CS369H: Hierarchies of Integer Programming Relaxations	Spring 2016-2017

Lecture 6: General Sum-of-Squares and Tensor Decomposition

Professor Moses Charikar — *Scribes: Jimmy Wu*

Overview We discuss the sum of squares algorithm over non-Boolean domains and with constraints. We also introduce an application of sum-of-squares to tensor decomposition.

1 Sum-of-Squares over General Domains

Over the past few lectures, we have used the sum-of-squares (SOS) algorithm to perform unconstrainted polynomial optimization over the Boolean hypercube. However, it turns out that SOS is capable of dealing with an arbitrary domain $\Omega \subseteq \mathbb{R}^n$ described by polynomial inequalities.

Concretely, suppose we are given a set of variables $x = (x_1, \dots, x_n)$ and a set of constraints $A = \{f_1 \geq 0, \dots, f_m \geq 0\}$, where each $f_i \in \mathbb{R}[x]$ (the ring of polynomials with real coefficients). We would like to be able to do the following:

- Decide if A has a solution.
- Given some $g \in \mathbb{R}[x]$, decide if g is nonnegative over the set of solutions to A.

1.1 General Sum-of-Squares Proofs

We say that a polynomial $p \in \mathbb{R}[x]$ is *sum-of-squares* if $\exists q_1, \dots, q_r \in \mathbb{R}[x]$ such that $p = \sum_{i=1}^{r} q_i^2$.

Definition 6.1 (sum-of-squares proof)**.** *A sum-of-squares proof that the constraints A imply the nonnegativity of a polynomial g consists of SOS polynomials $(p_S)_{S \subseteq [m]}$ such that*

$$g = \sum_{S \subseteq [m]} p_S \cdot \prod_{i \in S} f_i \tag{1}$$

We say this proof has degree ℓ if each term in the above has degree at most ℓ, in which case we write

$$A \vdash_\ell \{g \geq 0\} \tag{2}$$

To see that an SOS proof is indeed a certificate of nonnegativity, consider any point x satisfying the constraints A. Since the p_S's are sum-of-squares and the f_i's are nonnegative by choice, $g \geq 0$. To emphasize this pointwise view, we may write (2) as

$$\{f_1(x) \geq 0, \dots, f_m(x) \geq 0\} \vdash_{x,\ell} \{g(x) \geq 0\} \tag{3}$$

So, an SOS proof is *sound* by definition. But is it *complete*? That is, can the SOS proof system always decide whether a given set of constraints is (in)feasible? The following theorem of Krivine [Kri64] and Stengle [Ste74] answers this in the affirmative:

Theorem 6.1 (Positivstellensatz)**.** *For every system of polynomial constraints* $A = \{f_1 \geq 0, \ldots, f_m \geq 0\}$*, either there exists a solution, or there exists an SOS proof that* $A \vdash_\ell \{-1 \geq 0\}$ *(i.e. a contradiction) for some* $\ell \in \mathbb{N}$*. We call this a* degree-ℓ sum-of-squares refutation *for* A*.*

Sum-of-squares proofs exhibit some nice composition properties:

- If $A \vdash_\ell \{f \geq 0, g \geq 0\}$, then we also have $A \vdash_\ell \{f + g \geq 0\}$.
- If $A \vdash_\ell \{f \geq 0\}$ and $A \vdash_{\ell'} \{g \geq 0\}$, then $A \vdash_{\ell \cdot \ell'} \{f \cdot g \geq 0\}$. To see this, take the SOS proofs for f and g and multiply them.
- Let A, B, and C be sets of polynomial constraints. If $A \vdash_\ell B$ and $B \vdash_\ell C$, then $A \vdash_{\ell \cdot \ell'} C$. Here, the proof for $A \vdash C$ comes from replacing the f_i's in the proof of $A \vdash B$ with degree-ℓ' polynomial from the $B \vdash C$ proof.

1.2 Pseudo-Distributions

Recall that given a function $\mu : \mathbb{R}^n \to \mathbb{R}$, we define the *formal expectation with respect to* μ as

$$\tilde{\mathbb{E}}_\mu = \sum_{x \in \text{support}(\mu)} f(x) \cdot \mu(x) \tag{4}$$

Definition 6.2 (pseudo-distribution)**.** $\mu : \mathbb{R}^n \to \mathbb{R}$ *is a* degree-d pseudo-distribution *if the following hold:*

- $\tilde{\mathbb{E}}_\mu 1 = 1$.
- $\tilde{\mathbb{E}}_\mu f^2 \geq 0$ *for all polynomials* f *with deg*$(f) \leq d/2$.

