Limits of Approximation Algorithms 1 April, 2010 (TIFR)

Lec. 9: PCPs Constructions via low degree polynomials

Lecturer: Prahladh Harsha *Scribe: Srikanth Srinivasan*

In today's lecture [1], we will see how we can use the results on low-degree testing (which we discussed in detail in the earlier lectures) to construct PCPs. Towards this end, we will first show how local checkers for low-degreeness can be extended to check if a given function is a low-degree polynomial that vanishes on a pre-specified subcube. We will then "arithmetize" the NP-complete problem 3SAT in such a manner that the above checker yields a 2-query projective PCP.

9.1 Recap of the Low degree test

In the last lecture, we considered the low-degree *Plane-point* test for low degree polynomials over a finite field $\mathbb{F}$. We were given a function $f : \mathbb{F}^m \to \mathbb{F}$ that we called a "points oracle" and we wished to check if f was a polynomial of degree at most d by only querying f at a few points. To make this task easier, we were also given a "planes oracle" $\mathcal{A}$ that gives, for each plane in $\mathbb{F}^m$, a bivariate polynomial of degree at most d that is purportedly the restriction of f to the plane. The test that we performed was simple:

- Pick a plane s at random and a point x at random from it.
- Query the planes oracle $\mathcal{A}$ for the polynomial $\mathcal{A}(s)$ and the points oracle for the value $f(x)$.
- Accept iff $\mathcal{A}(s)(x) = f(x)$.

In the last lecture, we proved the following for the above plane-point test.

Theorem 9.1.1 (low degree test – standard formulation)**.** *Let $\mathbb{F}$ be a field of size q. Let $m, d \in \mathbb{Z}^{\geq 0}$ and $\delta \in (0,1)$ such that $\delta > \mathrm{poly}(d/q, m)$. Given any function $f : \mathbb{F}^m \to \mathbb{F}$, if there exists a planes oracle $\mathcal{A}$ satsisfying*

$$\Pr_{s,x}[\mathcal{A}(s)(x) = f(x)] \geq \gamma,$$

then there exists a degree d m-variate polynomial Q (i.e., $Q \in P_d^m$ [2]) such that

$$\Pr_{x}[f(x) = Q(x)] \geq \gamma - \delta.$$

Or equivalently,

$$\mathrm{agr}(f, P_d^m) \geq \mathop{\mathbb{E}}_{s}[\mathrm{agr}(f|_s, P_d^2)] - \delta.$$

[1] Prahladh: These notes are far more detailed than the lecture it corresponds to. Thanks to the scribe Srikanth for filling in all the missing details in the lecture.

[2] Recall that P_d^m refers to the set of all m-variate degree polynomials of degree at most d.

Above, the agreement $\mathrm{agr}(f,g)$ between $f,g:\mathbb{F}^m\to\mathbb{F}$ is defined to be the fraction of points in the domain where they agree and $\mathrm{agr}(f,S)$ refers to $\max_{g\in S}\mathrm{agr}(f,g)$.

We also showed that the above statement is equivalent to the following list-decoding formulation.

Theorem 9.1.2 (low degree test – list decoding formulation)**.** *Let $\mathbb{F}$ be a field of size q. Let $m,d\in\mathbb{Z}^{\geq 0}$ and $\delta\in(0,1)$ such that $\delta>\mathrm{poly}(d/q,m)$. For any function $f:\mathbb{F}^m\to\mathbb{F}$ there exists a list of polynomials $Q_1,Q_2,\ldots,Q_t$ of degree at most d where $t=O(\frac{1}{\delta})$ such that the following holds for any (even randomized) planes oracle $\mathcal{A}$,*

$$\Pr_{s,x}[\mathcal{A}(s)(x)\neq f(x)\ \vee\ \exists i\in[t],\ Q_i|_s\equiv\mathcal{A}(s)]\geq 1-\delta.$$

Finally, we state the version we will use. This is an extension of the above low-degree test from functions of the form $f:\mathbb{F}^m\to\mathbb{F}$ to $\overline{f}:\mathbb{F}^m\to F^k$ for some $k\in\mathbb{Z}^{\geq 0}$. That is, the points oracle is now a function $\overline{f}:\mathbb{F}^m\to\mathbb{F}^k$ and we need to verify that each coordinate of $\overline{f}$ is a polynomial of degree at most d. In this case, given a plane s, we expect the planes oracle $\overline{\mathcal{A}}$ to supply us with a k-tuple of polynomials of degree at most d. Finally, the planes oracle is also allowed to be *randomized*. The plane-point test remains exactly the same:

- Pick a plane s at random and a point x at random from it.
- Query the planes oracle $\overline{\mathcal{A}}$ for the k-tuple of polynomials $\overline{\mathcal{A}}(s)$ and the points oracle for the value $\overline{f}(x)$.
- Accept iff $\overline{\mathcal{A}}(s)(x)=\overline{f}(x)$.

It is easy to see the analysis for the case $k=1$ also works for arbitrary k. We thus have the following theorem which we refer to as the high dimensional version of the low-degree test.

Theorem 9.1.3 (high dimensional LDT)**.** *Let $\mathbb{F}$ be a field of size q. Let $m,d,k\in\mathbb{Z}^{\geq 0}$ and $\delta\in(0,1)$ such that $\delta>\mathrm{poly}(d/q,m)$. For any function $\overline{f}:\mathbb{F}^m\to\mathbb{F}^k$ there exists a list of k-tuples of polynomials $\overline{Q}_1,\overline{Q}_2,\ldots,\overline{Q}_t$ of degree at most d where $t=O(\frac{1}{\delta})$ such that the following holds for any (even randomized) planes oracle $\overline{\mathcal{A}}$,*

$$\Pr_{s,x}\left[\overline{\mathcal{A}}(s)(x)\neq\overline{f}(x)\ \vee\ \exists i\in[t],\ \overline{Q}_i|_s\equiv\overline{\mathcal{A}}(s)\right]\geq 1-\delta.$$

9.2 Label Cover and Robust PCPs

Now we come to the main focus of this lecture: constructing PCPs using Low degree tests. Recall two equivalent formulations of the PCP theorem. The first posits the existence of a PCP system with some nice properties, and the second states that a gap version of an optimization problem is NP-hard.

- **Formulation 1**: There exist $r=O(\log n)$, $\delta>0$, and $q=O(1)$ such that there is a polynomial-time verifier V that on input a CNF formula φ of size n and a proof π of size $m=\mathrm{poly}(n)$, does the following:

- Looking at the input and the outcomes of r many random bits, V produces a collection of indices $I = \{i_1, i_2, \ldots, i_q\} \subseteq [m]$ and a predicate $C : \{0,1\}^q \to \{0,1\}$.
- V then examines π at locations $i_1, i_2, \ldots, i_q$ and accepts iff $C(\pi_I) = 1$.

Furthermore, V satisfies the following:

- **Completeness**: If φ is satisfiable, then there is a proof π such that $\Pr_r[C(\pi_I) = 1] = 1$.
- **Soundness**: If φ is unsatisfiable, then for any proof π, it is the case that $\Pr_r[C(\pi_I) = 1] \leq 1 - \delta$, for some fixed constant $\delta > 0$.

- **Formulation 2**: We state this formulation in terms of the 2-query projective PCPs, which are more popularly called *Label Cover problem*. Recall the definition of this problem:

  Definition 9.2.1 (Label-Cover)**.** *An instance I of the* Label-Cover *problem is specified by a quadruple $(G, \Sigma_1, \Sigma_2, \Pi)$ where $G = (L, R, E)$ is a bipartite graph, Σ_1 and Σ_2 are two finite sized alphabets and $\Pi = \{\pi_e : \Sigma_1 \to \Sigma_2 | e \in E\}$, is a set of functions (also called projections), one for each edge $(u, v) \in E$.*

  A labeling $A : L \to \Sigma_1, B : R \to \Sigma_2$, is said to satisfy an edge (u, v) iff $\pi_{(u,v)}(A(u)) = B(v)$. The value *of an instance is the maximal fraction of edges satisfied by any such labeling.*

  For any $\delta \in (0, 1)$, the gap problem GAP_ε-LC is the promise problem of deciding if a given instance has value 1 or at most ε. More precisely, the YES *and* NO *of GAP_ε-LC are given as follows.*

$$\begin{aligned} \mathsf{YES} &= \left\{ I : \exists\, (A : L \to \Sigma_1,\ B : R \to \Sigma_2) \ \textit{such that}\ \forall (u,v) \in E,\ \pi_{(u,v)}(A(u)) = B(v) \right\} \\ \mathsf{NO} &= \left\{ I : \forall\, (A : L \to \Sigma_1,\ B : R \to \Sigma_2),\ \left| \{(u,v) \in E : \pi_{(u,v)}(A(u)) = B(v)\} \right| \leq \varepsilon |E| \right\} \end{aligned}$$

  Now, we can state the second formulation of the PCP theorem: SAT reduces in polynomial time to $\text{GAP}_{1-\delta}$-LC, for some fixed constant $\delta > 0$.

