Hardness Of Approximation

Lecture 7: [Hardness of Label Cover and Agreement Tests]

Instructor: Irit Dinur and Amey Bhangale *Scribe: [Irit Dinur]*

1 Label Cover

Recall the definition of the label cover problem. An instance is given by a bipartite graph $G = (A, B, E)$. Each vertex v is associated with a label set L_v. Each edge $uv \in E$ is associated with a relation $\pi_{uv} \subset L_u \times L_v$. A labeling $f = \{f_v \in L_v : v \in A \cup B\}$ is an assignment of one label per each vertex. The value of the assignment is

$$val(f) = \Pr_{uv \sim E}[(f_u, f_v) \in \pi_{uv}]$$

If edges in E have weights then the probability of choosing an edge is proportional to the edge weights. In this lecture we will sketch a proof for the following theorem

Theorem 1.1 (Label Cover theorem)**.** *For every $\varepsilon > 0$, Gap $(1, \varepsilon)$-label cover is NP-hard, with projection constraints and labels of size at most $(\varepsilon)^{O(1)}$.*

The proof proceeds by reduction from the basic PCP theorem, namely,

Theorem 1.2 (Basic PCP theorem)**.** *Gap $(1, 0.999)$-3SAT is NP-hard.*

In more words: it is NP-hard to decide if a given 3SAT formula is completely satisfiable, or every assignment satisfies at most 0.999 fraction of clauses. Moreover, the result holds for 3SAT formulae where each variable occurs in the same number of clauses.

We have seen in the exercise that the basic PCP theorem immediately implies a gap of $(1, 0.9999)$ for label cover. Our goal is to amplify the gap through a direct product construction.

$$gap(1, 1 - \frac{1}{|E|}) \overset{Basic\ PCP\ Thm}{\longrightarrow} gap(1, 0.9999) - gap \overset{direct\ product}{\longrightarrow} gap(1, \varepsilon) - gap$$

Construction of Label Cover. Given a 3SAT instance φ with n variables $V = \{v_1, \ldots, v_n\}$ and m clauses $\{C_1, \ldots, C_m\}$. Consider the following random process:

- Choose k clauses uniformly and independently, $(C_1, \ldots, C_k)$. Let $s \subset V$ be the set of all variables participating in the clauses $C_1, \ldots, C_k$. (Clearly $|s| = 3k$ unless there are unlikely collisions.)
- For each i with probability 0.9 choose a variable $v_i \in C_i$; otherwise choose nothing. Let $t \subset V$ be the set of chosen variables. (Clearly $|t| \approx 0.9k$)
- Output (s, t).

The label cover instance will have vertex sets S and T where S is the set of all possible sets s in the random process and T is the set of all possible sets t in the random process. An edge (s, t) will have weight according to the probability it is output in the random process.

The label set corresponding to $t \in T$ is $L_t = \{0,1\}^t$, the set of all possible assignments to the variables in t. The label set corresponding to $s \in S$ is L_s the set of all assignments for the variables of s, such that these assignments satisfy every clause contained in s.

$$L_s = \{f : s \to \{0,1\} \mid \ \forall C \subset s, f \text{ satisfies } C\}$$

The relation on the edge (s,t) allows only pairs of consistent assignments:

$$\pi_{st} = \{(f,g) \in L_s \times L_t \mid \ f|_t = g\}$$

Let us denote by LC_k the resulting label cover instance.

Claim 1.3 (Completeness)**.** *If $val(\varphi) = 1$ then for all k $val(LC_k) = 1$.*

Proof. Let $g : V \to \{0,1\}$ be an assignment that satisfies all of the clauses of φ. Label each $s \in S$ by $g|_s \in L_s$ and each $t \in T$ by $g|_t$. Clearly this satisfies all of the edge constraints and has value 1. □

Lemma 1.4 (Soundness)**.** *If $val(\varphi) < 0.999$ then $val(LC_k) \leq \exp(-k)$.*

Together Lemma 1.4 and Claim 1.3 prove Theorem 1.1.

Why is the soundness lemma true? We have seen in exercise 1 that if we consider only labelings that are globally consistent with some underlying assignment to the variables, then the soundness must be exponentially small. Indeed fix $h : V \to \{0,1\}$, and suppose that we label every $t \in T$ with $h|_t \in L_t$. Assuming $val(\varphi) < 0.999$, h must falsify a constant fraction of clauses. A Chernoff bound shows that nearly all $s \in S$ have a constant fraction of variables that are assigned a value different from $h|_s$ (indeed $h|_s \notin L_s$ is not even a valid label). A random edge (s,t) will be inconsistent except with probability exponential in k.

We call labelings that are consistent with a global h *direct product labelings.*[1]

The main question is whether labelings can benefit significantly if they deviate from a direct product labeling. Naively, we would like to show that whenever a labeling has value above ε, it must be close to a direct product on some ε' fraction of the label cover nodes. Unfortunately, this is false. Instead, we will be able to prove a weaker statement that will suffice. We will show that if a labeling has large value, it must look like a global labelling on special sub-instances.

Each set $r \subset V$ defines a sub-instance consisting of all nodes that contain (as subsets) the set r,

$$S_r = \{s \in S \mid \ s \supset r\}.$$

For $|r| = 0.1k$ consider the following distribution on S_r: For each $v \in r$ we choose a random clause containing v. Choose another $0.9k$ clauses uniformly and independently. Output s, the set of variables of these clauses.