Lemma 6.1 (pseudo-moments)**.** *Let* $\mu : \mathbb{R}^n \to \mathbb{R}$ *be a function such that* $\tilde{\mathbb{E}}_\mu 1 = 1$*. Then*

$$\mu \textit{ is a degree-}d\textit{ pseudo-distribution} \iff \underbrace{\tilde{\mathbb{E}}_\mu \left((1,x)^{\otimes d/2}\right)\left((1,x)^{\otimes d/2}\right)^\top}_{\textit{degree-}d\textit{ moment matrix}} \textit{ is positive semidefinite} \tag{5}$$

where $(1, x)$ *is the vector* x *prepended with* 1*.*

Proof. ($\Longleftarrow$) Take any polynomial p and write it as $p(x) = \langle v, (1,x)^{\otimes d/2} \rangle$, where v is the vector of coefficients. Then since the moment matrix M is PSD, we have $\tilde{\mathbb{E}}_\mu (p(x))^2 = vMv^\top \geq 0$. Therefore, μ is a degree-d pseudo-distribution.

($\Longrightarrow$) Conversely if M is not PSD, then there exists v such that $vMv^\top < 0$, and examining the polynomial $p(x)$ with coefficients v, we see that $\tilde{\mathbb{E}}_\mu (p(x))^2 = vMv^\top < 0$. □

Definition 6.3 (pseudo-distribution satisfying constraints)**.** *Let* $A = \{f_1 \geq 0, \ldots, f_m \geq 0\}$ *be a set of constraints, and let* μ *be a degree-*d *pseudo-distribution. We say* μ satisfies A at degree $\ell \leq d$ *if for every set* $S \subseteq [m]$ *and SOS polynomial* h *such that deg*$(h) + \sum_{i \in S} \max(\text{deg} f_i, \ell)$ *satisfies*

$$\tilde{\mathbb{E}}_\mu \left[h \cdot \prod_{i \in S} f_i \right] \geq 0 \tag{6}$$

More succinctly, we write $\mu \vDash_\ell A$*.*

1.3 Duality

Theorem 6.2 (duality between SOS proofs and pseudodistributions)**.** *Let A be a set of polynomial constraints over $\mathbb{R}[x]$, and suppose $\|x\|^2 \leq M$ for some constant M (i.e. this is a constraint in A). For every even $d \in \mathbb{N}$ and every degree-d polynomial $f \in \mathbb{R}[x]$, either:*

- *$\forall \varepsilon > 0$, there exists a degree-d SOS proof that $A \vdash_d \{f \geq -\varepsilon\}$; or*
- *There exists a degree-d pseudo-distribution μ such that $\mu \vDash A$ and $\tilde{\mathbb{E}}_\mu f \leq 0$.*

See the sum-of-squares lecture notes by Barak and Steurer for a proof.

An alternate view of the duality theorem, implied by the above statement, is that

$$\sup\{c \in \mathbb{R} | A \vdash_d f \geq c\} = \min_{\substack{\text{degree-}d\text{ pseudo-}\\ \text{distributions } \mu \text{ s.t. } \mu \vDash_d A}} \tilde{\mathbb{E}}_\mu f. \tag{7}$$

Lemma 6.2 (completeness for composition of SOS proofs)**.** *Let $d \geq \ell' \geq \ell$. Let $A, B \subseteq \mathbb{R}[x]$ be systems of polynomial constraints, where A contains a boundedness constraint $M - \sum_{i=1}^{n} x_i^2 \geq 0$. Suppose every degree-$d$ pseudo-distribution μ that satisfies $\mu \vDash_\ell A$ also satisfies $\mu \vDash_{\ell'} B$. Then for every $\varepsilon > 0$, there exists an SOS proof that certifies*

$$A \vdash_d B_\varepsilon \tag{8}$$

where B_ε is obtained from B by weakening every constraint by ε.

Theorem 6.3 (general SOS algorithm)**.** *There exists an algorithm that, given $d \in \mathbb{N}$ and a satisfiable system of polynomial constraints A over $\mathbb{R}^n$, outputs in time $n^{O(d)}$ a degree-d pseudo-distribution that approximately satisfies $\mu \vDash A$ up to error 2^{-n}* [1].