Let us see how these two formulations are equivalent (in lecture 5, we showed the first implies the second, but we will go over it again more carefully this time).

Assuming that SAT reduces to $\text{GAP}_{1-\delta}$-LC for some fixed $\delta > 0$, we can construct a verifier V as above who, on input a CNF formula φ, computes the instance G of $\text{GAP}_{1-\delta}$-LC that φ reduces to. The verifier expects, as a proof, a labeling to the right vertices R of G that can be extended to a labeling of all of G that satisfies all the edges of G. Given a labeling of R, the verifier picks a random left vertex $u \in L$, queries the proof for the labels of the neighbourhood of u, and accepts if u can be labelled in any way so that all these edges can be satisfied. Note that the number of random coins used by V is $O(\log |\varphi|)$ and the number of queries is q, the constant left-degree of G. Moreover, if there is a labeling of the graph G that satisfies all the edges, then the restriction of this labeling to R makes V accept with probability 1. On the other hand, if the verifier accepts with probability p,

then there is a labelling of G that satisfies at least a p fraction of the edges; hence, if G is a negative instance of $\text{GAP}_{1-\delta}$-LC, then V accepts with probability at most $1-\delta$.

Now for the reduction in the other direction: if there is a verifier V with the above mentioned properties and soundness error at most $1-\delta$, then we construct a GAP-LC instance G as follows: for each setting R of the random coins of the verifier, we add a left vertex u_R to L, and for each location i of the proof, we add a right vertex v_i to R. There is an edge between u_R and v_i iff the ith location of the proof is queried when the random coins take value R. Assume that when the random coins take value R the verifier queries locations $i_1, i_2, \ldots, i_q$; the labeling for u_R is an assignment to these locations in the proof that satisfies the predicate C_R of the verifier on these random bits. The assignment to v_i is expected to be the ith bit of the proof. The label cover instance expects the assignments to the left vertices to be "induced" from the proof on the right: that is, for each edge (u_R, v_i), we have a constraint that forces a fixed label for v_i depending on the label of u_R. The completeness condition is easy to see. For the soundness, consider the proof given by the labeling of the vertices in R. Since the verifier accepts this proof with probability at most $1-\delta$, we see that for at least a δ fraction of the vertices in L must be inconsistent with their neighbours on the right. Hence, for each such vertex u, there is at least one edge e incident to it that is not satisfied. Since the degree of each left vertex is q, this implies that any labeling of the vertices of G satisfies at most a $1-\delta/q$ fraction of edges. Thus, G is an instance of $\text{GAP}_{1-\delta/q}$-LC.

The above arguments show that, in the regime where δ is close to 0 (and for small q), the two formulations are just about equivalent. However, when δ is close to 1, that is, when the soundness error is close to 0, then the equivalence is lost since a verifier with soundness error $1-\delta$ (which is close to 0) translates to an instance of $\text{GAP}_{1-\delta/q}$-LC, which is at least $1-1/q$. Since we will be interested in the low soundness error regime, this suggests that we need a slightly different notion of a PCP system that is equivalent to GAP_{δ}-LC for small δ also.

For this reason, we introduce the notion of a *robust PCP*, a prover-verifier system where the verifier behaves exactly as above except for a more stringent soundess condition.

- **(Robust) Soundness**: $\mathbb{E}_r[\text{agr}(\pi_I, SAT(C))] \leq \delta$, where $SAT(C)$ is the set of satisfying assignments to the predicate C. We will refer to δ as the (robust) soundness error of the PCP.

Note that the above is a stronger requirement that the above (robust) soundness condition, where the predicate C is satisfied with probability at most δ.

The point of defining a robust PCP is the following.

Lemma 9.2.2. *The following statements are equivalent, for any $\delta > 0$:*

- *There is a robust PCP for SAT with (robust) soundness error at most δ.*
- *SAT polynomial-time reduces to* GAP_{δ}-LC.

The above statement can be proved by following exactly the reductions stated above. The above equivalence theorem is due to Dinur and Harsha [DH09]. Our aim, in this lecture, will be to prove the existence of robust PCPs with soundness error δ (or equivalently, the NP-hardness of GAP_{δ}-LC) for an arbitrary small constant $\delta > 0$.

9.3 Robust PCPs from Low Degree Tests

We now start proving the existence of Robust PCPs for SAT. The basic idea is to encode a satisfying assignment x of a SAT formula φ in a manner such that it can be locally checked by a polynomial-time verifier. Recall that the low degree test shows that the property of being a low-degree polynomial over a sufficiently large field can be locally checked. We will use this crucially in designing the locally checkable encodings of satisfying assignments.

To see how low-degree polynomials along with the associated low-degree test might yield a PCP, let us see that there is a robust PCP for checking if a given function $f : \mathbb{F}^m \to \mathbb{F}$ is (close to) a polynomial of small degree d. This falls right out of the plane-point test described above. The verifier picks a plane at random, reads off the function f restricted to this plane and accepts iff this is a low-degree polynomial. Let us see that this yields a robust PCP. Completeness is trivial. To prove robust soundness, consider an associated planes oracle $\mathcal{A}$ such that for any plane s, $\mathcal{A}(s)$ is the polynomial of degree at most d that has the maximum agreement with the function obtained when f is restricted to s (the planes oracle represents the closest "satisfying assignments" for each question of the verifier). It is easy to see, from the analysis of the plane-point test mentioned above, that if the polynomial f is "far" (say $(1-1/q-\varepsilon)$-far) from *every* polynomial of degree at most d, then the robust soundness error of the verifier – which is exactly the quantity $\Pr_{s,x}[\mathcal{A}(s)(x) = f(x)]$ – is bounded by δ for δ close to 0. What we now plan to do in the rest of the lecture is to replace "low-degree polynomials" by "satisfying assignments of a SAT formula φ" in this entire paragraph, and we would then have a robust PCP for SAT.

We now proceed to the construction of the PCP. We start by introducing the *Zero on subcube test*, which is an extension of the property of being a low-degree polynomial that we will use later.

9.3.1 Zero on subcube test

Let $H \subseteq \mathbb{F}$. Given $d \in \mathbb{N}$, we say that a function $f : \mathbb{F}^m \to \mathbb{F}$ satisfies the *Zero on subcube* property if: (a) f is a polynomial of degree at most d, and (b) f is identically 0 on H^m.

We would like to design a local checker for the above property, similar to the plane-point test for degree d polynomials. That is, we would like our test to have the following property: accept any function that satisfies the above property with probability 1, and not accept any function with non-trivial probability unless it has some noticeable correlation with a function that satisfies this property.

A natural test that comes to mind is the following: Test if f is a low-degree polynomial using the plane-point test, and then check if f is 0 at a random point of H^m. But this test is flawed since it only tests if f is close to some low-degree polynomial f' and to a function f'' — not necessarily f' — that vanishes over all of H^m. However, it fails to check if f has correlation with some function that has *both* these properties. Hence, a different approach is necessary.

In what follows, we will need the following lemma, which gives an algebraic characterization of polynomials that vanish over H^m. Let $g_H(x)$ denote the univariate polynomial $\prod_{h \in H}(x - h)$.

Lemma 9.3.1. *A polynomial f of degree at most d is identically 0 over H^m iff there exist polynomials $P_1, P_2, \ldots, P_m$ of degree at most d such that $f(x) = \sum_{i=1}^{d} g_H(x_i) P_i(x_1, x_2, \ldots, x_m)$.*

Proof. It is easy to see that if there exist polynomials $P_1, P_2, \ldots, P_m$ with the above properties, then f does indeed vanish over all of H^m. To prove the converse, we proceed as follows.

