CS369H: Hierarchies of Integer Programming Relaxations Spring 2016-2017

Rounding Sum-of-Squares Relaxations

Professor Moses Charikar *Presenter: Andy Tsao*

Overview: these lecture notes will be following the paper [Bar]

1 Introduction

We have spent a majority of the quarter talking about the sum-of-squares (SOS) hierarchy. Given an optimization problem in variables $x_1, \dots, x_n$, with constraints $A = (f_1 \geq 0, \dots, f_m \geq 0)$, and an objective function $f(x)$, the SOS hierarchy tests whether the set of equations $(f_1 \geq 0, \dots, f_m \geq 0, f \leq c)$ is satisfiable over the solution space (typically nonconvex). It does this by outputting objects called *degree-d pseudodistributions* with associated *pseudoexpectations*, which are maps $\tilde{\mathbb{E}}$ that satisfy

- **Normalization**: $\tilde{\mathbb{E}}(1) = 1$
- **Linearity**: $\tilde{\mathbb{E}}(P + Q) = \tilde{\mathbb{E}}P + \tilde{\mathbb{E}}Q$ for every P, Q of degree $\leq d$.
- **Positivity**: $\tilde{\mathbb{E}}(P^2) \geq 0$ for all P of degree $\leq d/2$.

A pseudodistribution over the solution space that satisfies the above constraints is considered a proxy for solving the optimization problem.

Traditionally, the SOS framework was mostly used to produce *negative* results. In other words, it was used to show that $O(1)$ levels of the SOS hierarchy can't really improve on the known polynomial-time approximation guarantees for many NP-hard problems such as SAT, Independent set, and Max-cut. Unfortunately, there have been fewer *positive* results, and several of them can only show that the SOS hierarchy can match the performance of previously-known (and oftentimes more efficient) algorithms. For example, Karlin, Mathieu, and Nguyen showed that l levels of SOS can approximate the Knapsack problem up to a factor of $1 + 1/l$, approaching the performance of the standard dynamic program.

One reason for the lack of positive results is the inability to round solutions to convex programs.

2 General Technique

Traditionally, approximation algorithms work as follows: given some optimization problem O, one first decides what the relaxation is (LP, SDP, etc.). Then, after finding a solution $\tilde{x}$ in the relaxed space, one designs a *rounding algorithm* that maps $\tilde{x}$ into a solution $\hat{x}$ of the original problem with approximately the same value.

The technique that the authors use is different — they design and analyze the rounding algorithm first, and come up with the relaxation afterwards.

2.1 Rounding vs. combining algorithms

Definition 1.1. *A* rounding algorithm *is a map from the space of relaxed solutions to a solution to the original program. A* combining algorithm *takes as input a distribution $\mathcal{X}$ over solutions in S and maps it into a single element $C(\mathcal{X})$, such that the objective value of $C(\mathcal{X})$ is close to the expected objective value of a random element in $\mathcal{X}$.*

Observe that every rounding algorithm R yields a combining algorithm C. Indeed, if some embedding map f maps S into some convex domain T, then for every distribution $\mathcal{X}$ over S, we can define $y_{\mathcal{X}}$ to be $\mathbb{E}_{x \in \mathcal{X}} f(X)$. By convexity, $y_{\mathcal{X}} \in T$ and its objective value will be at most the average objective value of an element in $\mathcal{X}$. So defining $C(\mathcal{X}) = R(y_{\mathcal{X}})$ yields a combining algorithm that is at least as good as R. Conversely, because the set of distributions over S is convex and can be optimized over by a linear program, every combining algorithm can be viewed as a rounding algorithm for this program. However, $|S|$ is typically exponential in size, so the linear program will not be very useful. However, in some cases, nontrivial combining algorithms (for instance, one that does more than just sample x from $\mathcal{X}$ at random) can be turned into a rounding algorithm using an efficient convex program. In our case, a nontrivial combining algorithm C has the form $C(\mathcal{X}) = C'(M(\mathcal{X}))$, where C' is an efficient algorithm and $M(\mathcal{X})$ is the vector of moments of $\mathcal{X}$ of degree at most d. Typically, this framework also extends to pseudodistributions, since many of the tools we will use in the analysis (such as Cauchy-Schwarz and Holder) are robust to this difference.

