Lecture 10 (08.05.2013)

Scribe: Marcin Wrochna
Lecturer: Marek Cygan

Hardness of approximation

To prove that some problem is hard to approximate we need some assumptions – this lecture will use theorems that follow from assuming $P \neq NP$, later we will use some stronger, less universally accepted conjectures.

1 PCP theorem

NP is the class of languages that admit a polynomially bounded verifier – for positive instances there is a proof (witness), which is accepted by the verifier, for negative instances no witness is accepted. For example in 3-SAT witnesses are (polynomially bounded strings interpreted as) valuations and the verifier simply checks the value of each clause.

Definition 1. $PCP_{c,s}(r(n), q(n))$ is the class of problems for which:

- there is a poly time verifier which, using $r(n)$ random bits and reading $q(n)$ bits from the proof, decides whether to accept an instance,
- when the instance is positive, accepts some proof with probability $\geq c$,
- when the instance is negative, accepts each of the proofs with probability $\leq s$.

For example $NP = PCP_{1,0}(0, \text{poly}(n))$ from the verifier-based definition.

The following is one version of the PCP theorem, proved in 1992, with a significantly simpler proof (but still outside the scope of this lecture) discovered by Dinur in 2005 [1].

Theorem 2. $NP = PCP_{1,1/2}(\mathcal{O}(\log n), \mathcal{O}(1))$.

A later, different version says

Theorem 3. *For each $\varepsilon, \delta > 0$, $NP = PCP_{1-\varepsilon,\frac{1}{2}+\delta}(\mathcal{O}(\log n), 3)$, where the verifier is allowed only to return the xor (or negated xor) of the proof bits it reads.*

The verifier for a given instance and a sequence of random bits chooses three bits to read and one function (`even` or `odd`) and then reads the three bits and applies the chosen function.

2 MAX-E3-SAT

MAX-E3-SAT is the problem of maximizing the number of satisfied clauses in a CNF-SAT instances with *exactly* three literals in each clause. A random assignment satisfies $\frac{7}{8}$ clauses and the method of conditional expectation gives an easy derandomized $\frac{7}{8}$-approximation. We follow with a typical application of the PCP theorem.

Proposition 4. *There is no $\frac{7}{8} + \epsilon$ approximation for MAX-E3-SAT (assuming $P \neq NP$).*

Proof. Let L be any language in NP– we will show how we could decide it in P using a $\frac{7}{8}+\epsilon$-approximation.

L, as a language in NP, has a $PCP_{1-\varepsilon',1/2+\delta}(\mathcal{O}(\log(n),3)$ verifier, for any ε'. We construct a variable for each bit of the proof. For each possible random string (that is polynomially many possibilities) we see what proof bits the verifier would read – let that be i,j,k. Assuming the verifier uses the function odd (the even case is analogous) we construct clauses $x_i \vee x_j \vee x_k$, $x_i \vee \bar{x_j} \vee \bar{x_k}$, $\bar{x_i} \vee x_j \vee \bar{x_k}$, $\bar{x_i} \vee \bar{x_j} \vee x_k$ – if an odd number of the variables is true, all four clauses will be true, otherwise exactly three will be true (because each clause is false in exactly one valuation and we have chosen those that describe negations of even valuations). This ends the construction.

Thus when $x \in L$, the verifier accepts some proof with probability $\geq 1-\varepsilon'$, which implies there is a valuation of the constructed E3-SAT instance with at least $(1-\varepsilon')\frac{4}{4}+\varepsilon'\frac{3}{4} = 1-\frac{\varepsilon'}{4}$ satisfied clauses. When $x \notin L$, any valuation can satisfy at most $(\frac{1}{2}+\delta)\frac{4}{4}+(\frac{1}{2}-\delta)\frac{3}{4} = \frac{7}{8}+\frac{\delta}{4}$ clauses. Choosing ε', δ to be much smaller than ϵ, we can distinguish between these two cases with an $\frac{7}{8}+\epsilon$-approximation. □

Notice we actually proved a slightly stronger statement: we cannot tell the difference between instances with $\geq 1-\varepsilon$ satisfiable clauses and those with $\leq \frac{7}{8}+\delta$ satisfiable clauses. This stronger version is more useful for later reductions.