We define sequences of polynomials $P_1, \ldots, P_m, R_0, R_1, R_2, \ldots, R_m \in \mathbb{F}[x_1, x_2, \ldots, x_m]$ as follows: R_0 is simply the polynomial f and for $i \geq 1$, we divide R_{i-1} by $g_H(x_i)$ and set P_i to be the quotient and R_i to be the remainder; more formally, we write $R_{i-1} = g_H(x_i) P_i + R_i$ in the unique way so that the degree of x_i in R_i is less than $|H|$. Note that the degree of P_i is at most d.

By the definition of the polynomials above, we have $f = \sum_i g_H(x_i) P_i + R_m$, where R_m is a polynomial of degree at most $|H| - 1$ in each x_i. Note that since f and $g_H(x_i)$ (for each $i \in [m]$) vanish over H^m, so does R_m. Hence, R_m must in fact be the zero polynomial. This shows that $f = \sum_i g_H(x_i) P_i$, which proves the lemma. □

Now we design the Zero on subcube test as follows. The test is identical to the high-dimensional version of the plane-point test. To prove that $f : \mathbb{F}^m \to \mathbb{F}$ is a polynomial of degree at most d that vanishes over H^m, we expect the prover to provide us with the polynomials $P_1, P_2, \ldots, P_m$ mentioned in Lemma 9.3.1 and prove that $f, P_1, \ldots, P_m$ are low-degree polynomials.

More formally, as in the low-degree test, we expect a planes oracle $\overline{\mathcal{A}}$ such that for each plane s, $\overline{\mathcal{A}}(s)$ is an $(m+1)$-tuple $(p_0, p_1, p_2, \ldots, p_m)$ of polynomials of degree at most d such that $p_0 = \sum_i g_H(x_i) p_i$. We also expect a points oracle that is an $(m+1)$-tuple of functions $\overline{f} = (f_0 = f, f_1, \ldots, f_m)$ from $\mathbb{F}^m$ to $\mathbb{F}$. The test itself is exactly the same as the plane-point test.

The Zero on subcube test:

- Pick a plane s uniformly at random and a random point x from it.
- Query the planes oracle for $\overline{\mathcal{A}}(s)$ and the points oracle for $\overline{f}(x)$.
- Accept iff $\overline{\mathcal{A}}(s)(x) = \overline{f}(x)$.

Since the above test is just the plane-point test applied to a planes oracle and points oracle, we know that for any $\delta > \text{poly}(m, d/q)$, there exists a short list of $(m+1)$-tuples of polynomials $\overline{Q^{(1)}}, \overline{Q^{(2)}}, \ldots, \overline{Q^{(t)}}$ — with $\overline{Q^{(i)}} = (Q_0^{(i)}, \ldots, Q_m^{(i)})$ and $t = O(1/\delta)$ — of degree at most d such that

$$\Pr_{s,x}\left[\overline{\mathcal{A}}(s)(x) \neq \overline{f}(x) \ \vee \ \exists i \in [t] \ \overline{Q^{(i)}}|_s \equiv \overline{\mathcal{A}}(s)\right] \geq 1 - \delta \tag{9.3.1}$$

However, this is not sufficient for our purposes, since the above does not state that the polynomials $Q_0^{(i)}$ are identically zero on the subcube H^m. Let us call $\overline{Q^{(i)}}$ for $i \in [t]$ *good* if $\overline{Q_0^{(i)}}$ is indeed identically zero in H^m and *bad* otherwise. In particular, for any bad $\overline{Q^{(i)}}$, we must have

$$\overline{Q_0^{(i)}} \neq \sum_{j=1}^{m} g_H(x_j) \overline{Q_j^{(i)}} \tag{9.3.2}$$

In fact, we will extend the definition of "good" as follows: the $(m+1)$-tuple $\overline{Q^{(i)}}$ is good if it satisfies (9.3.2) and bad otherwise.

Consider the new list of tupless obtained by throwing away all but the good tuples. We would like to say that Inequality 9.3.1, or some approximation of it, continues to hold.

Note that if this were *not* to be the case, then it must be true that the bad $\overline{Q^{(i)}}$ must agree on many planes with the planes oracle $\overline{\mathcal{A}}(s)$. However, for any s, if $\overline{\mathcal{A}}(s) = (p_0, p_1, \ldots, p_m)$, then we know that $p_0 = \sum_i g_H(x_i)p_i$. This implies that for many planes s, the bad $\overline{Q^{(i)}}$ satisfy the analogue of Equation 9.3.2 on the plane s. But a simple application of the Schwartz-Zippel lemma shows that this is not possible.

Let us state the above formally. Fix any bad $\overline{Q^{(i)}}$. For a plane s, let $(q_{s,0}, q_{s,1}, \ldots, q_{s,m})$ be the restriction of $\overline{Q^{(i)}}$ to the plane s. An application of the Schwartz-Zippel lemma shows that for a random choice of plane s,

$$\Pr_s \left[q_{s,0} = \sum_j g_H(x_j) q_{s,j} \right] \leq \frac{d + |H|}{q}$$

However, if the above event does not occur, then the restriction of $\overline{Q^{(i)}}$ to the plane s cannot agree with $\overline{\mathcal{A}}(s)$.

Let ε denote $(d + |H|)/q$. The above shows that:

$$\Pr_{s,x} \left[\overline{\mathcal{A}}(s)(x) \neq \overline{f}(x) \ \vee \ \exists i \text{ s.t } \overline{Q^{(i)}} \text{ good and } \overline{Q^{(i)}}|_s \equiv \overline{\mathcal{A}}(s) \right] \geq 1 - \delta - t\varepsilon \tag{9.3.3}$$

This concludes the proof of correctness of the Zero on subcube test.

9.3.2 Arithmetization

We now turn to the actual construction of the PCP: how one can encode a satisfying assignment of a SAT formula φ using a low-degree polynomial so that it is locally checkable. This process is known as *arithmetization*. Roughly, we will proceed as follows. We will first show how to encode any assignment to the variables of a formula as a low degree polynomial f so that f evaluates to zero on a predetermined subcube H^m iff f was obtained from a satisfying assignment of φ. We can then use the Zero on subcube test above to show that this can be locally checked. However, this entire procedure only works for polynomials constructed from satisfying assignments in the way described above. To ensure that the prover hasn't cheated and given us other polynomials, we will demand more structure of the planes oracle – ultimately, this will lead us to modify the low degree test in a fundamental way. The arithmetization presented in this section is from lecture 18 of Sudan's course on inapproximability at MIT [Sud99].

Fix a 3-CNF formula φ over n variables. Let $\mathbb{F}$ be a field of size q (ultimately, q will be $\operatorname{poly}\log n$). Fix any subset H of $\mathbb{F}$ such that H contains $\{0, 1\}$ and there is an integer m such that $|H|^m = n$; we will identify $[n]$ with H^m. Consider a Boolean assignment $A : [n] \to \{0, 1\}$ to the variables of φ. We will think of A as a function mapping H^m to $\mathbb{F}$. Using standard interpolation techniques, it is easy to prove that there is a polynomial $\tilde{A}$ of degree $O(m|H|)$ that agrees with A when evaluated on inputs from H^m. This defines the polynomial representation of the assignment A that we will work with.

Similarly, we will also need a polynomial representation of the formula φ. For each possible clause of 3 variables, the polynomial encodes whether the clause belongs to φ or not. We think of the formula φ as a function mapping $[n]^3 \times \{0,1\}^3$ to $\{0,1\}$ as follows:

$$\varphi(i,j,k,b_1,b_2,b_3) = \begin{cases} 1 & \text{if } x_i^{b_1} \vee x_j^{b_2} \vee x_k^{b_3} \text{ is a clause in } \varphi. \\ 0 & \text{otherwise.} \end{cases}$$

where x_i^0 and x_i^1 represent the negative and positive instances of x_i respectively. Since we have identified H^m with $[n]$ and H contains $\{0,1\}$, we can think of φ as a function from H^{3m+3} to $\mathbb{F}$ (define φ to be 0 outside the points mentioned above). As in the case of the assignment, we can define a polynomial $\tilde{\varphi}$ over $3m+3$ variables of degree $O(m|H|)$ that agrees with φ on H^{3m}.