3 Examples

3.1 Nonnegative tensor maximization

Definition 1.2. *We define the* spectral norm *of a degree-$2t$ homogeneous polynomial M in $x = (x_1, \ldots, x_n)$, denoted by $\|M\|_{spectral}$, to be the minimum of the spectral norm of Q taken over all quadratic forms Q over $(\mathbb{R}^n)^{\otimes t}$ such that $Q(x^{\otimes t}) = M(x)$ for every x.*

Definition 1.3. *The* Hellinger distance *between two distributions p and q is defined by*

$$d_H(p, q) = \sqrt{1 - \sum_i \sqrt{p_i q_i}}.$$

In particular, $d_H(p, q)$ is $1/\sqrt{2}$ times the Euclidean distance of the unit vectors $\sqrt{p}$ and $\sqrt{q}$.

The main theorem for nonnegative tensor maximization is as follows:

Theorem 1.1. *Let M be a degree-$2t$ homogeneous polynomial in $x = (x_1, \ldots, x_n)$ with nonnegative coefficients. Then, there is an algorithm, based on $O(t^2 \log n/\epsilon^2)$ rounds of SOS, that finds a unit vector $x^* \in \mathbb{R}^n$ such that*

$$M(x^*) \geq \max_{x \in \mathbb{R}^n, \|x\|=1} M(x) - \epsilon \|M\|_{spectral}.$$

Our proof strategy will be as follows: we first come up with a combining algorithm, which takes as input a distribution over unit vectors $x \in \mathbb{R}^n$ such that $M(x) \geq v$ and outputs a unit vector x^* such that $M(x) \geq$

$v - \epsilon$. We then show that our algorithm still works even if given a level $O(t \log n/\epsilon^2)$ pseudodistribution.
Combining algorithm:
Given a distribution $\mathcal{X}$ over $x \in \mathbb{R}^n$ with $M(x) = v$, do the following for $t^2 \log n/\epsilon^2$ steps:

1. For $i \in [n]$, let $x_i^* = \sqrt{\mathbb{E}_{x \sim \mathcal{X}} x_i^2}$. If $M(x^*) \geq v - 4\epsilon$, then output x^*.
2. Otherwise, set $\mathcal{X}$ to $\mathcal{X}_{i_1,\ldots,i_{t-1}}$, where $\mathcal{X}_{i_1,\ldots,i_{t-1}}$ satisfies

   $$\Psi(\mathcal{X}_{i_1,\ldots,i_{t-1}}) \leq \Psi(\mathcal{X}) - \epsilon^2/t^2,$$

   and

   - $\mathcal{X}_{i_1,\ldots,i_{t-1}}$ is defined by setting $P(X_{i_1,\ldots,i_{t-1}})$ to be proportional to $P(\mathcal{X} = x) \cdot \prod_{j=1}^{t-1} x_{i_j}^2$ for every $x \in \mathbb{R}^n$.
   - $\Psi(\mathcal{X}) = H(A(\mathcal{X}))$, where $H(\cdot)$ is the Shannon entropy function and $A(\mathcal{X})$ is the distribution over $[n]$ obtained by letting $P(A(\mathcal{X}) = i) = \mathbb{E}_{x \sim \mathcal{X}} x_i^2$ for every $i \in [n]$.

The idea behind the algorithm is as follows: either we can construct a suitable x^*, or we can decrease a function of the entropy by a fixed amount. Since $\Psi(\mathcal{X}) \in [0, \log n]$ by Jensen's inequality, after a finite number of iterations we can be guaranteed to find a suitable x^*.
The proof of Theorem 1.1 goes roughly as follows:

Proof sketch. Given $\mathcal{X}$, we define the following random variables $A_1, \ldots, A_t$ over $[n]$ to have the following property:

$$P((A_1, \ldots, A_t) = (i_1, \ldots, i_t)) = \mathbb{E}_{x \sim \mathcal{X}} x_{i_1}^2 \ldots x_{i_t}^2.$$

Showing that the above algorithm succeeds requires two lemmas. The first establishes a sufficient condition on the random variables $\{A_i\}$ (heuristically, they must be almost independent) for finding x^* such that $M(x^*) \geq v - 4\epsilon \|M\|_{\text{spectral}}$.

