6.893 Approximability of Optimization Problems	December 6, 1999
Lecture 24	
Lecturer: Madhu Sudan	*Scribe: Prahladh Harsha*

In today's lecture, we shall show that SET-COVER cannot be approximated to within a factor of $(1-\epsilon)\log_4 n$ for any $\epsilon > 0$ unless $\mathbf{NP} \in \mathbf{DTIME}(n^{\log\log n})$. As in earlier lectures, properties of PCPs will be exploited to obtain this result. Just as the small answer size was used to show the hardness of approximating MAX-SAT and the c/s ratio for CLIQUE, we shall employ MIP with 2 provers to obtain the hardness for SET-COVER. A few other approximation results which also use this as a basis are

- Hardness of Positive Quadratic Programming. This problem is to maximise $x^T Qx$ such that $Ax \le b$ where A, Q, b are non-negative matrices.
- Closest Vector Lattice Problem.

The SET-COVER hardness result, that we shall discuss in today's lecture is due to Lund and Yannakakis [LY 94]. A strengthening of this result can be found in [F 98].

1 Starting Proof System

We shall use a 2-prover 1-round $\mathrm{MIP}_{1,\frac{1}{n}}[\log^2 n, 2, \log n]$ (ie., a MIP proof system that throws $\log^2 n$ random coins, asks the 2 provers a question each and obtains answers of size $\log n$). Such a proof system can be obtained from the regular $\mathrm{MIP}_{1,1-\epsilon}[\log n, 2, 1]$ and then reducing the soundness error using Raz's Parallel Repetition Theorem. We can assume the following properties regarding the verifier in the MIP system. Let L be any language in $\mathbf{NP}$ and let V be the verifier for the $\mathrm{MIP}_{1,\frac{1}{n}}[\log^2 n, 2, \log n]$ system for L. Then on any input $x \in \Sigma^*$(where Σ is the alphabet), (let $n = |x|$)

- The verifier V tosses $r(n) = (O(\log^2 n))$ random coins to obtain the random string R.
- Based on the input x, the random string R, the verifier V computes the queries q_1, q_2 to be asked to the provers P_1, P_2 respectively and the projection function π.
- If a_1 and a_2 are the answers provided by the provers to the queries q_1 and q_2 respectively, then the verifier V accepts if $\pi(a_2) = a_1$.

For convenience, let Q_1, Q_2, A_1 and A_2 represent the question spaces and answer spaces for a particular input $x \in \Sigma^*$ over all random strings. (ie., $Q_i = \{q_i | V$ outputs query q_i to prover P_i on input x and some random string $R\}$, $A_i = \{a_i | a_i$ is the answer of Prover P_i for query q_i of V on input $x\}$). We can easily modify the MIP system to accommodate the following regularity conditions. We can assume that there exist integers k_1, k_2 such that.

- The question spaces Q_1 and Q_2 are of equal size, ie., $|Q_1| = |Q_2| (= Q$ say).
- For all random strings R and all answers a_1 of prover P_1, there exist exactly k_1 strings a_2 such that $\pi(a_2) = a_1$.
- $\forall q_1 \in Q_1$, there exist exactly k random strings R such that the verifier V on input x and random string R produces query q_1.

2 Ingredients for the Gadget

In the reduction to SET-COVER from L, we shall build an instance of SET-COVER from any input string $x \in \Sigma^*$ such that the Universe is roughly the set of all random strings and the subsets considered relate to the choice of answers to the queries.

The basic ingredient behind the gadget will involve the following: A universe $\mathcal{X} = \{1, 2, \ldots, n\}$ and subsets $S_1, S_2, \ldots, S_m, S_{m+1}, S_{m+2}, \ldots, S_{2m}$ such that $S_i = \overline{S_{m+i}}$

An easy way to cover the set $\mathcal{X}$ is obviously a pair (S_i, S_{m+i}). We shall call such a covering a good covering. The covering $(S_{i_1}, S_{i_2}, \ldots, S_{i_k})$ is called a bad covering if both a set and its complement do not occur in the covering (ie., there does not exist a pair (i_p, i_q) such that $i_q = m + i_p$). It can easily be shown by greedy property that if $m > \log n$ then there exists a bad covering. $(\mathcal{X}, S_1, \ldots, S_{2m})$ is called a (n, m, k)-gadget if there does not exist any bad covering of size less than k. It can easily be shown by probabilistic methods that such (n, m, k)-gadgets exist.