Given the polynomials $\tilde{\varphi}$ and $\tilde{A}$ defined above, we are ready to define the polynomial on which a zero on subcube test will tell us if A is a satisfying assignment or not. This polynomial, defined on $3m+3$ variables, is denoted $p_{\varphi,A}$ and is defined below. We think of the input to $p_{\varphi,A}$ as three tuples i,j,k from $\mathbb{F}^m$ followed by three field elements b_1,b_2,b_3.

$$p_{\varphi,A}(i,j,k,b_1,b_2,b_3) = \tilde{\varphi}(i,j,k,b_1,b_2,b_3)(\tilde{A}(i)-b_1)(\tilde{A}(j)-b_2)(\tilde{A}(k)-b_3)$$

Clearly, $p_{\varphi,A}$ is a polynomial of degree $O(m|H|)$. We claim that moreover, $p_{\varphi,A}$ vanishes over the subcube H^{3m+3} iff A is a satisfying assignment for the formula φ. To see this, assume that $p_{\varphi,A}$ is evaluated on input $(i,j,k,b_1,b_2,b_3) \in H^{3m+3}$. Unless b_1,b_2,b_3 lie in $\{0,1\}$ and the clause $x_i^{b_1} \vee x_j^{b_2} \vee x_k^{b_3}$ is a clause in φ, the polynomial $\tilde{\varphi}(i,j,k,b_1,b_2,b_3)$ evaluates to 0 and hence so does $p_{\varphi,A}$. On the other hand, when the clause $x_i^{b_1} \vee x_j^{b_2} \vee x_k^{b_3}$ is in φ, then it is easy to see that $p_{\varphi,A}$ evaluates to 0 iff $\tilde{A}(i) = b_1$ or $\tilde{A}(j) = b_2$ or $\tilde{A}(k) = b_3$, which happens exactly when A satisfies this clause of φ. We have proved the following.

Lemma 9.3.2. *Let $\tilde{A}$ be any polynomial defined on m variables. Assume the polynomial $p_{\varphi,A}$ is constructed from $\tilde{A}$ as above. Then, $p_{\varphi,A}$ is identically zero on H^{3m+3} iff $\tilde{A}\,|_{H^m}$ is a satisfying assignment for the formula φ.*

Hence, it seems that it is enough for the prover to supply us with enough proof for to be able to verify that the polynomial $p_{\varphi,A}$, for some satisfying assignment A, vanishes over H^{3m+3}. Recall that to do this, the prover needs to supply us with $3m+3$ polynomials $P_1, P_2, \ldots, P_{3m+3}$ of degree $O(m|H|)$ such that $p_{\varphi,A} = \sum_i g_H(x_i) P_i$ and prove to us that they are low degree.

The entire proof is now the following:

- The points oracle: A collection of functions $\overline{f} : \mathbb{F}^{3m+3} \to \mathbb{F}^{3m+4}$. In the ideal proof, the prover would ensure that $\overline{f} = (f_0, \ldots, f_{3m+3})$ where $f_0 = p_{\varphi,A}, f_1 = P_1, \ldots, f_{3m+3} = P_{3m+3}$.
- The planes oracle: For each plane s, a $(3m+4)$-tuple of polynomials $(p_0, p_1, \ldots, p_{3m+3})$ of degree $O(m|H|)$ such that $p_0 = \sum_{1 \le i \le 3m+3} g_H(x_i) p_i$. Ideally, these would just be the restrictions of the functions in the points oracle to this plane.

However, this is not enough: the prover must also prove that the polynomial f_0 he supplies in place of $p_{\varphi,A}$ was indeed constructed from some assignment A in the manner

described above. More precisely, he needs to show that there is a polynomial $\tilde{A}$ such that f_0 is constructed from $\tilde{A}$ as above.

To do this, he also supplies the polynomial $\tilde{A}$ along with the points oracle and proves, using an additional coordinate in the planes oracle, that this is indeed a low-degree polynomial. Note that there is a slight type-mismatch here: $\tilde{A}$ is a polynomial over m variables, whereas all the other polynomials we have been working with are polynomials over $3m+3$ variables. To get around this, we work with an extended version $\overline{A}$ that applies $\tilde{A}$ to the first m variables among the $3m+3$ variables in its input. That is,

$$\overline{A}(x_1, x_2, \ldots, x_{3m+3}) = \tilde{A}(x_1, x_2, \ldots, x_m)$$

(There is a slight problem here: how do we know that the polynomial $\overline{A}$ the prover has supplied depends only on the first m variables? We will need to get around this in what follows, but we will ignore the problem for now.)

We need to check that the following identity holds:

$$\begin{aligned} f_0(i,j,k,b_1,b_2,b_3) &= \tilde{\varphi}(i,j,k,b_1,b_2,b_3)(\tilde{A}(i)-b_1)(\tilde{A}(j)-b_2)(\tilde{A}(k)-b_3) \\ &= \tilde{\varphi}(i,j,k,b_1,b_2,b_3)(\overline{A}(x_1)-b_1)(\overline{A}(x_2)-b_2)(\overline{A}(x_3)-b_3) \end{aligned}$$

where x_1 is some point in $\mathbb{F}^{3m+3}$ that contains i in its first m coordinates, and similarly x_2 and x_3 contain j and k respectively in their first m coordinates.

We will proceed as in the case of the zero on subcube test, where instead of actually checking that an identity held, we forced the planes oracle to only supply us with tuples of polynomials that satisfied the identity. We would like the planes oracle to always satisfy the above identity on each plane, but the problem with ensuring this is that most planes do not contain three points of the form on which $\overline{A}$ is evaluated. Hence, we have no way of restricting the planes oracle to only such polynomials, as long as we insist on continuing to use planes.

The answer, as we will see in the next few subsections, is to use a slightly larger object instead of planes in the low-degree test on which the above identity can indeed be verified. More precisely, define $\rho : \mathbb{F}^{3m+3} \to \mathbb{F}^{3m+3}$ to be the linear map such that $\rho(i,j,k,b_1,b_2,b_3) = (k,i,j,b_1,b_2,b_3)$. This larger object will contain, for some $x \in \mathbb{F}^{3m+3}$, the points $x, \rho(x)$, and $\rho^2(x)$. We will then check that:

$$f_0(x) = \tilde{\varphi}(x)(\overline{A}(x) - x_{3m+1})(\overline{A}(\rho(x)) - x_{3m+2})(\overline{A}(\rho^2(x)) - x_{3m+3}) \tag{9.3.4}$$

The larger object will also help us verify that the polynomial $\overline{A}$ does indeed depend on the first m variables only. In the next subsection, we will show that this version of the low-degree test can also be proved to work in the same sense as the plane-point test, for a rather general notion of "object": the proof proceeds by simply reducing the analysis of this test to that of the plane-point test. After that, we will formally define the objects we will work with, and show that they are sufficient to construct the PCP.

It is to be noted that the way we have circumvented the above problems in these notes is not the only one: a clever folding argument can also be used to circumvent the problem of checking the identity mentioned in Equation 9.3.4.

9.3.3 The Object-point test

As mentioned above, we now need to modify the low-degree test to work over more complicated objects than just planes. Without going into details right now about what exactly our object will be, we prove that the *object-point* low-degree test will work just as the plane-point test does, under a rather general definition of "object".

Each of our objects will be associated with some constant-dimensional (dimension less than or equal to about 7 will do) subspace from $\mathbb{F}^m$. For now, let us think of the object as just the subspace associated with it.

Let $\mathcal{D}_1$ be the distribution over tuples (Ω, s) of objects and planes induced by picking an object at random and then picking a random plane s contained in it. Let $\mathcal{D}_2$ be the distribution over pairs (Ω, s) induced by picking a plane s at random and then picking a random Ω containing s. For $\eta > 0$, we say that our collection of objects is η-good if the statistical distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is at most η.

Say our collection of objects is η-good for some $\eta > 0$. Fix any "objects oracle" $\overline{\mathcal{A}}$ (possibly randomized) that, when given an object Ω, outputs k polynomials of degree d. Also fix a points oracle $\overline{f} : \mathbb{F}^m \to \mathbb{F}^k$. The object-point test is defined as follows:

- Pick a random object Ω and a random $x \in \Omega$.
- Query the objects oracle for $\overline{\mathcal{A}}(\Omega)$ and the points oracle for $\overline{f}(x)$.
- Accept iff $\overline{\mathcal{A}}(\Omega)(x) = \overline{f}(x)$.

