AMS Short Course January 15th, 2019

Lecture 1: Rounding Sum-of-Squares

Prof. Ankur Moitra *Scribe:*

1 Introduction

In this lecture we will show how the Sum-of-Squares hierarchy can be used to give efficiently computable convex programming relaxations to NP-hard optimization problems.

2 MAXCUT

We will start with the famous MAXCUT problem:

Definition 1. *Given an unweighted graph $G = (V, E)$, the goal is to find a subset $U \subset V$ that maximizes $|E(U, V \setminus U)|$, where $E(S, T)$ is the set of edges with one endpoint in S and the other endpoint in T.*

This is a classic NP-hard optimization problem. In fact, it was on Karp's [8] original list of 21 NP-complete problems. (More precisely, he showed that the weighted version was NP-hard and this was relaxed to the unweighted problem in later work). You can also think of this problem as asking: What is the largest subgraph of G (where we count the number of edges included in the subgraph) that is bipartite?

Since MAXCUT is NP-hard, we will focus on designing approximation algorithms for it. More precisely, our goal is to give an algorithm that finds a cut $U \subset V$ where $|E(U, V \setminus U)|$ is at least an α factor times the optimal value. We call this an α-approximation algorithm. First, we note that it is trivial to get an $1/2$-approximation algorithm:

Claim 2. *There is a randomized polynomial time $1/2$-approximation algorithm for MAXCUT.*

Proof. We can construct U randomly as follows: For each node in V, include it in U with probability $1/2$ and otherwise leave it out. Do this independently for each node. Then it is easy to see that the probability that any edge (u, v) ends up crossing the cut – i.e. having one endpoint in U and the other in $V \setminus U$ – is exactly $1/2$. And now by linearity of expectation, we have that the expected number of edges in $|E(U, V \setminus U)|$ is at least $1/2|E|$, and certainly the optimum is at most $|E|$. □

For a long time, this was the best known approximation algorithm. Even various linear programming relaxations turn out to not do any better.

In a seminal work, that introduced semidefinite programming into approximation algorithms, Goemans and Williamson [6] gave a α-approximation algorithm for $\alpha \geq 0.878$. While their relaxation and its analysis does not need any of the machinery of Sum-of-Squares, it turns out to be a great place to start – both in understanding how Sum-of-Squares gives a general family of relaxations

for various NP-hard graph partitioning problems, and also in how to round a feasible solution in the semidefinite program into a feasible solution for the original problem. This latter part will be surprisingly subtle when we move beyond MAXCUT, and we will elaborate on this later.

We will start by formulating MAXCUT as a polynomial optimization problem. We can write it as

$$\begin{array}{ll} \max & \sum_{(i,j)\in E} (x_i - x_j)^2 \\ \text{subject to} & x_i^2 = x_i, \text{ for all } i \end{array}$$

The constraint $x_i^2 = x_i$ implies that x_i is either zero or one. Then we can think of the set of x_i's that are one as the set U, and the objective function contributes one for each edge in $E(U, V \setminus U)$ and zero otherwise.

We have already seen the Sum-of-Squares hierarchy introduced as a relaxation for solving polynomial optimization problems over polynomial constraints. Here we will take a dual view. Instead of thinking of it as a proof system that becomes more powerful as we increase the degree of the proofs that we allow, we will think of it as producing an operator that we will call a *pseudo-expectation* operator. You should think of this as acting like a distribution solutions – which in our case are assignments to the variables that we can map into cuts.

Definition 3. *A degree d pseudo-expectation $\widetilde{\mathbb{E}}$ is an operator*

$$\widetilde{\mathbb{E}} : \mathcal{P}_n^{\leq d} \to \mathbb{R}$$

where $\mathcal{P}_n^{\leq d}$ is the set of degree at most d polynomials in variables $x_1, x_2, \cdots, x_n$ that satisfies the following conditions:

(1) $\widetilde{\mathbb{E}}$ *is linear*

(2) $\widetilde{\mathbb{E}}[1] = 1$

(3) $\widetilde{\mathbb{E}}[p^2] \geq 0$ *for any polynomial p of degree at most $d/2$.*

(4) $\widetilde{\mathbb{E}}[x_i^2 p] = \widetilde{\mathbb{E}}[x_i p]$ *for any polynomial p of degree at most $d-2$.*

