Georgia Tech Fall'04
CS 8002 : PCPs and Hardness of Approximation

Lecture 3: Hardness of Set Cover

Lecturer: *Subhash Khot* Scribe: *Deeparnab Chakrabarty*

1 Set Cover Problem

Problem Statement: Given a universe $\mathcal{U} = \{1, \cdots, n\}$ and sets $S_1, S_2, \cdots S_m \subseteq [n]$ such that $\bigcup S_i = [n]$, the goal is to find minimum t and sets $S_{i_1}, \cdots, S_{i_t}$ such that $S_{i_1} \cup \cdots \cup S_{i_t} = [n]$.

This problem has a greedy $\ln n$ factor approximation algorithm. (see for example [Vaz01]). From Feige's result it follows that this is the best one can hope for.

Theorem 1
[Fei98] It is hard to approximate Set Cover *within* $(1-\epsilon)\ln n$ *if* $\mathbf{NP} \subsetneq \mathbf{DTIME}(n^{O(\log\log n)})$

In this lecture we shall derive a weaker result, that is, a $\Omega(\log n)$ hardness result for the Set Cover problem. The reduction is essentially same as the reduction of Lund and Yannakakis [LY94] and the following shows the sequence of reductions used.

Gap-MAX3SAT(5) $\rightarrow$ Label Cover $\xrightarrow[\text{Raz's Parallel Repetition}]{\text{Hardness Amplification}}$ Label Cover $\xrightarrow[\text{Systems}]{\text{Partition}}$ Set Cover

2 Gap-MAX3SAT(5) and Label Cover

2.1 MAX3SAT(5)

Problem Statement: Given a $3SAT$ formula with the further restriction that each variable is in exactly 5 clauses, MAX3SAT(5) asks whats the maximum fraction of clauses that can be satisfied by any assignment.

Fact: Gap-MAX3SAT(5) is **NP**-hard. Equivalently, it is **NP**-hard to tell if an instance ϕ has

Yes: $OPT(\phi) = 1$ or
No: $OPT(\phi) \le c < 1$ for some constant c

2.2 Label Cover

Problem Statement: An instance of Label Cover is denoted by $\mathcal{L}(G(V, W, E), [M], [N], \{\pi_{vw}\}_{(v,w)\in E})$, where

- $G(V, W, E)$ is a regular bipartite graph.
- $[M], [N]$ are sets of labels, and V, W get labels from $[N], [M]$ respectively.
- $\{\pi_{vw}\}_{(v,w)\in E}$ denote the constraints on each edge. For every edge (v, w) in E, we have a map $\pi_{vw} : [M] \to [N]$

A labeling $l : V \to [N], W \to [M]$ satisfies (the constraint on) an edge (v, w) if $\pi_{vw}(l(w)) = l(v)$. Given an instance $\mathcal{L}$, the goal is to find a labeling which satisfies the maximum fraction, denoted by $OPT(\mathcal{L})$, of edges.

THEOREM 2
It is **NP**-*hard to tell whether a label cover instance* $\mathcal{L}(G(V, W, E), [7], [2], \{\pi_{vw}\}_{(v,w)\in E})$ *has*
YES: $OPT(\mathcal{L}) = 1$ *or*
NO: $OPT(\mathcal{L}) \leq c' < 1$ *for some constant* c'

PROOF:Reduction from GAP-MAX3SAT(5)

Given an instance ϕ of MAX3SAT(5) we construct an instance $\mathcal{L}_\phi$ of LABEL COVER as follows. Let ϕ have variables $\{x_1, \cdots, x_n\}$ and clauses $\{C_1, \cdots, C_m\}$. Then the graph $G_\phi(V, W, E)$ has $V := \{x_1, \cdots, x_n\}$ and $W := \{C_1, \cdots, C_m\}$. Also $(x_i, C_j) \in E$ if and only if $x_i \in C_j$. Note that the degree of a vertex in V is 5 while that in W is 3.

We now define the constraints. Since V gets labels from [2] we can assume it to be from the set $\{0, 1\}$. Let [7] denote the seven satisfying assignments that a clause $C_j \in W$ can have in some order. $\pi_{x_i,C_j} : [7] \to [2]$ be the assignment of variable x_i by the satisfing assignment. For instance, if the clause C_j were $x_1 \vee x_2 \vee x_3$ and the third satisfying assignment were $x_1 = 1, x_2 = 0, x_3 = 1$, then $\pi_{x_1,C_j}(3) = 1$ while $\pi_{x_2,C_j}(3) = 0$.