Another stronger version of the PCP theorem allows the make the same statement with 1 instead of $1-\varepsilon$.

3 Independent Set

During the exercises we will prove the following lemma with a reduction from MAX-E3-SAT:

Lemma 5. *There is an $\alpha > 1$ such that Independent Set is not α-approximable (that is, Independent Set is APX-hard – there are stronger results known).*

We now use this lemma to prove no constant α-approximability, with a gap amplification lemma.

Lemma 6. *If there is an α-approximation for IS, then there is a $\sqrt{\alpha}$-approximation for IS.*

Proof. From a given instance $G=(V,E)$ we simply construct the graph $G'=(V',E')$ with $V' = V \times V$ and $E' = \{(a,b)(c,d) : ac \in E \vee bd \in E\}$ (yes, that's a lot of edges).

We show that $OPT(G') = OPT(G)^2$, from which the lemma immediately follows. If S is an independent set in G, then $S \times S$ is easily seen to be an independent in G'.

Conversely, if S' is an independent set in G', then define the projections
$S_1 = \{u \in V : \exists_v (u,v) \in S'\}$, $S_2 = \{v \in V : \exists_u (u,v) \in S'\}$. Now S' is a subset of $S_1 \times S_2$ and it's independence implies the independence of S_1 and S_2. $|S'| \leq |S_1| \cdot |S_2|$, so one of S_1, S_2 must be at least $\sqrt{|S'|}$ in size. □

Nowadays we know we cannot even make the difference between instances with a maximum independent set of size $n^{1-\varepsilon}$ and instances with at an IS of size at most n^{ε} (Håstad [3] assuming $NP \neq ZPP$, then Zuckerman [5] assuming only $P \neq NP$).

4 Label Cover

LABEL COVER
Input: a bipartite graph $G = (X \uplus Y, E)$, a set of labels $[m]$ and
for each edge xy a constraint $\psi_{xy} : [m] \to [m]$.
Question: Give an assignment of labels to vertices $f : X \cup Y \to [m]$ maximizing the number of satisfied edges (that is edges xy having $\psi_{xy}(f(x)) = f(y)$).

A simple reduction from MAX-E3-SAT shows the APX-hardness of LABEL COVER, and similarly as for INDEPENDENT SET one can use (more complicated) gap amplification to achieve the following theorem (which uses a stronger assumption, because we need to create $\log\log n$-tuples instead of pairs like in IS).

Theorem 7. *Assuming $NP \not\subseteq DTIME(n^{\mathcal{O}(\log\log n)})$, for all $c > 0$, we cannot distinguish between fully satisfiable and $\leq \frac{1}{\log^c n}$-satisfiable instances of* LABEL COVER*, even in regular graphs.*

5 Set Cover

SET COVER (unweighted)
Input: a universe U of n elements and a family of subsets $\mathcal{F} = \{F_1, \ldots, F_m\}$, $F_i \subseteq U$ satisfying $\bigcup_i F_i = U$.
Question: Give a minimum set of indices $I \subseteq [m]$ satisfying $\bigcup_{i \in I} F_i = U$.

5.1 Approximation algorithms

There is a $\ln n$-approximation of SET COVER. The algorithm greedily selects the set that covers the most uncovered elements. If we assign each set a cost of 1 and divide it equally among its elements, the i-th covered element costs us $\leq OPT \cdot \frac{1}{n+1-i}$ – indeed, the optimal solution gives a set that covers at least $\frac{n}{OPT}$ elements (and then there is a set which covers $\frac{n-i+1}{OPT}$ elements, and so on). So the sum of costs in the greedy algorithm is bounded by $H_n \cdot OPT$, where H_n is the n-th harmonic number, asymptotically $\ln n + \mathcal{O}(1)$.