Now we can write down our relaxation for MAXCUT:

$$\begin{array}{ll} \max & \widetilde{\mathbb{E}}[\sum_{(i,j)\in E} (x_i - x_j)^2] \\ \text{subject to} & \widetilde{\mathbb{E}} \text{ is a degree } d \text{ pseudo-expectation for MAXCUT} \end{array}$$

Now, how should you think about this complicated object? What it's trying to do is act like a distribution on feasible solutions. So let's see how a distribution on feasible solutions leads to a pseudo-expectation operator:

Lemma 4. *If there is a distribution on cuts $U \subset V$ where the expected value of $|E(U, V \setminus U)| \geq k$, then there is a degree d pseudo-expectation that satisfies*

$$\widetilde{\mathbb{E}}[\sum_{(i,j)\in E} (x_i - x_j)^2] \geq k$$

Proof. We can construct the pseudo-expectation operator explicitly. For any polynomial $p = p(x_1, \cdots, x_n)$ we will set

$$\widetilde{\mathbb{E}}[p] = \mathbb{E}_U[p(1_{x_1 \in U}, \cdots, 1_{x_n \in U}]$$

We leave it as an exercise to the reader to check that all of the conditions in Definition 3 are satisfied, and that the objective value is at least k. □

Now if we could maximize the polynomial $\sum_{(i,j)\in E} (x_i - x_j)^2$ over all cuts, we would be done. We could solve MAXCUT optimally and $P = NP$ and we could all go home. But there are some issues. You cannot even write down a distribution on cuts succinctly because you would need to specify 2^n values, one for each cut. What the pseudo-expectation does to circumvent this is it only asks you to write down enough information about the distribution so that we can evaluate it on degree up to d polynomials. More precisely, knowing that the expectation is linear (and same with the pseudo-expectation) you could write down one number for each of the $O(n^d)$ monomials.

The rest of the constraints in the pseudo-expectation are just safety checks. If $\widetilde{\mathbb{E}}$ really were the operator that maps a polynomial to its expectation under a distribution on cuts, then certainly $\widetilde{\mathbb{E}}[1] = 1$. Also $\widetilde{\mathbb{E}}[x_i^2 p] = \widetilde{\mathbb{E}}[x_i p]$ because the distribution only places non-zero probability on assignments to the variables x_i that are zero or one valued. The most powerful constraint is that $\widetilde{\mathbb{E}}[p^2] \geq 0$. This is both what leads to state-of-the-art algorithms but also the hardest condition to verify when you want to prove lower bounds.

3 Rounding SOS for MAXCUT

We will show the following theorem:

Theorem 5. *Any feasible solution to the degree 2 Sum-of-Squares relaxation for MAXCUT that achieves objective value k can be rounded to a cut U where*

$$\mathbb{E}[|E(U, V \setminus U)|] \geq \alpha k$$

where $\alpha = \min_{-1 \leq \rho \leq 1} \frac{2 \arccos \rho}{(1-\rho)\pi} \geq 0.878$.

The constant above is often called the Goemans-Williamson [6] constant. It turns out to be at the heart of some of the deepest mysteries in theoretical computer science, like the validity of the *Unique Games Conjecture*.

The main question is: How do we round a pseudo-expectation to find a cut? In the case of a degree 2 pseudo-expectation this turns out to be quite easy. But even for degree 3 the fundamental approach we are using will not work. With this as a caveat, the idea is to sample from a Gaussian distribution whose moments match the pseudo-moments (that is, the evaluation of the pseudo-expectation operator on degree up to two monomials).

Claim 6. *Without loss of generality, we can assume that for all i*

$$\widetilde{\mathbb{E}}[x_i] = 1/2$$

The intuition for this claim is that the cuts U and $V \setminus U$ have the same objective value. So we might as well as put the same probability on both, in which case for each node i, the probability it is in U is exactly $1/2$. Anyways, the useful information about the cut that pertains to its objective value is the correlation between pairs of nodes of which side they belong to.