Claim: If $OPT(\phi) = 1$, then $OPT(\mathcal{L}_\phi) = 1$. If $OPT(\phi) \leq c$ then $OPT(\mathcal{L}_\phi) \leq 1 - \frac{1-c}{3}$

PROOF:If $OPT(\phi) = 1$, then there is an assignment which satisfies all the clauses. Thus labeling x_i as per the satisfying assignment, and giving C_j the label corresponding to the satisfying assignment, we see that we can satisfy all the edges. Hence $OPT(\mathcal{L}_\phi) = 1$.

On the other hand, say $OPT(\phi) \leq c$. Consider any labeling l of $\mathcal{L}$. Since the labeling of V can be thought of as an assignment, we know that at least $1 - c$ fraction of clauses C_j are unsatisfied by this assignment. Pick one such clause $C_j = x_1 \vee x_2 \vee x_3$. Since C_j is unsatisfied, whatever label it might have, one of its three neighbors (which are exactly x_1, x_2, x_3) must be violating the constraint (otherwise the clause C_j would be satisfied). Thus at least $\frac{1-c}{3}$ of the edges are violated and hence we have $OPT(\mathcal{L}_\phi) \leq 1 - \frac{1-c}{3}$.

The proof of the theorem follows from the claim. ∎

3 Raz's Parallel Repetition Theorem and Gap amplification

In this section we shall see how the gap obtained for LABEL COVER in the previous section can be amplified using Raz's parallel repetition theorem.

Consider the following reduction from one instance of LABEL COVER to another. Let $\mathcal{L}(G(V,W,E),[M],[N],\{\pi_{vw}\})$ be an instance. Consider the instance $\mathcal{L}' = \mathcal{L}^{\otimes k}(G'(V',W',E'),[M'],[N'],\{\pi'_{v'w'}\})$ where

- $V' := V^k$, k times set cartesian product.
- $W' := W^k$
- $[M'] = [M]^k, [N'] := [N]^k$
- $(v',w') \in E'$ iff $(v_i,w_i) \in E \; \forall \; 1 \leq i \leq k$ where $v' = (v_1,\cdots,v_k), w' = (w_1,\cdots,w_k)$
- $\pi'_{v'w'}(b_1,\cdots,b_k) := (\pi_{v_1w_1}(b_1),\cdots,\pi_{v_kw_k}(b_k))$

We would like to know how $OPT(\mathcal{L}')$ is related to $OPT(\mathcal{L})$. It's not hard to see that if $OPT(\mathcal{L}) = c$ then $OPT(\mathcal{L}') \geq c^k$. It was believed for a long time that $OPT(\mathcal{L}') = OPT(\mathcal{L})^k$ but this was proved wrong by Fortnow. The following theorem of Raz shows that although it is possible to have $OPT(\mathcal{L}') > OPT(\mathcal{L})^k$, one does have an upper bound on $OPT(\mathcal{L}')$ that is exponentially small in k.

THEOREM 3
[Raz98] If $OPT(\mathcal{L}) \leq c < 1$, then there exists a $c' < 1$ depending only on c, M, N such that $OPT(\mathcal{L}^{\otimes k}) \leq c'^k$

Combining the above theorem with the results of the previous section, we get the following theorem :

THEOREM 4
*There is a reduction from SAT (or any other **NP**-complete problem) to a gap version of LABEL COVER $\mathcal{L}(G(V,W,E),[7^k],[2^k],\{\pi_{vw}\})$, such that*

YES INSTANCE: $OPT(\mathcal{L}) = 1$ *or*

NO INSTANCE: $OPT(\mathcal{L}) \leq 2^{-\gamma k}$ *for some absolute constant $\gamma > 0$*

where $|V|,|W| = n^{O(k)}$ and n is the size of the SAT instance.

4 Reduction to Set Cover

In this concluding section we shall reduce the instance of label cover to an instance of set cover.

We first define a partition system.