Another way to obtain an $\mathcal{O}(\log n)$-approximation is by LP-rounding. Take the linear program

$$\text{minimize} \sum_{s \in [m]} x_s$$

$$\sum_{\substack{F_s \in \mathcal{F} \\ v \in F_s}} x_s \geq 1 \qquad \forall\ v \in U$$

take each set $F_s \in \mathcal{F}$ with probability x_s, and repeat $2\ln n$ times. The expected number of sets taken is, by the linearity of expected value, $2\ln n \cdot \sum x_s \leq 2\ln n \cdot OPT$. Each element u of U is left uncovered in one step with probability $\prod_{F_s \ni u}(1 - x_s) \leq \frac{1}{e}$ (from $\sum x_s \geq 1$ and Jensen's inequality). So each element is left uncovered in all steps with probability at most $\frac{1}{e}^{2\ln n} = n^{-2}$ and by the union bound, there is an uncovered element with probability at most $\frac{1}{n}$.

5.2 Hardness of approximation

Theorem 8 (Feige [2]). *Assuming $NP \not\subseteq DTIME(n^{\mathcal{O}(\log\log n)})$, there is no $(1-\varepsilon)\ln n$-approximation for* Set Cover *for any $\varepsilon > 0$.*

We'll prove only a weaker statement (the above for some ε, namely $\frac{15}{16}$):

Theorem 9 (Lund, Yannakakis [4]). *Assuming $NP \not\subseteq DTIME(n^{\mathcal{O}(\log\log n)})$, there is no $\frac{\ln n}{16}$-approximation for* Set Cover.

We reduce from Label Cover. The following gadget is crucial to the proof.

Definition 10. An (m, l)-system consists of a universe B and a collection of m subsets of that universe $\{C_1, \ldots, C_m\}$, such that if a choice of l sets of the form C_i or $\overline{C_i}$ covers all the universe, then this choice must contain both C_i and $\overline{C_i}$ for some i.

Theorem 11. *For all m, l there is an (m, l)-system of size $|B| = \mathcal{O}(2^{2l}m^2)$. Such a system can be constructed in time polynomial in $|B|$.*

There is an easy randomized construction (which requires us to change our assumption from $DTIME$ to $ZTIME$ (expected time, instead of deterministic, but results with probability 1)) – simply take a universe of size $|B| = 2^l \ln(2m^l)$ and random sets C_i (each element with probability $\frac{1}{2}$), the probability of a failure (all elements covered by only l sets) is $(1 - \frac{1}{2^l})^{|B|}$. It could be derandomized with *universal functions*.

We are now ready to reduce an instance $(G = (X, Y, E), [m], \Psi)$ of Label Cover. Let $(B, C_1, \ldots, C_m)$ be an (m, l)-system and let the universe be $U = E \times B$. We add to the family $\mathcal{F}$:

- for all $x \in X$ and $i \in [m]$, the set $S_{x,i} = \sum_{xy \in E} xy \times C_{\Psi_{xy}(i)}$,
- for all $y \in Y$ and $i \in [m]$, the set $S_{y,i} = \sum_{xy \in E} xy \times \overline{C_i}$.

Lemma 12. *If the instance of* Label Cover *was fully satisfiable, then the instance of* Set Cover *has a solution of size $\leq |X| + |Y|$.*

Proof. Let $f : X \cup Y \to [m]$ be an assignment satisfying all edge constraints of Label Cover. To cover the elements of $U = E \times B$ we choose the sets $\{S_{v,f(v)} : v \in X \cup Y\}$. This gives us for each edge xy the set $S_{x,f(x)}$ containing $\{xy\} \times C_{\Psi_{xy}(f(x))}$ and the set $S_{y,f(y)}$ containing $\{xy\} \times \overline{C_{f(y)}}$. Because $\Psi_{xy}(f(x)) = f(y)$, this gives a set and it's complement on each edge. □