Proof. Given a pseudo-expectation operator $\widetilde{\mathbb{E}}$ we can construct a new one as

$$\widetilde{\mathbb{E}}'[p(x_1, \cdots, x_n)] = \frac{1}{2}\widetilde{\mathbb{E}}[p(x_1, \cdots, x_n)] + \frac{1}{2}\widetilde{\mathbb{E}}[p(1 - x_1, \cdots, 1 - x_n)]$$

Again, we leave it as an exercise to the reader to check that if $\widetilde{\mathbb{E}}$ satisfies the conditions of being a degree 2 pseudo-expectation operator then $\widetilde{\mathbb{E}}'$ does too. Moreover one can also check that $\widetilde{\mathbb{E}}'$ achieves the same objective value. □

Now we get to the most interesting part of the proof. We can sample a Gaussian random variable y in n dimensions with mean μ and covariance Σ where

$$\mu = \widetilde{\mathbb{E}}[x] \text{ and } \Sigma = \widetilde{\mathbb{E}}[(x - \mu)(x - \mu)^T]$$

where x is a vector with variables $x_1, \cdots, x_n$. Now we construct the set U by including nodes for which $y_i \leq 1/2$ and leaving out the rest. This is called Gaussian rounding. We will show that

$$\mathbb{E}[|E(U, V \setminus U)|] \geq \alpha k$$

where α is the Goemans-Williamson constant and k is the objective value of the pseudo-expectation. We will break up the analysis into two parts. First we will calculate the contribution of an edge (i, j) to the objective of the relaxation. For notational convenience, we define

$$\rho = 4\widetilde{\mathbb{E}}[x_i x_j] - 1$$

Lemma 7. $\widetilde{\mathbb{E}}[(x_i - x_j)^2] = \frac{1-\rho}{2}$

Proof. We can compute

$$\begin{aligned} \widetilde{\mathbb{E}}[(x_i - x_j)^2] &= \widetilde{\mathbb{E}}[x_i^2] - 2\widetilde{\mathbb{E}}[x_i x_j] + \widetilde{\mathbb{E}}[x_j^2] \\ &= \widetilde{\mathbb{E}}[x_i] - 2\widetilde{\mathbb{E}}[x_i x_j] + \widetilde{\mathbb{E}}[x_j] = \frac{1 - \rho}{2} \end{aligned}$$

where the second equality comes from the constraint $\widetilde{\mathbb{E}}[x_i^2] = \widetilde{\mathbb{E}}[x_i]$ for all i and the last equality comes from using the fact that $\widetilde{\mathbb{E}}[x_i] = 1/2$. □

Now to complete the analysis, we need to calculate the contribution of the same edge to the number of edges in the cut:

Lemma 8. $\mathbb{P}[(i, j) \in E(U, V \setminus U)] = \frac{\arccos \rho}{\pi}$

Proof. First we compute the variance of y_i:

$$\text{Var}(y_i) = \widetilde{\mathbb{E}}[(x_i - \frac{1}{2})^2] = \widetilde{\mathbb{E}}[x_i] - \frac{1}{4} = \frac{1}{4}$$

and similarly

$$\mathrm{Cov}(y_i, y_j) = \frac{\rho}{4}$$

which follows from the way that we have defined Σ and ρ. Finally we can use a standard fact about correlated Gaussian random variables that

$$\mathbb{P}[\mathrm{sgn}(y_i - \frac{1}{2}) \neq \mathrm{sgn}(y_j - \frac{1}{2})] = \mathbb{P}[\mathrm{sgn}(s) \neq \mathrm{sgn}(\rho s + \sqrt{1-\rho^2}t)]$$

where s and t are independent Gaussians with mean zero and variance one. The right hand side is well known to be $\frac{\arccos \rho}{\pi}$, which completes the proof. □

Now we are ready to prove Theorem 5:

Proof. For every edge (i, j) we have

$$\mathbb{P}[(i,j) \in E(U, V \setminus U)] \geq \frac{2 \arccos \rho}{(1-\rho)\pi} \widetilde{\mathbb{E}}[(x_i - x_j)^2]$$

and, summing over all edges, and by linearity of expectation and pseudo-expectation, this completes the proof. □

3.1 Open Questions

Despite its simplicity, this is the best known approximation algorithm for MAXCUT. In fact any approximation algorithm that achieves a strictly better approximation ratio (independent of the number of nodes) would have spectacular consequences.