PARTITION SYSTEM: A partition system $\mathcal{P}(\mathcal{U},m,h,t)$ consists of

- $\mathcal{U}$: Universe on m elements
- t pairs of sets $(A_1,\overline{A_1}),\cdots,(A_t,\overline{A_t})$, with $A_i \subseteq [m]$

Furthermore it has the property that for any indices $i_1 < i_2 \cdots < i_h$, $A^*_{i_1} \cup \cdots \cup A^*_{i_h} \subsetneq [m]$ where A^*_i is either A_i or $\overline{A_i}$. In words, no h sets, one from each pair, can cover the whole universe.

LEMMA 5
Given h, t there exists a partition system with $m = O(2^h h \log t)$. Moreover it can be constructed efficiently with high probability.

PROOF:Pick $A_1, A_2 \cdots, A_t$ at random from $2^{\mathcal{U}}$. Fix any h of the sets, say for notational ease $A_1, \cdots, A_h$. Then

$$\begin{aligned}\Pr[\bigcup_{i=1}^{h} A_i = \mathcal{U}] &= \Pr[x \in \bigcup_{i=1}^{h} A_i]^m \\ &= (1 - \frac{1}{2^h})^m\end{aligned}$$

Now since the total number of ways of fixing h sets is $\binom{t}{h}2^h$, we have

$$\Pr[(A_1, A_2, \cdots, A_t)\text{does not form a partitioning system}] \leq \binom{t}{h} 2^h (1 - \frac{1}{2^h})^m$$

Thus if $m = O(2^h h \log t)$, this probability is very small, and with high probability we have a partition system. ■

(δ, h')-PSEUDOLABELING: Given a label cover instance $\mathcal{L}$, an assignment $v \to T_v, w \to T_w$ for all $v \in V, w \in W$ where $T_v \subseteq [N], T_w \subseteq [M]$ is called a (δ, h')-Pseudolabeling if for *at least* δ fraction of the edges (v, w) of $\mathcal{L}$ we have $|T_v|, |T_w| \leq h'$ and $\pi_{vw}(T_w) \cap T_v \neq \varnothing$. That is, for atleast δ fraction of the edges (v, w), the sets T_v, T_w are bounded in size by h' and there is a label $j \in T_w$ and a label $i \in T_v$ such that $\pi_{vw}(j) = i$.

LEMMA 6
If a label cover instance $\mathcal{L}$ has a (δ, h')-Pseudolabeling, then there is a labeling which satisfies more than $\frac{\delta}{h'^2}$ fraction of the edges and therefore $OPT(\mathcal{L}) \geq \frac{\delta}{h'^2}$.

PROOF:Given any pseudolabeling, for a vertex v (vertex w resp.) pick an element at random from T_v (T_w resp.) and let it be the label. Now since for at least δ fraction of the edges (v, w) at least one pair of labels from $T_v \times T_w$ is "consistent", we shall satisfy an edge with probability δ/h'^2. Thus the expected fraction of satisfied edges is δ/h'^2, which implies there is a labeling acheieving atleast this value. Thus $OPT(\mathcal{L}) \geq \delta/h'^2$. ■

We are now ready to give the reduction from the label cover instance $\mathcal{L}(G(V, W, E), [7^k], [2^k], \{\pi_{vw}\})$ to SET COVER. In particular we shall show that if $OPT(\mathcal{L}) = 1$ then our SET COVER instance would have a cover of size $|V| + |W|$. On the other hand, if $OPT(\mathcal{L}) \leq 2^{-\gamma k}$, our instance would have a cover larger than $\rho h |W|$, where h would be a parameter in our construction and ρ would be a constant. Thus from Theorem 4 we would have a $\Omega(h)$ hardness result for set cover. We now describe the reduction. We assume w.l.o.g. that $|V| = |W|$ (make a suitable number of copies of every vertex in V. This preserves the value of $OPT(\mathcal{L})$).

We shall call the reduced SET COVER instance $SC_{\mathcal{L}}$. Let h be a parameter. For each edge (v, w) we define a subuniverse as follows. From Lemma 5 we know that we have a partition system $\mathcal{P}_{vw}(\mathcal{U}_{vw}, m, h, t = N = 2^k)$, where $m = 2^h O(hk)$ is the size of the subuniverse $\mathcal{U}_{vw}$. Let the universe of $SC_{\mathcal{L}}$ be the disjoint union of all these subuniverses.