Lemma 1.1. *If $d_H(\{A_1 \ldots A_t\}, \{A_1\} \ldots \{A_t\}) \leq \epsilon$, then the unit vector x^* with $x_i^* = \sqrt{\mathbb{E}_{x \sim \mathcal{X}} x_i^2}$ satisfies $M(x^*) \geq v - 4\epsilon \|M\|_{spectral}$. Moreover, this holds even if $\mathcal{X}$ is a level $l \geq 2t$ pseudodistribution.*

The second lemma states that we can decrease the entropy if the condition in Lemma 1.1 is not satisfied.

Lemma 1.2. *If $d_H(\{A_1 \ldots A_t\}, \{A_1\} \ldots \{A_t\}) \geq \epsilon$, then $H(A_t \mid A_1 \ldots A_{t-1}) \leq H(A) - 2\epsilon^2/t^2$.*

It is then easy to check that the conditional distribution $(A_t \mid A_1 = i_1, \ldots, A_{t-1} = i_{t-1})$ has the desired properties of $\mathcal{X}_{i_1,\ldots,i_{t-1}}$. □

3.2 Finding a sparse vector in a subspace

Finding a sparse nonzero vector inside a d dimensional linear subspace V of $\mathbb{R}^n$ arises in many applications in machine learning and optimization. Formally, we consider the following problem $P(\mu, d, |\mathcal{U}|, \epsilon)$:
Input: An arbitrary basis for a linear subspace $V = \text{span}(V' \cup \{f_0\})$, where $V' \subseteq \mathbb{R}^{\mathcal{U}}$ is a random d-dimensional subspace, chosen as the span of d vectors drawn independently frmo the standard Gaussian

distribution on $\mathbb{R}^{\mathcal{U}}$, and f_0 is an arbitrary μ-sparse vector, i.e., $S = \text{supp}(f_0)$ has $|S| \le \mu|\mathcal{U}|$.
Goal: Find a vector $f \in V$ with $\langle f, f_0 \rangle^2 \ge (1-\epsilon)\|f\|_2\|f_0\|_2$.
The main result for this problem is as follows:

Theorem 1.2. *For some absolute constant $K > 0$, there is an algorithm that solves $P(\mu, d, |\mathcal{U}|, \epsilon)$ with high probability in time poly$(|\mathcal{U}|, \log(1/\epsilon))$ for any $\mu < K\mu_0(d)$, where*

$$\mu_0(d) = \begin{cases} 1 & \textit{if } d \le \sqrt{|\mathcal{U}|} \\ n/d^2 & \textit{otherwise} \end{cases}.$$

The algorithm will work as follows: it will first solve a constant-degree SOS relaxation to find a noisy approximate solution. Then, it will solve an auxiliary linear program that converts any sufficiently good approximate solution into an exact one. The following theorem makes more precise the notion of an "approximate" solution.

Theorem 1.3. *Let $V = span(V' \cup \{f_0\})$, where $f_0 \in \mathbb{R}^{\mathcal{U}}$ is a vector with $\|f_0\|_4/\|f_0\|_2 \ge C$, and $V' \subseteq \mathbb{R}^{\mathcal{U}}$ is a linear subspace with*

$$\max_{0 \ne f \in V'} \frac{\|f\|_4}{\|f\|_2} \le c. \tag{1}$$

Furthermore, assume that (1) has a degree-4 sum of squares proof, i.e., that

$$\|\Pi_{V'} f\|_4^4 = c^4 \|\Pi_{V'} f\|_2^4 - S,$$

where $\Pi_{V'}$ is the orthogonal projection onto V', and S is a degree-4 sum of squares.
Then, there is a polynomial-time algorithm based on a constant-degree SOS relaxation that returns a vector $f \in V$ with $\langle f, f_0 \rangle \ge (1 - (c/C)^{\Omega(1)})\|f_0\|_2\|f\|_2$.