In 1994, Arora et al. [1, 2] showed that any proof of length n can be turned into a proof of length polynomial in n in such a way that you can verify its correctness, not by reading it all, but by querying a constant number of bits randomly. These are called probabilistically checkable proofs, and their famous PCP theorem turns out to be the key for showing that various optimization problems are not only NP-hard to solve exactly but even to solve approximately. Over the years, there were many improvements to their construction that had corresponding improvements in hardness of approximation. For example, Hastad [7] showed that for any $\epsilon > 0$, it is NP-hard to distinguish between 3-SAT formulas that are perfectly satisfiable and ones where it is possible to satisfy at most a $7/8 + \epsilon$ fraction of the clauses. This is tight in the sense that for any 3-SAT formula, you can always satisfy at least $7/8$ of its clauses.

In 2002, Khot [10] formulated a powerful conjecture called the *Unique Games Conjecture*. In some ways, it sounds like a slight modification of the types of NP-hard problems that arise in the PCP literature. It is still unknown whether the problems remain NP-hard under these modifications. However, over the years, the community has discovered that this one conjecture leads to an amazing number of tight hardness of approximation results. Most relevant to us, Khot et al. [11] and Mossel et al. [13] proved:

Theorem 9. *If the Unique Games Conjecture is true then for any $\epsilon > 0$, it is NP-hard to approximate MAXCUT within $\alpha + \epsilon$, where α is the Goemans-Williamson constant.*

Amazingly, for many of the most natural approaches towards improving on the Goemans-Williamson approximation algorithm and refuting the Unique Games Conjecture, it is not known whether they succeed or fail. We showed how to round the degree 2 Sum-of-Squares relaxation.

Open Question 10. *Does the Sum-of-Squares relaxation for MAXCUT give an approximation algorithm that beats the Goemans-Williamson constant for any constant degree?*

In fact, even for degree 4 this is open. In theoretical computer science, the Sum-of-Squares hierarchy plays a unique role: Lower bounds against it are often a source of evidence for a problem being computationally hard. There was a time when there were candidate hard instances of unique games for weaker semidefinite programming hierarchies, and then Barak et al. showed that Sum-of-Squares solves all of them. In this way, sometimes the beliefs of the community get swayed by how the story plays out for Sum-of-Squares.

4 Rounding Higher Degree SOS

There are a number of optimization problems where the state-of-the-art algorithms come from Sum-of-Squares. Many of these require degree higher than two. However rounding higher degree Sum-of-Squares relaxations is quite challenging. Our approach of finding a distribution (in our case a Gaussian) that matches the pseudo-moments is doomed to fail for degree three and higher (we will see this as a corollary of the lower bounds in the next lecture).

4.1 Inequalities Derivable in SOS

So how can we round higher degree Sum-of-Squares relaxations? Most of the rounding algorithms revolve around inequalities that are true for distributions, but can also be proven for pseudo-distributions. Following Barak et al. [3] we will show the Cauchy-Schwatz inequality can be derived in Sum-of-Squares, and then we'll use it to give an interesting algorithm for tensor optimization.

The classic Cauchy-Schwartz inequality tells us:

Fact 11. *If $p(x_1, \cdots, x_n)$ and $q(x_1, \cdots, x_n)$ are polynomials and μ is a distribution on $x_1, \cdots, x_n$ then*

$$\mathbb{E}_\mu[p(x)q(x)] \leq \sqrt{\mathbb{E}_\mu[p(x)^2]\mathbb{E}_\mu[q(x)^2]}$$

We will prove that this inequality is true for pseudo-expectations too:

Lemma 12. *If $p(x_1, \cdots, x_n)$ and $q(x_1, \cdots, x_n)$ are polynomials of degree at most $d/2$ and $\widetilde{\mathbb{E}}$ is a degree d pseudo-expectation then*

$$\widetilde{\mathbb{E}}[p(x)q(x)] \leq \sqrt{\widetilde{\mathbb{E}}[p(x)^2]\widetilde{\mathbb{E}}[q(x)^2]}$$

When we say pseudo-distribution above we are referring to just the constraints (1), (2) and (3). These are the constraints we will always enforce, regardless of what polynomial constraints come from the optimization problem we want to solve. When we defined the pseudo-expectation for MAXCUT we augmented these constraints with (4) which comes from the fact that we are looking for zero or one valued solutions, along with an objective function that is supposed to count the size of a cut.