Now we define the sets. Firstly note that for each edge (v, w), we have 2^k sets of the partition system $\mathcal{P}_{vw}$. Denote them as $\{A^{vw}\}_i$ with $1 \leq i \leq N$. We shall have $N|V| + M|W|$ sets as follows.

$$\forall v \in V, 1 \leq i \leq N \qquad S_{v,i} := \bigcup_{w:(v,w)\in E} A_i^{vw}$$

$$\forall w \in W, 1 \leq j \leq M \qquad S_{w,j} := \bigcup_{w:(v,w)\in E} \overline{A_{\pi(j)}^{vw}}$$

where for notational simplicity we let $\pi = \pi_{vw}$.

Claim: If $OPT(\mathcal{L}) = 1$, $SC_{\mathcal{L}}$ has a cover of size $|V| + |W|$

Proof:Let $l : V \to [N], W \to [M]$ be the correct labeling. Thus we have for all edges $(v, w) \in E$, $\pi_{vw}(l(w)) = l(v)$. Then the sets $\{S_{v,l(v)}, S_{w,l(w)} : v \in V, w \in W\}$ cover the universe. Indeed, the subuniverse $\mathcal{U}_{vw}$ will be covered by the sets $S_{v,l(v)}$ and $S_{w,l(w)}$ since $A_{l(v)}^{vw} \subseteq S_{v,l(v)}$ and $\overline{A_{\pi(l(w))}^{vw}} \subseteq S_{w,l(w)}$ and $\pi(l(w)) = l(v)$.

Claim: If $SC_{\mathcal{L}}$ has a set cover of size $\rho h|W|$, then there is a $(1-4\rho, h)$-Pseudolabeling of $\mathcal{L}$.

Proof:Suppose there is a set cover of size $\rho h|W|$. For every $v \in V$ assign it the set $T_v := \{i : S_{v,i}$ is in the cover$\}$. Similarly, for $w \in W$ assign it the set $T_w := \{j : S_{w,j}$ is in the cover$\}$. Disregard all those vertices for which $|T_v| \geq h/2$. Since there are atmost $\rho h|W|$ sets in all, we would be disregarding atmost 2ρ fraction of the vertices from either side. Thus there are greater than $1 - 4\rho$ pairs (v, w) such that both the sets T_v, T_w have size less than $h/2$.

Consider such a pair (v, w). Note that the sets $\{S_{v,i} : i \in T_v\}$ and $\{S_{w,j} : j \in T_w\}$ cover the subuniverse $\mathcal{U}_{vw}$. Thus the sets $\{A_i^{vw} : i \in T_v\}$ and $\{\overline{A_{\pi(j)}^{vw}} : j \in T_w\}$ cover $\mathcal{U}_{vw}$. But since there are atmost h sets in all, by the property of the partition system, there must be a complementary pair. In other words, there must be $i \in T_v$ and $j \in T_w$ such that $\pi(j) = i$. Thus the assignment was a $(1 - 4\rho, h)$-pseudolabeling.

Now we are almost done. By Lemma 6 we know that there exists a c/h^2 labeling, where $c = 1 - 4\rho$. Choose $\rho = 1/8$. Thus if $h = O(2^{\gamma k/2})$ we see that $OPT(\mathcal{L}) \leq 2^{-\gamma k}$ implies size of set cover of $SC_{\mathcal{L}}$ is atleast $h|W|/8$, which gives us, as noted above, a $\Omega(h)$ hardness result.

We set k so that $N' = n^{O(k)} \approx 2^h O(hk)$. Clearly, $k = O(\log\log n)$ suffices. The size of the set cover instance is $n^{O(k)} 2^h O(hk) \approx N'^2$ and the hardness factor is $\Omega(h) = \Omega(\log N')$.

Thus we have the following hardness result.

Theorem 7

It is hard to approximate Set Cover *within* $\Omega(\log n)$ *if* $\mathbf{NP} \subsetneq \mathbf{DTIME}(n^{O(\log\log n)})$.

References

[Fei98] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–653, 1998.

[LY94] Carsten Lund and Mihalis Yannakakis. On the hardness of approximating minimization problems. *JACM*, 41(5):960–981, 1994.

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM Journal of Computing*, 27(3):763–803, 1998.

[Vaz01] Vijay V. Vazirani. *Approximation Algortihms.* Springer-Verlag, 2001.