Lemma 1.5 (Global structure on restrictions)**.** *There is some $\gamma > 0$ such that if $\{f_s\}, \{g_t\}$ is a labeling for LC_k with value above $\varepsilon > (1-\gamma)^k$ then there is a restriction $r \subset V$, $|r| = 0.1k$, and a global assignment $h : V \to \{0,1\}$ such that*

$$Prob_{s \sim S_r}[h|_s \overset{\alpha k}{=} f_s] > poly(\varepsilon)$$

where the notation $x \overset{\alpha k}{=} x'$ means that x, x' differ on at most αk points, and we set $\alpha = 10^{-5}$.

[1] This name originates from a similar situation where the sets s are ordered as tuples, and then the assignment for s is really a direct product of h, and so the name remains.

We will discuss this lemma and its proof further below, and in the next lecture. First, let us see that it is useful. We show how to derive the soundness lemma from it.

Proof of Lemma 1.4. Suppose we are given a labelling $\{f_s, g_t\}$ with value above ε. By the structure lemma there is some restriction $r \subset V$, $|r| = 0.1k$, and a global assignment $h : V \to \{0, 1\}$ such that

$$Prob_{s \sim S_r}[h|_s \overset{\alpha k}{=} f_s] > poly(\varepsilon).$$

h must violate at least 0.001 fraction of the clauses of φ. Except for an exponentially small fraction of $s \in S_r$, s must contain at least $10^{-4}k$ violated clauses. This means that the label f_s must differ from $h|_s$ on at least that many elements, which is more than αk, a contradiction. □

2 Agreement tests

The structure lemma is really a kind of so-called agreement test. The setup for an agreement test is this.

- V - global set of vertices
- S a collection of subsets of V. We begin by focusing on $S = \binom{V}{\leq k}$ - the set of all $\leq k$-element sets in V.
- For every $s \in S$ a space of available local functions $L_s \subset \{0, 1\}^s$

A collection of local functions $\{f_s \in L_s\}$ is a perfect collection if there is some $h : V \to \{0, 1\}$ such that $f_s = h|_s$ for every $s \in S$. An agreement test tests, in the property testing sense, if a given collection is perfect. Here is the test relevant to us

Agreement test with parameter $\rho > 0$:

- Choose $v_1, \ldots, v_k \in V$ independently and uniformly. Let $s = \cup\{v_i\}$.
- For each $i = 1, ..., k$ let $v'_i = v_i$ with probability ρ and v'_i uniform in V otherwise. Let $s' = \cup_i \{v'_i\}$.
- Accept iff $f_s(v) = f_{s'}(v)$ for all $v \in s \cap s'$.

Typically one wants to prove that if the test succeeds with significant probability, then $\{f_s\}$ is close to perfect. We will prove

Lemma 2.1 (Global structure on restrictions). *There exists $\gamma > 0$ such that the following holds. Suppose that $\{f_s\}$ passes the agreement test with probability $\varepsilon > (1 - \gamma)^k$. Then there is a restriction $r \subset V$, $|r| = 0.1k$, and a global assignment $h : V \to \{0, 1\}$ such that*

$$Prob_{s \supset r}[h|_s \overset{\alpha k}{=} f_s] > poly(\varepsilon).$$

This statement might appear weak in that we only prove closeness to perfect on a tiny fraction of the domain. Indeed the fraction of sets s that contain a fixed r is $O(n^{-r})$ and certainly not polynomial in ε. We have already seen that such a theorem is useful nevertheless. Let us also see that there is no real hope of proving anything stronger. The reason is the small set expansion behavior of the graph underlying the test.

2.1 The Direct Product graph DP_ρ

Let us look at the graph underlying the test. The direct product graph, "the DP_ρ graph", has vertex set $[n]^k$ and the edge weights are described as in the test: starting at a vertex $(v_1, \ldots, v_k)$ we move to a neighbor $(v_1', \ldots, v_k')$ by choosing for each i independently with probability ρ $v_i' = v_i$ and with the remaining probability v_i' is chosen uniformly from $[n]$.

Remark 2.2 (The Johnson graph $J(n,k)$). *The Johnson graph is the graph whose vertices are the sets $s \in \binom{V}{k}$ and edges connect two sets if their intersection has size $k-1$. If we think of random walks of appropriate length t in this graph with $e^{-t/k} = \rho$ (so that the intersection between s, s' is expected to be ρk) we get a graph that has similar behaviour to the direct product graph.*

Definition 2.3 (The folded DP graph). *The folded DP graph is the graph obtained from DP_ρ by gluing together all vertices $(v_1, \ldots, v_k)$ for which $\cup\{v_i\}$ is the same. We denote this graph by $\widehat{DP}_\rho$.*

This is the graph we are really looking at, and because it is obtained by folding a larger graph, its expansion behavior is governed by that of the DP graph (why?).

Suppose we are given that the first set chosen by the test lands in S_r. What is the probability of s' also landing in S_r? This is easy to compute, and it is exactly $\rho^{|r|}$. Observe that S_r is tiny with respect to $V(DP_\rho)$ the entire collection of subsets even for $|r| = 1$, yet it has many edges that stay inside of it. Thus, we can construct a labeling $\{f_s\}$ that demonstrates that there can be no qualitative improvement to Lemma 2.1.

- Choose r at random, with $|r| = \lfloor \log_\rho \varepsilon \rfloor$.
- Choose a random function $h_r : V \to \{0,1\}$ and assign all $s \supset r$ that weren't previously assigned with $f_s = h|_s$
- Repeat the above until all of the s's are assigned.

It is easy to see that this labeling passes the agreement test with probability $O(\varepsilon)$ yet its correlation to any perfect labeling is $o_n(1)$ (much smaller than any polynomial in ε as $n \to \infty$).