Proof. The key step is the following lemma about (pseudo)distributions supported on L_4/L_2-sparse functions in V'.

Lemma 1.3. *Let $V' \subseteq \mathbb{R}^{\mathcal{U}}$ be a linear subspace such that*

$$\max_{0 \ne f \in V'} \frac{\|f\|_4}{\|f\|_2} \le c.$$

Let f_0 be a unit function in $V'^{\perp}$ with $\|f_0\|_4 = C > 100c$, and let $\mathcal{X}$ be a distribution over $\mathbb{R}^{\mathcal{U}}$ over unit functions $f \in span(V \cup \{f_0\})$ satisfying $\|f\|_4 \ge C$. Then

$$\mathbb{E}\langle x, f_0 \rangle^2 \ge 1 - O(c/C).$$

Moreover, this holds even if $\mathcal{X}$ is a pseudodistribution of level $l \ge 8$, as long as (1) has a degree-4 SOS proof.

We first argue that the lemma implies the theorem. Solving a degree-8 SOS program that maximizes $\|f\|_4^4$ over $f \in V$ with the constraint $\|f\|_2^2 = 1$ gives us a pseudodistribution $\mathcal{X}$ that meets the requirements of the lemma. Next, we sample a random Gaussian with the same first two moments as $\mathcal{X}$. This gives us a vector g such that $\mathbb{E}\|g\|_2^2 = 1$ and $\mathbb{E}\langle g, f_0 \rangle = (1 - o(1))\|f_0\|$. This concludes the proof of the theorem. □

Proof of lemma 1.3. To prove this lemma, we first write every vector f in the support of $\mathcal{X}$ in the form $f = \alpha f_0 + f'$, where $f' \in V'$ and $\alpha = \langle f, f_0 \rangle$. Then, by the definition of f, we have that

$$C = \|f\|_4 \leq \alpha \|f_0\|_4 + \|f'\|_4 \leq \alpha C + c\|f'\|_2 \leq \alpha C + c \tag{2}$$

which means that $\alpha \geq 1 - c/C$. This concludes the proof for actual distributions $\mathcal{X}$. To extend this result to pseudodistributions, we write

$$\begin{aligned} C^4 &= \|f\|_4^4 = \tilde{\mathbb{E}}_f \mathbb{E}_\omega (\alpha f_0(\omega) + f'(\omega))^4 \\ &= \tilde{\mathbb{E}}_f \alpha^4 \|f_0\|_4^4 + 4\tilde{\mathbb{E}}_f \alpha^3 \langle f_0^3, f' \rangle + 6\tilde{\mathbb{E}}_f \alpha^2 \langle f_0^2, f'^2 \rangle + 4\tilde{\mathbb{E}}_f \alpha \langle f_0, f'^3 \rangle + \tilde{E}_f \|f'\|_4^4. \end{aligned}$$

If V' is a random subspace of dimension d, [BBH] showed that (1) has a degree-4 sum-of-squares proof with high probability for $c = O(1)$ when $d \leq \sqrt{|\mathcal{U}|}$, and for $c = O(d^{1/2}/|\mathcal{U}|^{1/4})$ otherwise.For the purposes of this theorem, we merge these two cases together and say that $c = O(\mu_0(d)^{-1/4})$.
The existence of a degree-4 SOS proof of (2) means that $\mathcal{X}$ must be consistent with the constraint $\|f\|_4^4 \leq c^4$. An application of Cauchy-Schwarz and Holder's inequality gives

$$C^4 \leq \tilde{\mathbb{E}}_f \alpha^4 C^4 + 15|\alpha|^3 C^3 c.$$

Since the pseudoexpectation is consistent with the constraint $|\alpha| \leq 1$, this simplifies to $\tilde{\mathbb{E}}\alpha^4 \geq 1 - 15c/C$, after which we can apply Cauchy-Schwarz again:

$$\tilde{\mathbb{E}}\alpha^4 \leq \sqrt{\tilde{\mathbb{E}}\alpha^2}\sqrt{\tilde{\mathbb{E}}\alpha^6} \leq \sqrt{\tilde{\mathbb{E}}\alpha^2}.$$