Lemma 13. *If there is a set cover $\mathcal{S}$ of size $\leq \frac{l}{8}(|X| + |Y|)$, then there is a label assignment satisfying a fraction of $\geq \frac{2}{l^2}$ of all edges.*

Proof. Given such a cover $\mathcal{S}$, we define for each vertex v the set of potential labels

$$L_v = \{i : i \in [m], S_{v,i} \in \mathcal{S}\}.$$

Each set contributes one label to one vertex, thus $\sum_{v \in X \cup Y} |L_v| = |\mathcal{S}|$. We say a vertex i *bad* when $|L_v| > \frac{l}{2}$. At most $\frac{1}{4}$ of all vertices can be bad (because $|\mathcal{S}| \leq \frac{1}{4} \cdot \frac{l}{2}(|X| + |Y|)$). Throw away edges incident to *bad* vertices – vertices with $|L_v| > \frac{l}{2}$. This is the place where we use the assumption that the graph is regular – observe that we throw away at most $\frac{1}{2}$ of all edges because at most

$n/4$ vertices are bad, which touch at most $dn/4$ endpoints of edges, hence at most $dn/4 = |E|/2$ edges are thrown away. We thus leave at least $\frac{|E|}{2}$ edges, all of them have both endpoints satisfying $|L_v| \leq \frac{l}{2}$.

Randomly select each $f(v)$ from the set L_v (no label, if it's empty). We want to show that f satisfies any edge xy with probability at least $\frac{4}{l^2}$. The subset $\{xy\} \times B$ of the universe could have been covered only by sets $S_{x,*}$ and $S_{y,*}$. But since $|L_x|, |L_y| \leq \frac{l}{2}$, we took at most l such sets. By the definition of an (m, l)-system, we must have taken a set C_i and it's complement $\overline{C_i}$ into $\mathcal{S}$, for some i. This means there are labels $i_x \in L_x, i_y \in L_y$ satisfying $\Psi_{xy}(i_x) = i_y$ – our random f will select both these labels with probability at least $\frac{4}{l^2}$. The expected fraction of satisfied edges is then at least $\frac{2}{l^2}$, so there is an f that satisfies at least that much. □

It now remains to connect those lemmas with the inapproximability of Label Cover (Theorem 7 with $c = 3$). We reduce $\text{GapLabelCover}_{1,1/\log^3 n}$ (the problem of distinguishing between a fully satisfiable instance and a $\leq \frac{1}{\log^3 n}$-satisfiable instance) with regular graphs to Set Cover using the above reduction with $l = \beta \log n$ (for some $\beta > 0$, chosen later). The universe constructed has size $N = |U| = |E| \cdot |B| = n^{\mathcal{O}(1)} 2^{2l}$, which is polynomial in n.

Suppose we have an $\alpha \ln N$-approximation for Set Cover, for $\alpha \leq \frac{1}{16}$. For appropriate β, $\alpha \ln N = \alpha \cdot (2l \ln 2 + \mathcal{O}(\ln n)) \leq \alpha 2l \leq \frac{l}{8}$, so if we get a fully satisfiable instance of Label Cover, the construction gives a Set Cover instance with a solution $\leq |X| + |Y|$ and the approximation would give a result of at most $\frac{l}{8}(|X| + |Y|)$.

On the other hand, if we get a $\leq 1/\log^3 n$-satisfiable instance of Label Cover, it means no label assignment can satisfy $\geq \frac{2}{l^2} = \frac{2}{\beta^2 \ln^2 n} > 1/\log^3 n$ (for large enough n) edges, so by the previous lemma no set cover of size $\leq \frac{l}{8}(|X| + |Y|)$ exists – thus we can distinguish between the two cases and solve $\text{GapLabelCover}_{1,1/\log^3 n}$, a contradiction.

References

[1] Irit Dinur. The pcp theorem by gap amplification. *J. ACM*, 54(3):12, 2007.

[2] Uriel Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[3] Johan Håstad. Clique is hard to approximate within n$^{\text{1-epsilon}}$. In *FOCS*, pages 627–636, 1996.

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

[5] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007.