Proof. Suppose that $\widetilde{\mathbb{E}}[p(x)^2], \widetilde{\mathbb{E}}[q(x)^2] > 0$. In this case, by rescaling, we can assume without loss of generality that

$$\widetilde{\mathbb{E}}[p(x)^2] = \widetilde{\mathbb{E}}[q(x)^2] = 1$$

Now using constraint (3) on pseudo-expectations we have $\widetilde{\mathbb{E}}[(p(x) - q(x))^2] \geq 0$ and by expanding we have

$$\widetilde{\mathbb{E}}[p(x)q(x)] \leq \frac{1}{2}(\widetilde{\mathbb{E}}[p(x)^2] + \widetilde{\mathbb{E}}[q(x)^2]) = 1 = \sqrt{\widetilde{\mathbb{E}}[p(x)^2]\widetilde{\mathbb{E}}[q(x)^2]}$$

which completes the proof. □

In general, when you can prove an inequality using a low degree Sum-of-Squares proof then it will hold when you replace expectations by pseudo-expectations. Barak et al. [3] showed that you can prove Holder's inequality in Sum-of-Squares. More powerfully, they also showed that you can prove the hypercontractive inequality which implies that for any degree polynomial p of degree at most d

$$\widetilde{\mathbb{E}}[p(x)^4] \leq 9^d(\widetilde{\mathbb{E}}[p(x)^2])^2$$

for any degree $4d$ pseudo-expectation. Kauers et al. [9] gave a Sum-of-Squares proof for a reverse form of the hypercontractive inequality. De et al. [5] gave a Sum-of-Squares proof for the Majority is Stablest Theorem which states that for any function whose influences are going to zero, majority is the function that has the largest chance of not changing its value when you apply the noise stability operator. Finally, Lei and Sheng [12] gave a Sum-of-Squares proof for the powerful Brascamp-Lieb inequality.

4.2 Tensor Optimization

Here we will explain the main ideas in an algorithm of Barak, Kelner and Steurer [4] that uses higher degree Sum-of-Squares to approximately optimize the injective norm of a tensor. More precisely, they show:

Theorem 13. *Given an entry-wise nonnegative, symmetric matrix $M \in \mathbb{R}^{n^2 \times n^2}$, there is an $n^{O(\log n/\epsilon^2)}$ time algorithm to find a unit vector x with*

$$\langle x^{\otimes 2}, Mx^{\otimes 2}\rangle \geq \max_{z \ s.t. \ \|z\| \leq 1} \langle z^{\otimes 2}, Mz^{\otimes 2} - \epsilon\|M\|_1$$

Here $\|M\|_1$ is the trace norm and is the sum of the entries along the diagonal (equivalently, it is the sum of the singular values). Even though the result is described as an optimization problem over matrices, you can think of M as coming from an $n \times n \times n \times n$ tensor, in which case we are approximating the quartic form over all unit vectors.

We can once again define the notion of a pseudo-distribution, taking constraints (1), (2) and (3) as usual and instead using

(4') $\widetilde{\mathbb{E}}[p(x)\sum_{a=1}^n x_a^2] = \widetilde{\mathbb{E}}[p(x)]$

which comes from the polynomial constraint $\sum_{a=1}^n x_a^2 = 1$ specific to the optimization problem at hand.

The basic idea pioneered by Barak, Kelner and Steurer [4] is to pretend that the pseudo-distribution is actually a distribution and use its moments to hone in on a feasible solution. And then, if the only inequalities you used are derivable in Sum-of-Squares, reformulate it as a way to round. Essentially all of the later algorithms for utilizing Sum-of-Squares follow this paradigm. It feels almost like a cheat the first time (or few times) you see it.

