Pseudo-calibration for Planted Clique Sum-of-Squares Lower Bound

Pasin Manurangsi[*]
UC Berkeley

August 13, 2016

Abstract

Recently, Barak, Hopkins, Kelner, Kothari, Moitra and Potechin [BHKKMP16] proved an essentially tight Sum-of-Squares lower bound for the *planted clique* problem. Their result can be divided into two main parts: coming up with the *pseudo-distribution* and proving positivity of such pseudo-distribution. In this short blog, we summarize the first part of the paper, which provides a general systematic way to come up with pseudo-distributions for problems other than the planted clique problem, without going into details of the proof. We do not touch on the second part, which is more technically involved, here but we will hopefully do so in future posts.

1 SoS Lower Bounds and the Planted Clique Problem

In this section, we provide some background for readers unfamiliar with proving Sum-of-Squares lower bounds and the planted clique problem. Those who are accustomed to the topic can skip this section. For SoS, we use notations from Tselil's blog on Sum-of-Squares Hierarchy, which is also a good place to start for those unfamiliar with SoS Hierarchy.

In this blog, we do not need the optimization version of SoS Hierarchy but we will only use a feasibility one. Recall that, given a polynomial feasibility problem of the form

$$Q = \{x \in \mathbb{R}^n : \forall i \in [m], g_i(x) = 0\},$$

the degree-$2d$ Sum-of-Squares relaxation of the problem, which can be solved in $n^{O(d)}$ time, is[1]

$$\mathrm{sos}_d(Q) = \left\{ \tilde{\mathbb{E}} : \begin{array}{c} \tilde{\mathbb{E}} : \{q : \deg(q) \leq 2d\} \to \mathbb{R} \text{ is a linear operator with } \tilde{\mathbb{E}}[1] = 1, \\ \forall q \text{ with } \deg(q) \leq d, \tilde{\mathbb{E}}[q^2] \geq 0, \\ \forall i \in [m]\ \forall q \text{ with } \deg(q) \leq 2d - \deg(g_i), \tilde{\mathbb{E}}[g_i q] = 0. \end{array} \right\}. \tag{1}$$

Roughly speaking, if we want to show that degree-$2d$ SoS fails to certify that a polynomial feasibility problem Q is infeasible, we need to come up with a degree-$2d$ pseudo-distribution $\tilde{\mathbb{E}}$ that satisfies the conditions in (1). For concreteness, let us consider the *planted clique* problem defined as follows.

Definition 1 (Planted Clique(n, k)) *Given as an input a graph $G = (V, E)$ drawn from one of the two following distributions (each with probability $1/2$):*

(1) $\mathcal{G}(n, 1/2)$: the Erdos-Renyi random graph of n vertices where each edge is included with probability 1/2,

[*] Email: `pasin@berkeley.edu`.

[1] Note that, in Tselil's blog, the positivity condition is written as the pseudo-moment matrix being positive semidefinite but it is not hard to see that this is the same as requiring that $\tilde{\mathbb{E}}[q^2] \geq 0$ for every q with $\deg(q) \leq d$.

(2) $\mathcal{G}(n, 1/2, k)$: the planted distribution, in which a graph G is first drawn from $\mathcal{G}(n, 1/2)$. Then, k vertices of G are chosen uniformly at random and an edge between each pair of chosen vertices are added to G.

The goal is to determine, with correctness probability $1/2 + \varepsilon$ for some constant $\varepsilon > 0$, which distribution G is drawn from.

In this blog, we always restrict ourselves to the case where $k \gg \log n$ so that the maximum clique sizes of the two cases are different. Since the largest clique in $\mathcal{G}(n, 1/2)$ is of size $O(\log n)$ with high probability, brute-force search solves the planted clique problem with high probability in $n^{O(\log n)}$ time. On the other hand, the best known polynomial-time algorithm works only when $k = \Omega(\sqrt{n})$ [AKS98]. A natural question is of course whether the SoS Hierarchy can do any better than this.

The most widely-used formulation of planted clique in terms of polynomial feasibility, and the one used in [BHKKMP16], is to formulate it as "does G have a clique of size k?". For convenience, let $V = [n] = \{1, \ldots, n\}$. This formulation can be written as follows.

$$\mathrm{CLIQUE}_k(G) = \left\{ x \in \mathbb{R}^n : \begin{array}{c} \forall i \in [n], x_i^2 = x_i, \\ \forall (i,j) \notin E, x_i x_j = 0, \\ \sum_{i \in [n]} x_i = k \end{array} \right\}$$

When the constraints are satisfied, x_i is simply a boolean indicator variable whether i is included in the clique. If we can solve $\mathrm{CLIQUE}_k(G)$ in polynomial time, then we are done because $G \sim \mathcal{G}(n, 1/2, k)$ always has clique of size k whereas the maximum clique of $G \sim \mathcal{G}(n, 1/2)$ is of size $O(\log n)$ w.h.p. Thus, there is always a solution in $\mathrm{CLIQUE}_k(G)$ for $G \sim \mathcal{G}(n, 1/2, k)$ but, w.h.p., there is no feasible solution for $G \sim \mathcal{G}(n, 1/2)$. But of course solving $\mathrm{CLIQUE}_k(G)$ is NP-hard so we will try to relax it using degree-$2d$ SoS which we can solve in n^{2d} time.