Proposition 1 *For all m, k, there exists a (n, m, k)-gadget for some $n = \mathrm{poly}(m, 2^k)$.*

It is to be noted that even deterministic constructions are known for the above.

3 Construction of gadget

The basic idea behind the construction is as follows: Let L be any language in **NP** as before. For any $x \in \Sigma^*$, we shall construct an instance $\mathcal{S}_x$ of Set-Cover such that

- $x \in L \Rightarrow \text{Set-Cover}(\mathcal{S}_x) = |Q_1| + |Q_2|$.
- $x \notin L \Rightarrow \text{Set-Cover}(\mathcal{S}_x) \geq (1 - \epsilon) \log_4 n(|Q_1| + |Q_2|)$ for any $\epsilon > 0$

Let $m = |A_1|$ and k be an integer to be fixed later. Consider the (N, m, k)-gadget $\mathcal{X}$ generated by proposition 1. The universe $\mathcal{U}$ for $\mathcal{S}_x$ will involve $2^{r(n)}$ disjoint copies of $\mathcal{X}$ where $r(n)$ is the number of random coins tossed by the verifier V on input x.

$$\mathcal{U} = \{(x, R) | x \in \mathcal{X}, R \in \{0, 1\}^{r(n)}\}$$

For notational convenience let us denote the copy of the (N, m, k)-gadget corresponding to the random string R by $(\mathcal{X}^{(R)}, S_1^{(R)}, \ldots, S_{2m}^{(R)})$. Note that as $m = |A_1|$, we can index the subsets $S_1^{(R)}, \ldots, S_m^{(R)}$ by members of A_1 instead of $1, \ldots, m$. Let $q[R, i]$ denote the query that V asks prover P_i on input x and random string R. The subsets of $\mathcal{U}$ in the Set-Cover instance $\mathcal{S}_x$ are of two types - $S_{(q_1, a_1)}, \forall q_1 \in Q_1, a_1 \in A_1$ and $S_{(q_2, a_2)}, \forall q_2 \in Q_2, a_2 \in A_2$. These subsets are defined as follows:

$$\begin{aligned} S_{(q_1, a_1)} &= \bigcup_{R : q[R,1] = q_1} S_{a_1}^{(R)} \\ S_{(q_2, a_2)} &= \bigcup_{R : q[R,2] = q_2} S_{\pi(a_2)+m}^{(R)} \end{aligned}$$

It is clear from the above definitions and from the property of (N, m, k)-gadgets that the following is true.

Lemma 2 *Completeness : If $x \in L$, then* $\text{Set-Cover}(\mathcal{S}_x) = |Q_1| + |Q_2| = 2Q$.

To prove the the soundness counterpart of the above lemma, we have to argue that if $x \notin L$, then no small covering covers $\mathcal{U}$. Let $\mathcal{C}$ be the optimal covering for $\mathcal{S}_x$ in this case. If for each $q_1 \in Q_1$ (similarly $q_2 \in Q_2$), there exists at most one $a_1 \in A_1$ ($a_2 \in A_2$) such that $S_{(q_1, a_1)} \in \mathcal{C}$ ($S_{(q_2, a_2)} \in \mathcal{C}$), then it can be argued that $\mathcal{C}$ is large. However this need not be true. To overcome this problem, we slightly alter the definition of MIP system. In this altered model, the 2 provers instead of returning an answer each to their respective queries return l answers and the verifier V accepts if any of the l^2 pairs of answers leads it to accept. It is easy to prove that if the soundness in the old model is ϵ then the soundness in the new model is at most ϵl^2. Working with this altered MIP, let us divide the random strings into 3 different classes as shown.

$$
\begin{aligned}
R \in R_{\mathrm{BAD}-1} &\iff |\{S_{(q_1,a_1)} : q[R,1] = q_1\}| > \frac{k}{2} \\
R \in R_{\mathrm{BAD}-2} &\iff |\{S_{(q_2,a_2)} : q[R,2] = q_2\}| > \frac{k}{2} \\
R \in R_{\mathrm{GOOD}} &\iff R \notin R_{\mathrm{BAD}-1} \cup R_{\mathrm{BAD}-2}
\end{aligned}
$$