In this specific tensor optimization problem, we will first construct a unit vector from the pseudo-moments in a simple way. If it works, then great! If not, we will have to find some other way to make progress. In particular suppose that we have found a pseudo-expectation (of some unspecified degree, for now) with objective value γ – i.e.

$$\widetilde{\mathbb{E}}[\sum_{a,b,c,d} M_{a,b,c,d} x_a x_b x_c x_d] = \gamma$$

Then we will define a n dimensional vector t with $t_a = \sqrt{\widetilde{\mathbb{E}}[x_a^2]}$. First, we will prove:

Claim 14. *t is a unit vector*

Proof. We can compute

$$\sum_a t_a^2 = \sum_a \widetilde{\mathbb{E}}[x_a^2] = \widetilde{\sum_a} \mathbb{E}[x_a^2] = 1$$

where second equality comes from constraint (1) and the third comes from constraint (4′). □

Now suppose we get lucky and t satisfies

$$\langle t^{\otimes 2}, M t^{\otimes 2} \rangle \geq \gamma - \epsilon \|M\|_1$$

Then we are done because

$$\gamma \geq \max_{z \text{ s.t. } \|z\| \leq 1} \langle z^{\otimes 2}, M z^{\otimes 2}$$

because the maximum over pseudo-distributions is at least as large as the maximum over unit vectors. The main question is: What do we do if we do not get lucky?

We can define an n^2 dimensional vector s with $s_{a,b} = \sqrt{\widetilde{\mathbb{E}}[x_a^2 x_b^2]}$. We leave it as an exercise to the reader to check that s is also a unit vector.

Lemma 15. *If $\langle t^{\otimes 2}, M t^{\otimes 2} \rangle \leq \gamma - \epsilon \|M\|_1$ then*

$$\|s - t^{\otimes 2}\|_2 \geq \frac{\epsilon}{2}$$

Proof. First we have

$$\gamma = \widetilde{\mathbb{E}}[\sum_{a,b,c,d} M_{a,b,c,d} x_a x_b x_c x_d] \leq \sum_{a,b,c,d} M_{a,b,c,d} \sqrt{\widetilde{\mathbb{E}}[x_a^2 x_b^2] \widetilde{\mathbb{E}}[x_c^2 x_d^2]} = s^T M s$$

The inequality above comes from Lemma 12. Now rearranging, we have

$$\epsilon \|M\|_1 \leq s^T M s - \langle t^{\otimes 2}, M t^{\otimes 2} = \langle s - t^{\otimes 2}, M(s + t^{\otimes 2}) \rangle \leq 2 \|M\|_1 \|s - t^{\otimes 2}\|_2$$

where the last inequality uses the fact that $\|s + t^{\otimes 2}\|_2 \leq 2$ because both s and t are unit vectors. This completes the proof. □

Now it turns out that we can interpret $\|s - t^{\otimes 2}\|_2$ information theoretically. In particular, we can define two random variables W and W' on $[n] \times [n]$ as follows:

- $\mathbb{P}[W = (a,b)] = \widetilde{\mathbb{E}}[x_a^2 x_b^2] = s_{a,b}^2$
- $\mathbb{P}[W' = (a,b)] = \widetilde{\mathbb{E}}[x_a^2]\widetilde{\mathbb{E}}[x_b^2] = t_a^2 t_b^2$

The important point is that both W and W' have the same marginal distribution on their first (and second) coordinate. By construction, the two coordinates of W' are independent of each other. Now the point is that $\|s - t^{\otimes 2}\|_2$ being lower bounded is the same thing as W and W' being far from each other, in an appropriate choice of distance between random variables. And the only way for them to be far is if the two coordinates of W are somewhat dependent. More precisely we have

$$\|s - t^{\otimes 2}\|_2 = \sum_{(a,b)} (\sqrt{\mathbb{P}[W = (a,b)]} - \sqrt{\mathbb{P}[W' = (a,b)]})$$

which is (up to a factor of two) exactly the squared Hellinger distance between W and W'. Using various relationships between Hellinger distance and mutual information, it can be shown that:

Fact 16. *$\|s - t^{\otimes 2}\|_2 \geq \frac{\epsilon}{2}$ implies $I(w_1; w_2) \geq \frac{\epsilon^2}{8}$ where w_1 and w_2 are the two coordinates of W.*

Recall that $I(w_1; w_2) = H(w_1) - H(w_1|w_2) = H(w_2) - H(w_2|w_1)$ where H is the entropy function, and so the mutual information can be thought of as how much the variability of w_1 reduces (in expectation) when you condition on w_2.