Again, when $G \sim \mathcal{G}(n, 1/2, k)$, $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ remains feasible. If we want to tell which distribution G is drawn from by looking only at whether $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ is feasible, we need that, when $G \sim \mathcal{G}(n, 1/2)$, $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ is infeasible with probability at least ε. The main result of [BHKKMP16] is that this is impossible. In particular, they show the following:

Theorem 1 ([BHKKMP16]) *For every $d \ll \log n$, when $k \leq n^{1/2 - O(\sqrt{d/\log n})}$ and G is drawn from $\mathcal{G}(n, 1/2)$, $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ is feasible with high probability.*

In other words, Barak et al.'s result says that the SoS approach to planted clique is no better (up to the $O(\sqrt{d/\log n})$ factor in the exponent) than the known algorithm from [AKS98].

From how $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ is defined, proving Theorem 1 boils down to find a linear operator $\tilde{\mathbb{E}}_G : \{q : \deg(q) \leq 2d\} \to \mathbb{R}$ for each graph G such that, if $G = ([n], E)$ is drawn from $\mathcal{G}(n, 1/2)$, the following conditions are satisfied with high probability:

1. $\tilde{\mathbb{E}}_G[1] = 1$,
2. $\forall i \in [n]\ \forall q$ with $\deg(q) \leq 2d - 2, \tilde{\mathbb{E}}_G[x_i^2 q] = \tilde{\mathbb{E}}_G[x_i q]$,
3. $\forall (i,j) \notin E\ \forall q$ with $\deg(q) \leq 2d - 2, \tilde{\mathbb{E}}_G[x_i x_j q] = 0$,
4. $\tilde{\mathbb{E}}_G[\sum_{i \in [n]} x_i] = k$,
5. $\forall q$ with $\deg(q) \leq d, \tilde{\mathbb{E}}_G[q^2] \geq 0$.

2 Pseudo-calibration for Planted Clique

Coming up with degree-$2d$ pseudo-distribution $\tilde{\mathbb{E}}_G$ with desired properties stated in the previous section is particularly hard for planted clique and past attempts often involve some ad-hoc fixes that prevent them from getting tight bound for large d. This is where Barak et al.'s so-called *pseudo-calibration* method, which is a systematic way to derive $\tilde{\mathbb{E}}_G$, comes in. Since the method is more of an intuitive heuristic rather than a provable approach, we will be informal here. We also note that the explanation given here is somewhat different than that in [BHKKMP16] and the readers should consult the full paper for a more thorough view of pseudo-calibration.

Let us take a step back and think about our algorithm for planted clique for a moment. Given G, we try to solve $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$. If it is infeasible, then we know for certain that G is drawn from $\mathcal{G}(n, 1/2)$. Otherwise, we do not seem to gain anything. However, this may not be entirely true; we actually get back $\tilde{\mathbb{E}}_G$. One thing we can do here is to pick f_G (which can depend on G) of degree (with respect to x) at most $2d$ as a test function and ask for $\tilde{\mathbb{E}}_G[f_G]$. If the distributions of $\tilde{\mathbb{E}}_G[f_G]$ under $G \sim \mathcal{G}(n, 1/2)$ and $G \sim \mathcal{G}(n, 1/2, k)$ are "very different"[2], then we should be able to tell G's from the two distributions apart by just looking at $\tilde{\mathbb{E}}_G[f_G]$. Hence, not only that $\mathrm{sos}_d(\mathrm{CLIQUE}_k(G))$ must be feasible with high probability when $G \sim \mathcal{G}(n, 1/2)$ but the distributions of $\tilde{\mathbb{E}}_G[f_G]$ when $G \sim \mathcal{G}(n, 1/2)$ and when $G \sim \mathcal{G}(n, 1/2, k)$ must also be indistinguishable in polynomial time for every test function f_G. An implication of this is that the expectation of $\tilde{\mathbb{E}}_G[f_G]$ over the two distributions are roughly equal, i.e.,

$$\mathop{\mathbb{E}}_{G \sim \mathcal{G}(n,1/2)} \tilde{\mathbb{E}}_G[f_G] \approx \mathop{\mathbb{E}}_{G \sim \mathcal{G}(n,1/2,k)} \tilde{\mathbb{E}}_G[f_G].$$

We of course do not know what $\tilde{\mathbb{E}}_G$ is even when G is drawn from $\mathcal{G}(n, 1/2, k)$ so the above equality does not tell us much yet. But recall that $\tilde{\mathbb{E}}_G$ is our fake solution and we want it to resemble the actual solution as much as possible. Hence, a reasonable heuristic here is to try to make $\mathbb{E}_{G \sim \mathcal{G}(n,1/2,k)} \tilde{\mathbb{E}}_G[f_G]$ roughly equal to $\mathbb{E}_{G \sim \mathcal{G}(n,1/2,k)} f_G(x_G)$ where x_G denote the actual solution, i.e., the indicator vector for the maximum clique in G.