We wish to prove that an analogue of Theorem 9.1.3 holds for the above test. We can prove this by simply reducing to the plane-point test. Consider a randomized planes $\tilde{\mathcal{A}}$ oracle defined as follows: Given a plane s, $\tilde{\mathcal{A}}$ queries $\overline{\mathcal{A}}(\Omega)$ where Ω is a random object containing the plane s and outputs the restriction of $\overline{\mathcal{A}}(\Omega)$ to the plane s.

Fix $\delta > \text{poly}(m, d/q)$. By Theorem 9.1.3, there is a short list of $t = O(1/\delta)$ degree-d polynomial maps $\overline{Q_1}, \overline{Q_2}, \ldots, \overline{Q_t}$ such that

$$\Pr_{\tilde{\mathcal{A}},s,x}\left[\tilde{\mathcal{A}}(s)(x) = \overline{f}(x) \vee \ \exists i\ \overline{Q_i}|_s \equiv \tilde{\mathcal{A}}(s)\right] \geq 1 - \delta$$

For $i \in \{1, 2\}$, we use $(\Omega, s)_i$ to denote a pair (Ω, s) picked according to distribution $\mathcal{D}_i$. The above implies that

$$\Pr_{\overline{\mathcal{A}},(\Omega,s)_2,x}\left[\overline{\mathcal{A}}(\Omega)(x) = \overline{f}(x) \vee \ \exists i\ \overline{Q_i}|_s \equiv \overline{\mathcal{A}}(s)\right] \geq 1 - \delta$$

Since our objects are η-good, we have:

$$\Pr_{\overline{\mathcal{A}},(\Omega,s)_1,x}\left[\overline{\mathcal{A}}(\Omega)(x) = \overline{f}(x) \vee \ \exists i\ \overline{Q_i}|_s \equiv \overline{\mathcal{A}}(s)\right] \geq 1 - \delta - \eta$$

We almost have the analogue of Theorem 9.1.3 in the case of the object-point test. The only difference is that above we get the agreement of $\overline{Q_i}$ with $\overline{\mathcal{A}}(\Omega)$ on a random plane s chosen from Ω instead of over the entire object. However, by a standard Schwartz-Zippel

argument, this implies that most of the time, we must get agreement over the entire object. Formally, for any i and Ω and a random s picked from Ω, we have:

$$\Pr_{\overline{\mathcal{A}},s}\left[\overline{Q}_i|_\Omega \not\equiv \overline{\mathcal{A}}(\Omega) \ \wedge \ \overline{Q_i}|_s \equiv \overline{\mathcal{A}}(\Omega)|_s\right] \leq \frac{d}{q} \leq \varepsilon$$

Hence, by a union bound over i, we have:

$$\Pr_{\overline{\mathcal{A}},(\Omega,s)_1,x}\left[\overline{\mathcal{A}}(\Omega)(x) = \overline{f}(x) \vee \ \exists i \ \overline{Q}_i|_\Omega \equiv \overline{\mathcal{A}}(\Omega)\right] \geq 1 - \delta - \eta - t\varepsilon$$

We have proved the following, for any collection of η-good objects. (We have absorbed the $t\varepsilon$ term in the δ.)

Theorem 9.3.3. *Fix any points oracle $\overline{f} : \mathbb{F}^m \to \mathbb{F}^k$. Given any $\delta > \mathrm{poly}(m, d/q)$ and any objects oracle $\overline{\mathcal{A}}$, there exists a list of k-tuples of polynomials $\overline{Q}_1, \overline{Q}_2, \ldots, \overline{Q}_{O(\frac{1}{\delta})}$ of degree at most d such that*

$$\Pr_{\overline{\mathcal{A}},\Omega,x}\left[\overline{\mathcal{A}}(\Omega)(x) \neq \overline{f}(x) \ \vee \ \exists i \in [t] \ s.t. \ \overline{Q}_i|_\Omega \equiv \overline{\mathcal{A}}(\Omega)\right] \geq 1 - \delta - \eta$$

We leave it to the reader to check that the zero on subcube test works in exactly the same way for η-good objects as it does for planes (in fact, all we need is that a random point in an object looks like a random point from the entire space).

9.3.4 Defining our objects and completing the construction of the PCP

Our objects will essentially be constant-dimensional subspaces in $\mathbb{F}^m$. They will contain points of the form $z, \rho(z), \rho^2(z)$ so that the identity in Equation 9.3.4 can be checked. Moreover, to check that the polynomial $\overline{A}$ depends only on the first m variables, we will ensure that the object contains points z, z' that agree on the first m coordinates.

To pick an object at random, we pick $y, y', z \in \mathbb{F}^m$ independently and uniformly at random. Moreover, we also pick $z' \in \mathbb{F}^m$ such that the first m coordinates of z' are the same as the first m coordinates of z and the remaining coordinates of z' are chosen independently and uniformly at random from $\mathbb{F}$. The object Ω that we have picked is formally just the tuple (y, y', z, z'); we associate with Ω the subspace $L(\Omega)$ spanned by the vectors $z, z', \rho(z), \rho^2(z), y$, and y'. Note that the same subspace may be associated with different tuples and hence with different objects.

To prove that the object point test has the nice properties proved in the previous subsection, we need to show that our objects are well-behaved: that is, they are η-good for some small $\eta > 0$. We state this as a claim here, and postpone the simple but rather ugly details to the appendix.

Claim 9.3.4. *The objects Ω as defined above are $O(1/q)$-good.*

The PCP we construct is based on the zero-on-subcube test for objects. As in the case of planes, the proof consists of the following:

- The points oracle: A collection of functions $\overline{f} : \mathbb{F}^{3m+3} \to \mathbb{F}^{3m+5}$. In the ideal proof, the prover would ensure that $\overline{f} = (f_{-1}, f_0, \ldots, f_{3m+3})$ where $f_{-1} = \overline{A}, f_0 = p_{\varphi,A}, f_1 = P_1, \ldots, f_{3m+3} = P_{3m+3}$.

- The objects oracle: For each object $\Omega = (y, y', z, z')$, a $(3m+5)$-tuple of polynomials $(p_{-1}, p_0, p_1, \ldots, p_{3m+3})$ of degree $O(m|H|)$ defined on (the subspace given by) Ω such that
  - $p_0(x) = \sum_{1 \le i \le 3m+3} g_H(x_i) p_i(x)$ for each $x \in L(\Omega)$.
  - $p_{-1}(z) = p_{-1}(z')$. This helps us check that $\overline{A}$ depends only on the first m variables.
  - $p_0(z) = \tilde{\varphi}(z)(p_{-1}(z) - z_{3m+1})(p_{-1}(\rho(z)) - z_{3m+2})(p_{-1}(\rho^2(z)) - z_{3m+3})$. This helps us verify Equation 9.3.4.

  Ideally, these would just be the restrictions of the functions in the points oracle to this object.

Let $d = O(m|H|)$ be an upper bound on the degree of the above polynomials. Fix $\delta > \text{poly}(m, d/q)$. By results stated in the previous subsection about the zero on subcube test with objects, we know that there exists a short list of $(3m+5)$-tuples of polynomials $\overline{Q^{(1)}}, \overline{Q^{(2)}}, \ldots, \overline{Q^{(t)}}$ – here, $t = O(\frac{1}{\delta})$ – such that each $Q_0^{(i)}$ is zero on the subcube H^m and

$$\Pr_{\Omega, x}\left[\overline{\mathcal{A}}(\Omega)(x) \neq \overline{f}(x) \ \vee \ \exists i \ \overline{Q^{(i)}}|_\Omega \equiv \overline{\mathcal{A}}(\Omega)\right] \ge 1 - \delta - O(1/q) - t\varepsilon$$

where ε denotes $(d + |H|)/q$.

Now, we would like to say that we can prune the above list of polynomial maps so that we are only left with those tuples $\overline{Q^{(i)}}$ such that $Q_0^{(i)}$ is $p_{\varphi,A}$ for some satisfying assignment A of the formula φ, and yet the above condition (or a slight weakening) holds for this smaller list of polynomials.

How do we verify that $\overline{Q^{(i)}}$ is $p_{\varphi,A}$ for some satisfying assignment A? We need to verify that identity given in Equation 9.3.4 holds with $Q_0^{(i)}$ in place of $p_{\varphi,A}$ and $Q_{-1}^{(i)}$ in place of $\overline{A}$. Moreover, we also need to verify that $Q_{-1}^{(i)}$ is indeed a polynomial only in the first m variables (as $\overline{A}$ is supposed to be). Call $\overline{Q^{(i)}}$ ($i \in [t]$) *bad* if either of these conditions does not hold and *good* otherwise. We will now prove that pruning the above list by throwing away all the bad $\overline{Q^{(i)}}$ does not significantly change the above statement.

