pmatrix} \succeq 0, \qquad M_{kk} = 1.$$

(The diagonal constraint follows from $x_i^2 = y_j^2 = 1$.)

Krivine's idea is to build a new covariance matrix Σ by applying sinh to the diagonal blocks and sin to the off-diagonal blocks *entrywise.* Fix a parameter $c > 0$ and define

$$\Sigma \;=\; \begin{pmatrix} \sinh\big(c\,\widetilde{\mathbb{E}}[xx^\top]\big) & \sin\big(c\,\widetilde{\mathbb{E}}[xy^\top]\big) \\ \sin\big(c\,\widetilde{\mathbb{E}}[yx^\top]\big) & \sinh\big(c\,\widetilde{\mathbb{E}}[yy^\top]\big) \end{pmatrix}, \tag{13.8}$$

where $\sin(\cdot)$ and $\sinh(\cdot)$ are applied entrywise to matrices.

(PSD requirement.) Using the Taylor series for sin and sinh, Σ can be expanded as a nonnegative linear combination of Hadamard powers of M:

$$\Sigma = \sum_{k \geq 0} \frac{c^{2k+1}}{(2k+1)!} \left(M^{\circ(2k+1)} \;\circ\; S_k\right),$$

where S_k is the 2-block sign matrix that equals $+1$ on the (x,x) and (y,y) blocks and equals $(-1)^k$ on the cross blocks. Each S_k is PSD (it is rank-1, coming from the vector $(\mathbf{1},\,(-1)^k\mathbf{1})$), each Hadamard power $M^{\circ(2k+1)}$ is PSD by the Schur product theorem, and the Hadamard product of PSD matrices is PSD. Hence every summand is PSD and therefore $\Sigma \succeq 0$.

(Normalization.) Since $\widetilde{\mathbb{E}}[x_i^2] = \widetilde{\mathbb{E}}[y_j^2] = 1$, the diagonal entries of Σ are all $\sinh(c)$. Choose c so that $\sinh(c) = 1$, i.e.

$$c = \operatorname{arcsinh}(1) = \ln(1+\sqrt{2}),$$

which ensures $\Sigma_{kk} = 1$ for all k.

(Gaussian rounding.) Now sample a Gaussian pair $(g,h) \sim \mathcal{N}(0,\Sigma)$ and output

$$x_i = \operatorname{sign}(g_i), \qquad y_j = \operatorname{sign}(h_j).$$

By Sheppard's lemma,

$$\mathbb{E}[x_i y_j] = \frac{2}{\pi}\arcsin(\mathbb{E}[g_i h_j]) = \frac{2}{\pi}\arcsin\Big(\sin(c\cdot\widetilde{\mathbb{E}}[x_i y_j])\Big).$$

Because $c \leq \pi/2$ and $\widetilde{\mathbb{E}}[x_i y_j] \in [-1,1]$, we have $\arcsin(\sin(ct)) = ct$ for $t \in [-1,1]$, hence

$$\mathbb{E}[x_i y_j] = \frac{2c}{\pi}\,\widetilde{\mathbb{E}}[x_i y_j].$$

Therefore,

$$\mathbb{E}[x^\top By] = \sum_{i,j} B_{ij}\mathbb{E}[x_i y_j] = \frac{2c}{\pi}\sum_{i,j} B_{ij}\widetilde{\mathbb{E}}[x_i y_j] = \frac{2c}{\pi}\,\widetilde{\mathbb{E}}[x^\top By].$$

Since $\widetilde{\mathbb{E}}$ was optimal for the SDP, $\widetilde{\mathbb{E}}[x^\top By] = \mathrm{SDP}_2(B)$, and with $c = \ln(1+\sqrt{2})$ we obtain the factor

$$\frac{2c}{\pi} = \frac{1}{K_{\mathrm{G}}} \qquad \text{where} \qquad K_{\mathrm{G}} = \frac{\pi}{2\ln(1+\sqrt{2})}.$$

This gives the claimed rounding guarantee. ∎

13.6 Cut norm vs. $\ell_\infty \to \ell_1$ norm

Alon-Naor was studying algorithmic weak regularity. They were interested in approximate the so called cut norm algorithmically, which is constantly related to $\ell_\infty \to \ell_1$ norm. Therefore can be approximated by SDP up to constant approximation ratio.

Definition 13.11 (Cut norm). For a matrix $M \in \mathbb{R}^{m\times n}$, define the *cut norm*

$$\|M\|_\square \;=\; \max_{S\subseteq[m],\,T\subseteq[n]} \left|\langle \mathbf{1}_S, M\mathbf{1}_T\rangle\right|,$$

and the operator norm

$$\|M\|_{\infty\to 1} \;=\; \max_{u\in\{\pm1\}^m,\,v\in\{\pm1\}^n} \left|u^\top M v\right|.$$

Fact 13.12. *For every matrix M,*

$$\|M\|_\square \;\le\; \|M\|_{\infty\to1} \;\le\; 4\,\|M\|_\square\,.$$

Proof. For the first inequality, let (S,T) achieve $\|M\|_\square = |\,\langle \mathbf{1}_S, M\mathbf{1}_T\rangle\,|$. Define random sign vectors $u \in \{\pm1\}^m$ and $v \in \{\pm1\}^n$ by

$$u_i = \begin{cases} 1, & i\in S,\\ \text{uniform in } \{\pm1\}, & i\notin S,\end{cases} \qquad v_j = \begin{cases} 1, & j\in T,\\ \text{uniform in } \{\pm1\}, & j\notin T.\end{cases}$$

Then $\mathbb{E}[u] = \mathbf{1}_S$ and $\mathbb{E}[v] = \mathbf{1}_T$, so by linearity,

$$\mathbb{E}[u^\top Mv] = \langle \mathbb{E}[u], M\mathbb{E}[v]\rangle = \langle \mathbf{1}_S, M\mathbf{1}_T\rangle\,.$$

Hence there exists a *deterministic* choice of signs with $|u^\top Mv| \ge |\,\langle \mathbf{1}_S, M\mathbf{1}_T\rangle\,| = \|M\|_\square$, proving $\|M\|_{\infty\to1} \ge \|M\|_\square$.

For the second inequality, fix any $u \in \{\pm1\}^m$ and $v\in\{\pm1\}^n$. Let $S^+ = \{i : u_i = 1\}$, $S^- = \{i : u_i = -1\}$ and similarly T^+, T^- for v. Then

$$u^\top Mv = \langle \mathbf{1}_{S^+} - \mathbf{1}_{S^-}, M(\mathbf{1}_{T^+} - \mathbf{1}_{T^-})\rangle$$

expands into a sum of four cut terms:

$$u^\top Mv = \langle \mathbf{1}_{S^+}, M\mathbf{1}_{T^+}\rangle - \langle \mathbf{1}_{S^+}, M\mathbf{1}_{T^-}\rangle - \langle \mathbf{1}_{S^-}, M\mathbf{1}_{T^+}\rangle + \langle \mathbf{1}_{S^-}, M\mathbf{1}_{T^-}\rangle\,.$$

Taking absolute values and using the definition of $\|M\|_\square$ gives $|u^\top Mv| \le 4\,\|M\|_\square$. Maximizing over u, v completes the proof. ∎