This yields $\tilde{\mathbb{E}}\alpha^2 \geq 1 - 30c/C$ and concludes the proof. □

For the second stage, we consider the following linear program:

$$\begin{array}{rl} \text{minimize} & \|y\|_1 \\ \text{such that} & \langle y, f \rangle = 1 \end{array}$$

One can think of this as searching for a sparse vector in V with large inner product with f. The next theorem provides conditions under which the linear program will exactly recover f_0 from any f that is reasonably correlated to it. We state it without proof, as the proof does not utilize any sum-of-squares machinery.

Theorem 1.4. *Let $V = span(V' \cup \{f_0\})$, and suppose that the following conditions hold:*

- *f_0 is a μ-sparse vector. That is, $supp(f_0) = S$, $|S| = \mu n$*
- *V' doesn't contain any $1/\alpha^2$-L_2/L_1-sparse vectors. This means that $\|V'\|_{2:1} \leq \alpha$, where $\|V'\|_{2:1} = \max \|f'\|_2/\|f'\|_1$ for all $0 \neq f' \in V'$.*
- *f is sufficiently correlated with f_0. That is, $\langle f_0, f \rangle \geq (1-\epsilon)\|f_0\|_2\|f\|_2$.*
- *f is not very correlated with anything in V'. That is, $\langle f', f \rangle \leq \eta \|f'\|_2 \|f\|_2$ for all $f' \in V'$.*

Then, if

$$\frac{\eta}{1-\epsilon} < \frac{1}{\alpha\sqrt{\mu}} - 2,$$

then $f_0/\langle f_0, f \rangle$ is the unique optimal solution to the above linear program.

The following theorem deals with the case where there are no constraints on d.

Theorem 1.5. *There is a constant $\epsilon > 0$ and a polynomial-time algorithm A, based on $O(1)$ levels of the SOS hierarchy, that on input a projector operator Π such that there exists a μ-sparse Boolean function f satisfying $\|\Pi f\|_2^2 \geq (1-\epsilon)\|f\|_2^2$, outputs a function $g \in Image(\Pi)$ such that*

$$\|g\|_4^4 \geq \Omega\left(\frac{\|g\|_2^4}{\mu(rank(\Pi))^{1/3}}\right) \tag{3}$$

As with nonnegative tensor maximization, we first construct the combining algorithm. Given a distribution $\mathcal{D}$ over Boolean functions $f \in L_2(\mathcal{U})$ that satisfy

- $\mu(f) = P(f(\omega) = 1) = 1/\lambda$
- $\|\Pi f\|_2^2 \geq (1-\epsilon)\|f\|_2^2$

we do the following:

(a) Let δ_ω be the coordinate vectors. That is, $\langle f, \delta_\omega \rangle = f(\omega)$ for all $f \in L_2(\mathcal{U})$. Go over all vectors of the form $g_\omega = \Pi\delta_\omega$, and if there is one that satisfies (3) then output it.

(b) Choose a random vector t, and if Πt satisfies (3) then output it.

(c) For each 4-tuple $\omega_1, \ldots, \omega_4$, define the distribution $\mathcal{D}_{\omega_1,\ldots,\omega_4}$ to be such that

$$P_{\mathcal{D}_{\omega_1,\ldots,\omega_4}}(f) \propto P_{\mathcal{D}}(f) \prod_{j=1}^{4} f(\omega_j)^2$$

for every f.

(d) For each $\mathcal{D}_{\omega_1,\ldots,\omega_4}$, let t be a random Gaussian that matches the first two moments. Output $g = \Pi t$ if it satisfies (3).

The analysis of these four steps is quite technical and not particularly enlightening. The key takeaway is that if all four steps fail, we will be able to find a function g that satisfies (3). Furthermore, all of the techniques used in the analysis extend quite naturally from distributions to pseudodistributions.