Fix any bad $\overline{Q^{(i)}}$. $\overline{Q^{(i)}}$ can be bad for two reasons:

- The first is that the following happens.

  $$Q_0^{(i)}(x) \not\equiv \tilde{\varphi}(x)(Q_{-1}^{(i)}(x) - x_{3m+1})(Q_{-1}^{(i)}(\rho(x)) - x_{3m+2})(Q_{-1}^{(i)}(\rho^2(x)) - x_{3m+3})$$

  However, note that this implies that with probability at least $1 - d/q$ over the choice of a random object Ω, the above inequality continues to hold when restricted to Ω. (This is because there is a random z such that $z, \rho(z)$, and $\rho^2(z)$ lie in Ω.) However, since our objects oracle $\overline{\mathcal{A}}$ always satisfies the above with equality, we see that the probability, for a random Ω, that $\overline{Q^{(i)}}|_\Omega \equiv \overline{\mathcal{A}}(\Omega)$ is at most $d/q \le \varepsilon$.

- The second reason why $\overline{Q^{(i)}}$ could be bad is that the polynomial $Q_{-1}^{(i)}$ is not a polynomial in just the first m variables. This implies that $Q_{-1}^{(i)}(z_1, z_2) \not\equiv Q_{-1}^{(i)}(z_1, z_3)$, where

z_1 represents a tuple of m variables and z_2 and z_3 represent disjoint tuples of $2m+3$ variables. Then, as above, this inequality continues to hold with probability at least $1 - d/q$ for a random Ω, since Ω contains a random z, z' that agree on the first m coordinates. Hence, the probability that $\overline{Q^{(i)}}|_\Omega \equiv \overline{\mathcal{A}}(\Omega)$ is at most $d/q \leq \varepsilon$.

Thus, we have shown that for any bad $\overline{Q^{(i)}}$,

$$\Pr_{\Omega}\left[\overline{Q^{(i)}}|_\Omega \equiv \overline{\mathcal{A}}(\Omega)\right] \leq \varepsilon$$

Hence, a simple union bound gives us

$$\Pr_{\Omega,x}\left[\overline{\mathcal{A}}(\Omega)(x) \neq \overline{f}(x) \ \vee \ \exists i \text{ s.t. } \overline{Q^{(i)}} \text{ good and } \overline{Q^{(i)}}|_\Omega \equiv \overline{\mathcal{A}}(\Omega)\right] \geq 1 - \delta - O(1/q) - 2t\varepsilon$$

Let us see that this implies what we were looking for all along: a locally checkable proof of the satisfiability of φ. Clearly, if φ were satisfiable, there would be an objects oracle and a points oracle that would make the test accept with probability 1. However, if φ is unsatisfiable, then there are *no* good $\overline{Q^{(i)}}$ of the kind defined above (since each is actually derived from a satisfying assignment). Thus, the above implies that

$$\Pr_{\Omega,x}\left[\overline{\mathcal{A}}(\Omega)(x) \neq \overline{f}(x)\right] \geq 1 - \delta - O(1/q) - 2t\varepsilon$$

Now that we have designed something like a low-degree test for satisfying assignments of a 3-CNF formula φ, let us see how this implies the existence of a robust PCP for SAT. This conversion is a generalization of the robust PCP we presented for checking if a given function $f : \mathbb{F}^m \to \mathbb{F}$ at the beginning of Section 9.3. The proof is just a map $\overline{f} = (f_{-1}, f_0, f_1, \ldots, f_{3m+3})$ from $\mathbb{F}^{3m+3}$ to $\mathbb{F}^{3m+5}$. The verifier picks an object $\Omega = (y, y', z, z')$ at random and queries the proof for the value of $\overline{f}$ at each point in Ω and accepts iff the following conditions hold:

- Each coordinate of $\overline{f}$ restricts to a polynomial of degree $O(m|H|)$ over Ω,
- $f_0(x) = \sum_{1 \leq i \leq 3m+3} g_H(x_i) f_i(x)$ for all $x \in L(\Omega)$,
- $f_{-1}(z) = f_{-1}(z')$, and
- $f_0(x) = \tilde{\varphi}(x)(f_{-1}(x) - x_{3m+1})(f_{-1}(\rho(x)) - x_{3m+2})(f_{-1}(\rho^2(x)) - x_{3m+3})$ for all $x \in L(\Omega)$.

To prove that this yields a robust PCP, let us fix, for each Ω, a closest satisfying assignment for the verifier when he queries the points in Ω. For each choice of Ω, this is a $(3m+5)$-tuple of polynomials over Ω of degree $d = O(m|H|)$ satisfying the above mentioned identities. In other words, this defines an object oracle $\overline{\mathcal{A}}$. Our analysis above yields that

$$\mathop{\mathbb{E}}_{\Omega}\left[\mathrm{agr}(\overline{\mathcal{A}}(\Omega), \overline{f}|_\Omega)\right] = \mathop{\mathbb{E}}_{\Omega}\left[\Pr_x\left[\overline{\mathcal{A}}(\Omega)(x) = \overline{f}(x)\right]\right] \leq \delta$$

for $\delta = \mathrm{poly}(m, d/q)$. This shows that the robust soundness error of the PCP is δ.

Let us now set the values for the above variables. We started with an instance of 3-SAT with n variables. We will set the values of $q, m, |H|$ so that the above proof goes through and

we get a good PCP with small soundness error δ, size $\text{poly}(n)$, and randomness $O(\log n)$. For now, the query complexity and alphabet size of the PCP will remain large. In the next lecture, we will bring these parameters under control.

Recall that we need $|H|^m = n$. We set $m = \frac{\log n}{\log \log n}$ and $|H| = \log n$ so that this holds. We will set $q = (\log n)^c$ for some large constant c so that we get robust soundness error $\delta = \frac{1}{\text{polylog}n}$. The size of the proof is $|\mathbb{F}|^{3m+3} = q^{O(m)} = \text{poly}(n)$. Finally, also note that the number of random bits used by the verifier is $O(m \log q) = O(\log n)$.

Note that the proof specifies, for each point $x \in \mathbb{F}^{3m+3}$, the values $\overline{f}(x)$, which is a $(3m+5)$-tuple of field elements. Hence, the size of the alphabet is $q^{O(m)} = \text{poly}(n)$. Also note that an object Ω contains $q^{O(1)}$ points and hence the query complexity of the PCP is $\text{polylog}(n)$.

To summarize, we have constructed a robust PCP for SAT with the following parameters:

Proof Size	:	$\text{poly}(n)$
Randomness	:	$O(\log n)$
Robust soundness error	:	$1/\text{polylog}(n)$
Query Complexity	:	$\text{polylog}(n)$
Alphabet Size	:	$\text{poly}(n)$

Thus, though we have a robust PCP with low soundness error, it uses a large (superconstant-sized) alphabet and has superconstant query complexity. Our aim in the next lecture will be bring these two down to constant size.

References

[DH09] Irit Dinur and Prahladh Harsha. *Composition of low-error 2-query PCPs using decodable PCPs.* In *Proc. 50th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 472–481. 2009. ecc:TR09-042, doi:10.1109/FOCS.2009.8.

[Sud99] Madhu Sudan. *6.893: Approximability of optimization problems*, 1999. (A course on Approximability of Optimization Problems at MIT, Fall 1999).

A Appendix: Proof of Claim 9.3.4

Let $\mathcal{D}_1$ be the distribution on pairs (Ω, s) such that $s \subseteq L(\Omega)$ defined by picking $\Omega = (y, y', z, z')$ at random and then picking a random plane s contained in $L(\Omega)$. Let $\mathcal{D}_2$ be the distribution on pairs (Ω, s) such that $s \subseteq L(\Omega)$ defined by picking a random plane from $\mathbb{F}^m$ and then choosing a random Ω such that s is contained in $L(\Omega)$. We want to argue that $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically close — more precisely, that their statistical distance is $O(1/q)$.

Surely, if $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically close, then so are their marginals in each co-ordinate. We prove this weaker statement below in the case of the marginal in the second co-ordinate. We will use this to prove that $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically close.