For convenience, let us write $(G, x) \sim \mathcal{G}(n, 1/2, k)$ to denote G drawn from $\mathcal{G}(n, 1/2, k)$ and x being the indicator vector whether each vertex is included as part of the planted k-clique. Under this notation, the aforementioned condition can be written as

$$\mathop{\mathbb{E}}_{G \sim \mathcal{G}(n,1/2,k)} \tilde{\mathbb{E}}_G[f_G] \approx \mathop{\mathbb{E}}_{(G,x) \sim \mathcal{G}(n,1/2,k)} f_G(x).$$

Combining the above two equations, we get

$$\mathop{\mathbb{E}}_{G \sim \mathcal{G}(n,1/2)} \tilde{\mathbb{E}}_G[f_G] \approx \mathop{\mathbb{E}}_{(G,x) \sim \mathcal{G}(n,1/2,k)} f_G(x). \tag{2}$$

Condition (2) is what Barak et al. called *pseudo-calibration*[3]. As noted in the paper, this condition is quite strong. For example, for fixed $i, j \in [n]$ and q with $\deg(q) \leq 2d - 2$, if we define f as

$$f_G(x) = \begin{cases} 0 & \text{if } (i,j) \in E, \\ x_i x_j q(x) & \text{otherwise,} \end{cases}$$

then $f_G(x)$ is always zero on the right hand side. Hence, $\mathbb{E}_{G \sim \mathcal{G}(n,1/2)} \tilde{\mathbb{E}}_G[f_G] \approx 0$. If we assume that $\tilde{\mathbb{E}}_G[f_G]$ is non-negative, then Condition 3 at the end of the previous section is almost immediately satisfied. In fact, as we will see next, Condition (2) almost fully determines $\tilde{\mathbb{E}}_G$ for every G.

2.1 From Pseudo-Calibration to Pseudo-Distribution

We will now see how to arrive at $\tilde{\mathbb{E}}_G$ from the pseudo-calibration condition. As stated earlier, the condition is quite strong; in fact, it is too that it cannot hold for every f_G. For instance, we can pick f_G to simply be the

[2] In other words, they are distinguishable in polynomial time.

[3] In [BHKKMP16], the pseudo-calibration condition is in fact slightly stronger that stated here; equality is required instead of approximate equality. However, it does not matter anyway since there will be approximations in subsequent calculations.

indicator function of whether G has a clique of size k. By doing so, the left hand side of (2) is approximately zero whereas the right hand side is one. However, we are "cheating" by picking such f_G because we do not even know how to compute this test function in polynomial time! Hence, roughly speaking, we need to restrict f_G to only those that are not more "powerful" that the SoS relaxation itself.

To state the exact condition we enforce on f_G, let us think of $f_G(x)$ as a function $f(G, x)$ of both G and x where the graph G is encoded naturally as a string in $\{\pm 1\}^{\binom{[n]}{2}}$, i.e., the (i, j)-index of the input is $+1$ if there is an edge between i and j and -1 otherwise. Now, we can write f as a polynomial on both G and x:

$$f(G,x) = \sum_{T \subseteq \binom{[n]}{2}, S \subseteq [n]} a_{(T,S)} \chi_T(G) x_S$$

where $\chi_T(G)$ and x_S denote $\prod_{e \in T} G_e$ and $\prod_{i \in S} x_i$ respectively, and, $a_{(T,S)}$'s are the coefficients of the polynomial. We will require the pseudo-calibration condition to hold only for f_G such that each monomial depends on at most τ vertices where $\tau = O(d)$ is a truncation threshold. In other words, we only restrict ourselves to f that can be written as

$$f(G,x) = \sum_{\substack{T \subseteq \binom{[n]}{2}, S \subseteq [n] \\ |\mathcal{V}(T) \cup S| \leq \tau}} a_{(T,S)} \chi_T(G) x_S$$

where $\mathcal{V}(T)$ is the set of all vertices which are endpoints of edges in T. The intuition behind this heuristic is that, in the conditions on $\tilde{\mathbb{E}}_G$ imposed by the SoS relaxation, each monomial involves at most $2d$ vertices because $\tilde{\mathbb{E}}_G$ is defined only on polynomials on x of degree at most $2d$. As a result, each monomial appearing in $f(G, x)$ should involve no more than $O(d)$ vertices in order to limit its "power" to be not much more than the SoS relaxation.