From above it is clear that if $R \in R_{\mathrm{GOOD}}$, then for $q_1 = q[R,1]$ and $q_2 = q[R,2]$, there exist at most k subsets $S_{(q,a)} \in \mathcal{C}$ such that $q = q_1$ or $q = q_2$. It is to be noted that only these subsets can cover the $\mathcal{X}^{(R)}$ portion of the universe $\mathcal{U}$. Hence, we have that $\mathcal{X}^{(R)}$ is covered by at most k subsets. So this covering of $\mathcal{X}^{(R)}$ cannot be a bad covering. Hence, $\mathcal{X}^{(R)}$ has a good covering, say $S_{(q_1,a_1)}, S_{(q_2,a_2)}$. But this means that the answers (a_1, a_2) lead the verifier to accept. Hence,

$$\Pr[R \in R_{\mathrm{GOOD}}] \le \frac{1}{N}\left(\frac{k}{2}\right)^2 \le \frac{1}{2}(\text{say})$$

Or in other words

$$\Pr[R \in R_{\mathrm{BAD}-1} \cup R_{\mathrm{BAD}-2}] > \frac{1}{2}$$

Without loss of generality, we can assume that

$$\Pr[R \in R_{\mathrm{BAD}-1}] > \frac{1}{4}$$

In this case, due to the regularity conditions of the MIP and from the definition of $R_{\mathrm{BAD}-1}$, we get that,

$$|\mathcal{C}| > \frac{|Q_1|}{4}\frac{k}{2} = \frac{Qk}{8}$$

Setting $m = |A_1| = n^{\log n}$ as mentioned earlier and $k = c \log n$, we get that the size of the universe, $\mathcal{U}$, is of the order 2^k (say M). The following lemma then follows from the above discussion.

Lemma 3 *Soundness: If $x \notin L$, then $|\mathcal{C}| \ge (1-\epsilon)\log_4 M(2Q)$, for any $\epsilon > 0$.*

Lemmas 2 and 3 prove that approximating Set-Cover to within a factor of $(1-\epsilon)\log_4 n$ is NP-hard, unless $\mathbf{NP} \subseteq \mathbf{DTIME}(n^{\log n})$, since the construction of the gadget involves $O(n^{\log n})$ time. By a more careful analysis, it can be shown that the same result holds unless $\mathbf{NP} \subseteq \mathbf{DTIME}(n^{\log\log n})$.

It is to be noted that we have a positive result proving that there exists an algorithm for approximating Set-Cover to within a factor of $\ln n$. Note that there is a gap of $\ln n \leftrightarrow \log_4 n$ between the positive and negative results. In the next section, we shall give a rough sketch of Feige's improvement [F 98] of the above negative result to the factor $(1-\epsilon)\ln n$, for any $\epsilon > 0$.

4 Feige's Improvement [F 98]

As before, a (n, m, k)-gadget is constructed, however with a few variations. Here, instead of the universe, $\mathcal{X}$, being partitioned into two subsets, S_i and S_{m+i}, it is partitioned into C subsets as shown below.

$$
\begin{array}{ccc}
S_{11} & \dots & S_{1m} \\
S_{21} & \dots & S_{2m} \\
\vdots & \vdots & \vdots \\
S_{C1} & \dots & S_{Cm}
\end{array}
$$

such that $\bigcup_{i=1}^{C} S_{ij} = \mathcal{X}$, for all j. (ie., the subsets along a column partition the set $\mathcal{X}$) Here the C subsets in any column is called a good covering and a covering such that no two sets of the cover are from the same column is called a bad covering. Such a gadget is called a $(n, m, k)_C$-gadget if there does not exist any bad cover of size less than k. As before it can be shown that for any m, k, a $(n, m, k)_C$-gadget exists for some $n = \left(\frac{C}{C-1}\right)^k \text{poly}(m)$. With such a gadget, a gap of $\left(\log_{\frac{C}{C-1}} n\right)/C$ can be shown using a construction similar to above. Since $\left(\log_{\frac{C}{C-1}} n\right)/C \to \ln n$ as $C \to \infty$, this gives us a $(1-\epsilon)\ln n$ hardness result for Set-Cover. However, it is to be noted that for this construction we need a C-prover MIP system instead of the 2-prover system used earlier. Also, a different semantics is needed for the C-prover MIP system that provides a notion of consistency between the provers similar to the good-bad covers. In this manner an optimal hardness result for Set-Cover can be obtained unless $\mathbf{NP} \subseteq \mathbf{DTIME}(n^{\log n})$.

References

[F 98] Uriel Feige. A Threshold of $\ln n$ for Approximating Set Cover. *Journal of the ACM*, volume 45, pages 634-652, July 1998.

[LY 94] Carsten Lund and Mihalis Yannakakis. On the hardness of Approximating Minimization Problems. *Journal of the ACM*, volume 41, pages 960-981, September 1994.