Claim A.1. *The marginal of $\mathcal{D}_1$ in the second co-ordinate (i.e., on the planes) is $O(1/q)$-close to uniform.*

Proof. Say a random object $\Omega = (y, y', z, z')$ is picked. To pick a random plane contained in $L(\Omega)$, we pick random vectors (a_1, a_2, a_3) contained in $L(\Omega)$ and consider the plane $\{a_1 + \beta a_2 + \gamma a_3 \mid \beta, \gamma \in \mathbb{F}\}$. Note that a random plane in the entire space can be assumed

to be chosen by picking a'_1, a'_2, a'_3 independently and uniformly at random from $\mathbb{F}^{3m+3}$ and considering the plane as defined above.[3] Hence, it suffices to show that for a random Ω the distribution of the three-tuple of vectors (a_1, a_2, a_3) is close to the uniform distribution over $(\mathbb{F}^{3m+3})^3$. To do this, we proceed as follows.

To pick the tuple (a_1, a_2, a_3), we pick independently and uniformly at random scalars $\alpha_i, \alpha_i^1, \alpha_i^2, \alpha'_i, \beta_i, \gamma_i$ $(i \in \{1, 2, 3\})$ and set $a_i = \alpha_i z + \alpha_i^1 \rho(z) + \alpha_i^2 \rho^2(z) + \alpha'_i z' + \beta_i y + \gamma_i y'$. We show that the following holds:

- For any *fixed* choice of the scalars outside a "bad set", the distribution of the vectors (a_1, a_2, a_3) (for a random $\Omega = (y, y', z, z')$) is exactly the uniform distribution over $(\mathbb{F}^{3m+3})^3$.
- The probability that the scalars lie in the bad set is $O(1/q)$.

The above will prove that the distribution of (a_1, a_2, a_3) is $O(1/q)$-close to uniform and hence complete the proof of Claim A.1.

Let us fix some choice of scalars and try to prove that the resulting distribution on (a_1, a_2, a_3) is indeed uniform. The definition of the bad set will fall out of our analysis.

Fix some set of scalars $\alpha_i, \alpha_i^1, \alpha_i^2, \alpha'_i, \beta_i, \gamma_i$ $(i \in \{1, 2, 3\})$. Note that even for such a fixed choice of scalars, there are dependencies across the co-ordinates of the vectors a_1, a_2, and a_3. This is because of we are taking linear combinations of $z, \rho(z)$, and $\rho^2(z)$. However, we note that these dependencies are rather local, in the following sense. Partition the $3m + 3$ co-ordinates of a_1, a_2, a_3 as follows. Define S_i for $i \in [m + 3]$ by

$$S_i = \begin{cases} \{i, m + i, 2m + i\} & \text{if } i \in [m], \\ \{3m + (i - m)\} & \text{if } i > m. \end{cases}$$

Note that the co-ordinates of a_1, a_2, and a_3 across the different S_i are mutually independent. Thus, to show that the distribution of a_1, a_2, and a_3 is uniform over $(\mathbb{F}^{3m+3})^3$ it suffices to show that for each $i \in [m + 3]$, it is uniform when restricted to the co-ordinates in S_i.

Let us first consider the singleton sets $S_i = \{3(i - 1) + j\}$ for $i > m$. The proof for each S_i is identical. Let j denote $3m + i - m$. We have the following.

$$\begin{aligned} a_{1,j} &= (\alpha_1 + \alpha_1^1 + \alpha_1^2) z_j + \alpha'_1 z'_j + \beta_1 y_j + \gamma_1 y'_j \\ a_{2,j} &= (\alpha_2 + \alpha_2^1 + \alpha_2^2) z_j + \alpha'_2 z'_j + \beta_2 y_j + \gamma_2 y'_j \\ a_{3,j} &= (\alpha_3 + \alpha_3^1 + \alpha_3^2) z_j + \alpha'_3 z'_j + \beta_3 y_j + \gamma_3 y'_j \end{aligned}$$

We want to prove that the distribution of $(a1, j, a_{2,j}, a_{3,j})$ is the uniform distribution over $\mathbb{F}^3$. Note that z'_j, y_j, y'_j are all independently and uniformly chosen from $\mathbb{F}$. Fix any value of z_j. Let M denote the matrix applied to the remaining field elements (z'_j, y_j, y'_j):

$$\begin{pmatrix} \alpha'_1 & \beta_1 & \gamma_1 \\ \alpha'_2 & \beta_2 & \gamma_2 \\ \alpha'_3 & \beta_3 & \gamma_3 \end{pmatrix}$$

[3] Note that we are ignoring what happens when a'_2, a'_3 are not linearly dependent. In this case, we generate a line and not a plane. But this happens with probability at most $\frac{1}{q^m}$. Hence, even if we output some default plane when this does happen, this distribution is $\frac{1}{q^m}$-close to the uniform distribution over planes.

Note that whenever M is a non-singular matrix, then $(a_{1,j}, a_{2,j}, a_{3,j})$ are indeed uniformly distributed over $\mathbb{F}^3$. We say that the choice of scalars is *bad* if M is singular. For a random choice of the scalars, this happens with probability at most $3/q$. Note that this bad event is independent of the choice of $i > m$.

Now fix some S_i for $i \in [m]$. We wish to prove that the distribution on these co-ordinates is uniform. Since the proof is identical for different i, we only prove the statement for $i = 1$. Moreover, by renaming co-ordinates, we assume that $S_i = \{1, 2, 3\}$. We want to show that the tuple $(a_{1,1}, a_{1,2}, a_{1,3}, a_{2,1}, a_{2,2}, a_{2,3}, a_{3,1}, a_{3,2}, a_{3,3})$ is uniform over $\mathbb{F}^9$.

A routine, ugly computation shows that the 9-tuple of field elements $(a_{1,1}, a_{1,2}, a_{1,3}, a_{2,1}, a_{2,2}, a_{2,3}, a_{3,1}, a_{3,2}, a_{3}$ can be written as

$$\begin{pmatrix} a_{1,1} \\ a_{1,2} \\ a_{1,3} \\ a_{2,1} \\ a_{2,2} \\ a_{2,3} \\ a_{3,1} \\ a_{3,2} \\ a_{3,3} \end{pmatrix} = \underbrace{\begin{pmatrix} (\alpha_1 + \alpha_1') & \alpha_1^2 & \alpha_1^1 & \beta_1 & 0 & 0 & \gamma_1 & 0 & 0 \\ \alpha_1^1 & \alpha_1 & \alpha_1^2 & 0 & \beta_1 & 0 & 0 & \gamma_1 & 0 \\ \alpha_1^2 & \alpha_1^1 & \alpha_1 & 0 & 0 & \beta_1 & 0 & 0 & \gamma_1 \\ (\alpha_2 + \alpha_2') & \alpha_2^2 & \alpha_2^1 & \beta_2 & 0 & 0 & \gamma_2 & 0 & 0 \\ \alpha_2^1 & \alpha_2 & \alpha_2^2 & 0 & \beta_2 & 0 & 0 & \gamma_2 & 0 \\ \alpha_2^2 & \alpha_2^1 & \alpha_2 & 0 & 0 & \beta_2 & 0 & 0 & \gamma_2 \\ (\alpha_3 + \alpha_3') & \alpha_3^2 & \alpha_3^1 & \beta_3 & 0 & 0 & \gamma_3 & 0 & 0 \\ \alpha_3^1 & \alpha_3 & \alpha_3^2 & 0 & \beta_3 & 0 & 0 & \gamma_3 & 0 \\ \alpha_3^2 & \alpha_3^1 & \alpha_3 & 0 & 0 & \beta_3 & 0 & 0 & \gamma_3 \end{pmatrix}}_{N} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ y_1 \\ y_2 \\ y_3 \\ y_1' \\ y_2' \\ y_3' \end{pmatrix} + \begin{pmatrix} 0 \\ \alpha_1' z_2' \\ \alpha_1' z_3' \\ 0 \\ \alpha_2' z_2' \\ \alpha_2' z_3' \\ 0 \\ \alpha_3' z_2' \\ \alpha_3' z_3' \end{pmatrix}$$

For any fixed value of z_2' and z_3', the above distribution is uniform over $\mathbb{F}^9$, as long as the matrix N is non-singular; this implies that if N is non-singular, the distribution of $(a_{k,l})_{k,l}$ is uniform. We would like to say that for a random choice of the scalars, N is non-singular with high probability. To see this, consider the determinant of N as a polynomial in the scalars. Clearly, it is a polynomial of degree at most 9. Note that it is a non-zero polynomial, since by appropriate substitutions of the scalars, one can actually make N the identity matrix. Hence, the probability that the determinant of N vanished for a random choice of the scalars is at most $9/q$. Let us call the choice of scalars *bad* if the determinant of N vanishes. Note that as before, the choice of the matrix N does not depend on the set of co-ordinates S_i that we are considering.