So now we can explain how to complete the rounding procedure. When our first attempt fails, we know that the pseudo-expectation is making coordinates non-negligibly dependent on each other. So we will construct another unit vector using the following two-step procedure:

- Choose i in $[n]$ according to the distribution
$$(\widetilde{\mathbb{E}}[x_1^2], \cdots, \widetilde{\mathbb{E}}[x_n^2])$$
- Define a n dimensional vector u with
$$u_a = \sqrt{\frac{\widetilde{\mathbb{E}}[x_i^2 x_a^2]}{\widetilde{\mathbb{E}}[x_i^2]}}$$

Using the same arguments as before it is easy to show that for any choice of i in the first step, u is a unit vector. Similarly to what we did before, we can use u to define a distribution W'' on $[n]$ where

$$\mathbb{P}(W'' = a) = u_a^2$$

The important point is the following:

Fact 17. *The distribution of W'' is the same as the distribution of w_1 conditional on $w_2 = i$.*

We leave this as an exercise to the reader. Now using what we know about the mutual information between w_1 and w_2, we have that

$$\mathbb{E}_i[H(W'')] = \mathbb{E}_i[H(w_1|w_2 = i)] = H(w_1|w_2) = H(w_1) - I(w_1; w_2) \leq \log n - \frac{\epsilon^2}{8}$$

Now our second attempt u could also fail and not achieve objective value at least $\gamma - \epsilon$. But we could continue the procedure, and every time we are making progress because we are reducing the entropy of some associated distribution on $[n]$. It starts out being at most $\log n$ and it is always nonnegative. So eventually, after at most $\frac{8 \log n}{\epsilon^2}$ rounds, it must terminate.

Now in order to run this procedure, we need our pseudo-expectation to be defined up to degree at least $\frac{16 \log n}{\epsilon^2}$. In the first step, we only used the pseudo-expectation on degree two polynomials. In the second step, we needed it on degree four polynomials, and so on.

This hopefully gives you some sense for how to go about rounding higher degree Sum-of-Squares relaxations. Not only that, but the more inequalities we know how to prove using low degree Sum-of-Squares proofs, the more sophisticated types of reasoning we can utilize when we are working with pseudo-expectations, which in turn yield powerful algorithms that would be quite hard to discover without this machinery in hand.

References

[1] S. Arora, C. Lund, R. Motwani, R. Sudan and M. Szegedy. Proof verification and the hardness of approximation problems. In *Journal of the ACM*, 45(3):501–555, 1998.

[2] S. Arora and S. Safra. Probabilistic checking of proofs: A new characterization of NP. In *Journal of the ACM*, 45(1):70–122, 1998.

[3] B. Barak, F. Brandao, A. Harrow, J. Kelner, D. Steurer and Y. Zhou. Hypercontractivity, Sum-of-Squares proofs, and their applications. In *Symposium on Theory of Computing*, pages 307–326, 2012.

[4] B. Barak, J. Kelner and D. Steurer. Rounding Sum-of-Squares relaxations. In *Symposium on Theory of Computing*, pages 31–40, 2014.

[5] A. De, E. Mossel and J. Neeman. Majority is stablest: Discrete and SoS. In *Symposium on Theory of Computing*, pages 477–486, 2013.

[6] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. In *Journal of the ACM*, 42(6):1115–1145, 1995.

[7] J. Hastad. Some optimal inapproximability results. In *Journal of the ACM*, 48(4):798–859, 2001.

[8] R. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103, 1972.

[9] M. Kauers, R. O'Donnell, L-Y Tan and Y. Zhou. Hypercontractive inequalities via SOS, and the Frankl-Rodl graph. In *Symposium on Discrete Algorithms*, pages 1644–1658, 2014.

[10] S. Khot. On the power of unique 2-prover 1-round games. In *Symposium on Theory of Computing*, pages 767–775, 2002.

[11] S. Khot, G. Kindler, E. Mossel and R. O'Donnell. Optimal inapproximability results for MAXCUT and other 2-variable CSPs? In *SIAM Journal on Computing*, 37(1):319–357, 2007.

[12] Z. Lei and Y. Sheng. Sum of Squares proof for Brascamp-Lieb type inequality. *ArXiv*:1710.01458, 2017.

[13] E. Mossel, R. O'Donnell and K. Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. In *Annals of Mathematics*, 171(1):295–341, 2010.