4 Application to small set expansion

We conclude by discussing a few applications of the results to the Small Set Expansion problem. This is the problem of deciding, given an input graph G and parameters δ, ϵ, whether there is a measure-δ subset S of G's vertices where almost all of S's edges connect to vertices in $G \backslash S$. In this section, we consider G to be a Cayley graph on $\mathbb{F}_2^l$ with $n = 2^l$ vertices. Let $V_{\geq\lambda}$ be the linear subspace spanned by the eigenfunctions of G with eigenvalue at least λ. Let P_λ be the degree-4 polynomial $P_\lambda(f) = \|\Pi_{\geq\lambda} f\|_4^4$, where $\Pi_{\geq\lambda}$ is the projector into $V_{\geq\lambda}$. Let $K_\lambda(G) = \|P_\lambda\|_{\text{spectral}}$.
The following theorem shows that low-degree SOS relaxations can detect L_4/L_2 sparse functions in the subspaces $V_{\geq\lambda}$. This result follows from the previous section on nonnegative tensor maximization.

Theorem 1.6. *SoS relaxations of degree $\epsilon^{-O(1)}K_\lambda(G)^{O(1)}\log n$ provide an additive ϵ-approximation to the maximum of $\|f\|_4/\|f\|_2$ over all non-zero functions $f \in V_{\geq\lambda}$.*

Proof. In order to apply our nonnegative tensor maximization result, we must verify that P_λ has nonnegative coefficients in an appropriate basis. Consider the basis of characters $\{\xi_\alpha\}_{\alpha\in\mathbb{F}_2^l}$. Then,

$$\begin{aligned} P_\lambda(f) = \|\Pi_{\geq\lambda}f\|_4^4 = \mathbb{E}\left(\sum_{\alpha\in S_{\geq\lambda}}\hat{f}_\alpha\chi_\alpha\right)^4 &= \sum_{\alpha,\beta,\alpha',\beta'}\hat{f}_\alpha\hat{f}_\beta\hat{f}_{\alpha'}\hat{f}_{\beta'}\mathbb{E}(\chi_\alpha\chi_\beta\chi_{\alpha'}\chi_{\beta'}) \\ &= \sum \hat{f}_\alpha\hat{f}_\beta\hat{f}_{\alpha'}\hat{f}_{\beta'}, \end{aligned}$$

where the last sum is over $\alpha,\beta,\alpha',\beta' \in S_{\geq\lambda}$ satisfying $\alpha+\beta = \alpha'+\beta'$. This shows that P_λ has nonnegative coefficients in the monomial basis corresponding to the eigenfunctions of G. Thus, we can apply our nonnnegative tensor maximization result to finish the proof. □

The preceding theorem also gives us the following approximation algorithm for small-set expansion on Cayley graphs.

Theorem 1.7. *For some absolute constant $C \geq 1$ and all $\mu,\epsilon > 0$ small enough, SOS relaxations of degree $K_\lambda(G)^{O(1)}\log n$ can distinguish between the following two cases with $\lambda = 1 - C\epsilon$.*

- *The Cayley graph G contains a vertex set of measure at most μ and expansion at most ϵ*
- *All vertex sets of measure at most $C/\sqrt{\mu}$ in G have expansion at least $1 - 1/C$.*

Proof. We claim that the maximum of $\|f\|_4/\|f\|_2$ over $f \in V_{\geq\lambda}$ distinguishes between the two cases. Combining with the previous theorem gives us a method of distinguishing between the two cases using the SOS framework.

- Let f be the indicator function of a set with measure at most μ and expansion at most ϵ. Then, $\|\Pi_{\geq\lambda}f\|^2 \geq 0.99\|f\|^2$. This means that $\|\Pi_{\geq\lambda}\|_4^4 \geq 0.9\|f\|_4^4$, so the L_4/L_2 ratio is bounded below by $\Omega(1)\cdot 1/\mu$.
- It is shown in [BBH] that graphs with this kind of small-set expansion satisfy $\|f\|_4^4/\|f\|_2^2 \ll 1/\mu$ for all functions $f \in V_{\geq\lambda}(G)$.

□

References

[Bar] Steurer Barak, Kelner. Rounding sum-of-squares relaxations.

[BBH] Hypercontractivity, sum-of-squares proofs, and their applications.