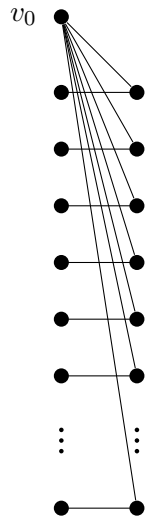
We have shown that as long as the choice of the scalars don't come from a bad set (as defined above), the resulting distribution of (a_1, a_2, a_3) is uniform in each S_i and hence uniform over $(\mathbb{F}^{3m+3})^3$. Moreover, by a union bound, the probability that a random choice of scalars lies in the bad set is at most $12/q$. Hence, the statistical distance of the distribution of (a_1, a_2, a_3) from uniform is at most $12/q$. □

Hence, we see that the process of picking a random object Ω and then picking a random plane s contained in $L(\Omega)$ generates a close-to-random plane. Does this mean that $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically close? A weaker question is whether this even implies that the marginals of $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same on the first co-ordinate (that is, on the objects). Before we answer this question, let us abstract out the situation a little bit. Consider the bipartite graph G_0 whose left-vertices are objects and right-vertices are planes.[4] We draw an edge (Ω, s) iff

[4] We use the self-explanatory terms "left-vertices" and "right-vertices" to denote the vertices in the two different partitions of the bipartite graph.

$s \subseteq L(\Omega)$. Let L denote the number of objects and R the number of planes. For each vertex v of the graph, we use $\deg(v)$ to denote the degree of v.

Note that both distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are distributions over the edge set of G_0. The distribution $\mathcal{D}_1$ (respectively, $\mathcal{D}_2$) is generated by picking a random left (respectively, right) vertex of the graph and then choosing a neighbour at random. We have just shown that the marginal of $\mathcal{D}_1$ on the right vertices is close to uniform. However, note that this by itself does not imply that $\mathcal{D}_1$ and $\mathcal{D}_2$ are close to uniform, or even just that their left marginals are close to uniform. To see this, consider the bipartite graph G whose left vertices consist of a collection of vertices of degree 1 and a single vertex v_0 of large degree (see figure). It is easy to see that in this case, the marginal of the distribution $\mathcal{D}_1$ on the right vertices is indeed uniform, but the marginal of $\mathcal{D}_2$ on the first co-ordinate is very different from that of $\mathcal{D}_1$ (the distribution $\mathcal{D}_2$ puts much more weight on v_0).



We need to show that such a situation does not arise in the case of the graph G_0. We do this by noting that G_0 is close to left-regular, in a well-defined sense. Let d_0 denote the maximum left-degree of G_0. Note that the left-degree of a left-vertex Ω is equal to the number of planes contained in $L(\Omega)$ and is hence completely determined by the dimension of the space $L(\Omega)$ (the more the dimension, the larger the degree). Moreover, for a randomly picked Ω, the space $L(\Omega)$ takes its maximum possible dimension (which is 6) with probability $1 - O(\frac{1}{q^m})$.

Armed with the above fact and Claim A.1, we are ready to show that $\mathcal{D}_1$ and $\mathcal{D}_2$ are indeed statistically close. Below, whenever Ω and s appear in a summation or double summation, they satisfy the property that (Ω, s) is an edge in the graph G_0.

$$\begin{aligned} |\mathcal{D}_1 - \mathcal{D}_2| &= \sum_{(\Omega,s)} |\mathcal{D}_1(\Omega, s) - \mathcal{D}_2(\Omega, s)| \\ &= \sum_s \sum_\Omega \left| \frac{1}{L \deg(\Omega)} - \frac{1}{R \deg(s)} \right| \end{aligned} \tag{A.1}$$

Claim A.1 tells us that the marginals of the distributions of $\mathcal{D}_1$ and $\mathcal{D}_2$ on the planes are statistically $O(1/q)$-close. In our graph terminology, this means that

$$\sum_s \left| \left(\sum_\Omega \frac{1}{L \deg(\Omega)} \right) - \frac{1}{R} \right| = O\left(\frac{1}{q}\right)$$
$$\sum_s \left| \sum_\Omega \frac{1}{L \deg(\Omega)} - \frac{1}{R \deg(s)} \right| = O\left(\frac{1}{q}\right) \tag{A.2}$$

Equation A.2 seems to be a "triangle-inequality" away from Equation A.1. To bridge the gap, we will use the fact that G_0 is close to left-regular.

Define T as follows.

$$T := \sum_s \sum_\Omega \left| \frac{1}{Ld_0} - \frac{1}{R \deg(s)} \right| = \sum_s \left| \sum_\Omega \frac{1}{Ld_0} - \frac{1}{R \deg(s)} \right|$$

We will show that T is close to both $|\mathcal{D}_1 - \mathcal{D}_2|$ and the expression in Equation A.2. First, let us consider $||\mathcal{D}_1 - \mathcal{D}_2| - T|$.

$$\begin{aligned} ||\mathcal{D}_1 - \mathcal{D}_2| - T| &= \left| \sum_{(s,\Omega)} \left| \frac{1}{L \deg(\Omega)} - \frac{1}{R \deg(s)} \right| - \sum_{(s,\Omega)} \left| \frac{1}{Ld_0} - \frac{1}{R \deg(s)} \right| \right| \\ &\leq \sum_{(s,\Omega)} \left| \left| \frac{1}{L \deg(\Omega)} - \frac{1}{R \deg(s)} \right| - \left| \frac{1}{Ld_0} - \frac{1}{R \deg(s)} \right| \right| \\ &\leq \sum_{(s,\Omega)} \left| \frac{1}{L \deg(\Omega)} - \frac{1}{Ld_0} \right| \end{aligned}$$

Since d_0 is the maximum left-degree of the graph G_0, the above can be bounded as follows

$$\begin{aligned} ||\mathcal{D}_1 - \mathcal{D}_2| - T| &\leq \sum_{(s,\Omega):\deg(\Omega)\neq d_0} \frac{1}{L \deg(\Omega)} \\ &= \sum_{\Omega:\deg(\Omega)\neq d_0} \frac{1}{L} = \Pr_\Omega [\deg(\Omega) \neq d_0] = O\left(\frac{1}{q^m}\right) \end{aligned}$$

We now show that T is close to expression in Equation A.2. The proof is almost exactly the same as the above proof, with the only difference being that we use a different expression

for T.

$$\begin{aligned}
\left|\sum_s \left|\sum_\Omega \frac{1}{L\deg(\Omega)} - \frac{1}{R\deg(s)}\right| - T\right| &= \left|\sum_s \left|\sum_\Omega \frac{1}{L\deg(\Omega)} - \frac{1}{R\deg(s)}\right| - \sum_s \left|\sum_\Omega \frac{1}{Ld_0} - \frac{1}{R\deg(s)}\right|\right| \\
&\le \sum_s \left|\left|\sum_\Omega \frac{1}{L\deg(\Omega)} - \frac{1}{R\deg(s)}\right| - \left|\sum_\Omega \frac{1}{Ld_0} - \frac{1}{R\deg(s)}\right|\right| \\
&\le \sum_s \left|\sum_\Omega \frac{1}{L\deg(\Omega)} - \frac{1}{Ld_0}\right| \\
&\le \sum_{(s,\Omega):\deg(\Omega)\neq d_0} \frac{1}{L\deg(\Omega)} \\
&= \sum_{\Omega:\deg(\Omega)\neq d_0} \frac{1}{L} = \Pr_\Omega\left[\deg(\Omega)\neq d_0\right] = O\left(\frac{1}{q^m}\right)
\end{aligned}$$

Combined with Equation A.2, the above implies that $|\mathcal{D}_1 - \mathcal{D}_2| = O(\frac{1}{q}) + O(\frac{1}{q^m}) = O(\frac{1}{q})$. This completes the proof of Claim 9.3.4.