2 Tensor Decomposition

Just as we would like to take a matrix $M \in \mathbb{R}^{n \times n}$ and write it as $M = \sum u_i v_i^\top = \sum u_i \otimes v_i$, we would like to be able to decompose tensors. For example, given a 3-tensor $T \in \mathbb{R}^{n \times n \times n}$, we would like to express it as $T = \sum u_i \otimes v_i \otimes w_i$.

Interest in tensor decomposition has come from a variety of fields, including psychometrics, chemometrics, and statistics.

For the rest of this section, we restrict our attention to $(n \times n \times n)$-dimensional tensors.

2.1 Tensor Rank

The rank of a tensor $T \in \mathbb{R}^{n \times n \times n}$ is trivially at most n^3. However, we can obtain a better upper bound of n^2. To see this, "slice" T into n matrices of size $n \times n$ and decompose each slice. This allows us to write

$$T = \sum_{i=1}^{n} (u_i \otimes v_i) \otimes \underbrace{(0 \cdots 0\ 1\ 0 \cdots 0)}_{1 \text{ in the } i\text{th coordinate}} \tag{9}$$

[1] Here, "approximately satisfies" means that we slightly weaken each of the constraints in (6).

The *tensor decomposition problem* is to compute, given a rank-r tensor, a rank-r decomposition. In general, this task is intractable. In particular, the following objective functions are $\mathbf{NP}$-hard to optimize:

[Hås90] $\min_{u_i, v_i, w_i} \left\| T - \sum_{i=1}^{k} u_i \otimes v_i \otimes w_i \right\|_2$

[HL13] $\max_{u,v,w} \langle u \otimes v \otimes w \rangle$

The reductions used to prove the above hardness results construct tensors with rank $\Omega(n^2)$; therefore we might hope for better structure among low-rank tensors. Indeed, for tensors of rank up to $3n/2$, reasonable conditions for having a unique decomposition are known [Kru77]. Efficient algorithms, meanwhile, are known for rank up to n. For the special case of *generic* tensors, uniqueness is known to hold up to rank $\Omega(n^2)$.

2.2 Jennrich's Algorithm

We now sketch an algorithm for decomposing a 3-tensor under mild assumptions. This algorithm is attributed to Robert Jennrich, but was first presented by Harshman [Har70] and later generalized by Leurgans, Ross, and Abel [LRA93].

Let T be a tensor that can be decomposed as $T = \sum_{i=1}^{r} a_i^{\otimes 3}$, where the a_i's are orthogonal.[2] To motivate the algorithm, examine the kth slice of T. Its contribution is

$$\sum_{i=1}^{r} (a_i \otimes a_i) a_{ik} \tag{10}$$

for some a_{ik}. An issue that arises, however, is that the a_{ik}'s are not necessarily unique. Jennrich's algorithm gets around this as follows:

1. Pick a random vector $v \in \mathbb{R}^n$. Flatten T into an $n^2 \times n$ matrix and compute

$$M := Tv = \sum_{i=1}^{r} (a_i \otimes a_i) \langle a_i, v \rangle \tag{11}$$

   Since the a_i's are orthogonal, this describes a singular value decomposition. Moreover since v is random, the $\langle a_i, v \rangle$'s are distinct with high probability, so the SVD is unique.

2. Use SVD to compute

$$M = \sum_{i=1}^{r} \lambda_i (w_i \otimes w_i) \tag{12}$$

[2]With some modifications, Jennrich's algorithm works more generally for tensors $T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i$, where $\{u_i\}$ is linearly independent, $\{v_i\}$ is linearly independent, and no w_i is a scalar multiple of another w_j.

3. Find coefficients $\alpha_1, \ldots, \alpha_r$ such that

$$M = \sum_{i=1}^{r} \alpha_i w_i^{\otimes 3} \tag{13}$$

which is a linear system in which the α_i's are the variables.

More algorithms for tensor decomposition are described in the lecture notes by Barak and Steurer. In particular, they describe a "brute data" algorithm that works when we are given a large number of observations, as well as an SOS algorithm that produces "fake moments" to compensate for lack of observations [BKS15].

References

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 143–151. ACM, 2015.

[Har70] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis. 1970.

[Hås90] Johan Håstad. Tensor rank is np-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

[HL13] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[Kri64] Jean-Louis Krivine. Anneaux préordonnés. *Journal d'analyse mathématique*, 12(1):307–326, 1964.

[Kru77] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

[LRA93] SE Leurgans, RT Ross, and RB Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

[Ste74] Gilbert Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974.