Now, let us use the pseudo-calibration condition to determine $\tilde{\mathbb{E}}_G$. Fixed a subset $S \subseteq [n]$ of size at most $2d$, we will compute $\tilde{\mathbb{E}}_G[x_S]$ for the monomial x_S. Note that, since $\tilde{\mathbb{E}}_G$ is linear and $\tilde{\mathbb{E}}_G[x_i^2] = x_i$ for all $i \in [n]$, these $\tilde{\mathbb{E}}_G[x_S]$'s uniquely determine $\tilde{\mathbb{E}}_G$. By viewing $\tilde{\mathbb{E}}_G[x_S]$ as a function of G, $\tilde{\mathbb{E}}_G[x_S]$ can be written as fourier expansion

$$\tilde{\mathbb{E}}_G[x_S] = \sum_{T \subseteq \binom{[n]}{2}} \widehat{\tilde{\mathbb{E}}_G[x_S]}(T) \chi_T(G).$$

The final heuristic employed by Barak et al. is to enforce $\tilde{\mathbb{E}}_G[x_S]$ to be low degree by letting $\widehat{\tilde{\mathbb{E}}_G[x_S]}(T) = 0$ for every T with $|\mathcal{V}(T) \cup S| > \tau$. This heuristic makes sense since $\tilde{\mathbb{E}}_G$ must be output by the SoS relaxation solver, which runs in $n^{O(d)}$ time; hence, $\tilde{\mathbb{E}}_G$ cannot be too hard to compute. More importantly, as we will see shortly, this condition allows us to almost uniquely determine $\tilde{\mathbb{E}}_G$ from the pseudo-calibration condition.

Recall that each fourier coefficient $\widehat{\tilde{\mathbb{E}}_G[x_S]}(T)$ is simply equal to $\mathbb{E}_{G \sim \mathcal{G}(n,1/2)} \tilde{\mathbb{E}}_G[x_S \chi_T(G)]$. Plugging in the pseudo-calibration condition with $f = x_S \chi_T(G)$, this is approximately $\mathbb{E}_{(G,x) \sim \mathcal{G}(n,1/2,k)}[x_S \chi_T(G)]$. It is not hard to see that this expression is equal to the probability that every vertex in $\mathcal{V}(T) \cup S$ is in the planted clique, which is roughly $(k/n)^{|\mathcal{V}(T) \cup S|}$ when $|\mathcal{V}(T) \cup S|$ is small. Indeed, we will set $\widehat{\tilde{\mathbb{E}}_G[x_S]}(T)$ to be exactly this. In other words, the final pseudo-distribution is

$$\tilde{\mathbb{E}}_G[x_S] = \sum_{\substack{T \subseteq \binom{[n]}{2} \\ |\mathcal{V}(T) \cup S| \leq \tau}} \left(\frac{k}{n}\right)^{|\mathcal{V}(T) \cup S|} \chi_T(G).$$

It is not hard to see that $\tilde{\mathbb{E}}_G$ indeed satisfies the pseudo-calibration condition for f's of our interest. As explained right before the beginning of this subsection, this almost immediately implies that the third condition required for $\tilde{\mathbb{E}}_G$ is satisfied; it is also pretty easy to check that the condition is indeed true (see Lemma 5.5 in the paper). Using concentration inequalities, Barak et al. also show that $\tilde{\mathbb{E}}_G[1] = 1 \pm o(1)$ and $\tilde{\mathbb{E}}_G[\sum_{i \in [n]} x_i] = k \pm o(1)$ (see full proof in Appendix A.2 of the paper). Note that while these two conditions are only approximately satisfied, $\tilde{\mathbb{E}}_G$ can be scaled so that they are exactly satisfied as well. As mentioned briefly earlier, the proof of the positivity condition $\tilde{\mathbb{E}}_G[q^2] \geq 0$ is much harder and is the paper's main technical contribution. We do not attempt to discuss it here but we will try to blog about it in the future.

3 Further Reading

The authors of [BHKKMP16] have given talks on the paper and some of them are available online, such as Moitra's and Kothari's. Barak also wrote a blog regarding pseudo-calibration. All the materials mentioned discuss the pseudo-calibration in much more detail than in this post. Moitra's talk also contains the proof sketch of positivity of the pseudo-distribution, which is not covered in this blog post.

Apart from the paper, I am not aware of the pseudo-calibration technique being used to prove new lower bounds for other problems yet. I will update this section when I come across new results based on pseudo-calibration.

References

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Struct. Algorithms*, 13(3-4):457–466, 1998.

[BHKKMP16] Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *CoRR*, abs/1604.03084, 2016.