Probabilistic analysis of the greedy algorithm[1]

N. N. Kuzjurin

Abstract. It is shown that the greedy algorithm in the average case (in some probabilistic model) finds almost minimum covers. It is shown also that in the average case the ratio of the size of minimum cover to the size of minimum fractional cover has logarithmic order in the size of the ground set.

1. Introduction

Set cover is one of the oldest and most studied NP-hard problems [7, 6, 8, 2, 4]. Given a ground set U of m elements, the goal is to cover U with the smallest possible number of subsets from a given family $S = \{S_1, \ldots, S_n\}$, $S_i \subseteq U$. A cover is arbitrary subfamily $S(I)$, $I \subseteq [n]$ such that

$$U = \cup_{i \in I} S_i,$$

where $[n] = \{1, 2, \ldots, n\}$.

The value $|I|$ is called the size of a cover. A cover of the smallest size is called minimum cover. The size of minimum cover is denoted by $C(S)$.

One of the best polynomial time algorithms for approximating set cover is the greedy algorithm: at each step choose the unused set from the family S which covers the largest number of remaining elements.

R is called an approximation ratio of an algorithm A if for all input data S the following holds

$$\frac{C_A(S)}{C(S)} \leq R,$$

where $C_A(S)$ denotes the size of a cover obtained by the algorithm A.

Lovasz [8] and Johnson [6] showed that the approximation ratio of the greedy algorithm is no worse than $H(m)$, where $H(m) = 1 + 1/2 + \ldots + 1/m$ is the mth harmonic number, a value which is clearly between $\ln m$ and $1 + \ln m$. Similar results were obtained in [9, 10]. These results were improved slightly by Slavik [11] who showed that the approximation ratio of the greedy algorithm is exactly $\ln m - \ln \ln m + \Theta(1)$. Feige [3] proved that for any $\varepsilon > 0$ no polynomial time algorithm can approximate set cover within $(1 - \varepsilon) \ln m$ unless $NP \subseteq DTIME[n^{O(\log \log n)}]$.

[1]Supported by RFBR, grants 02-01-00713 and 04-01-00359.

It is well-known that set cover forms an important class of integer programs

$$\min \mathbf{cx} |\ A\mathbf{x} \geq \mathbf{b}, \quad \mathbf{x} \in \{0,1\}^n, \tag{1}$$

where $\mathbf{c} = (1, ..., 1)$, $\mathbf{b} = (1, ..., 1)^T$ and $A = (a_{ij})$ is an arbitrary $m \times n$ $(0,1)$-matrix.

To see this it is sufficient to chose $(0,1)$-matrix $A = (a_{ij})$ such that $a_{ij} = 1$ iff $u_i \in S_j$, where $U = \{u_1, \ldots, u_m\}$. In such a way we correspond covering of (0,1)-matrix to covering by a family of subsets. The size of minimum cover we will denote by $C(A)$ as well.

In particular, it is known (see, [10]) that for any (0,1)-matrix of size $m \times n$ with at least k 1's in each row the size of the minimum cover $C(A)$ satisfies

$$C(A) \leq 1 + \frac{\ln \frac{mk}{n}}{\ln \frac{n}{n-k}}. \tag{2}$$

In fact, it is known that the size of any cover obtained by the greedy algorithm satisfies (2).

But all these investigations were related to the worst case performance of the greedy algorithm. In this paper we consider the average case and show that the asymptotic approximation ratio of the greedy algorithm in the average case is at most $1 + \varepsilon$ for arbitrary constant $\varepsilon > 0$. It is shown also that the ratio of the size of minimum cover to the size of the fractional cover is approximately $\ln mp$ in the average case.

2. Average case analysis of the greedy algorithm

In this section we consider a probabilistic model in which $A = (a_{ij})$ is a random (0,1)-matrix such that $\mathbf{P}\{a_{ij} = 1\} = p$ and $\mathbf{P}\{a_{ij} = 0\} = 1 - p$ independently for all i, j. Then, the value $C(A)$ becomes a random variable.

Lemma 1 [1]. Let Y be a sum of n independent random variables each taking the value 1 with probability p and 0 with probability $1 - p$. Then

$$\mathbf{P}\{|Y - np| > \delta np\} \leq 2 \exp\{-(\delta^2/3)np\}.$$

Let $L_0 = -\frac{\ln mp}{\ln(1-p)}$.

Theorem 1. Let the probability p be such that $0 < p < c < 1$, where c is a constant. Let

$$\frac{\ln \ln n}{\ln mp} \to 0 \text{ as } n \to \infty, \tag{3}$$

$$\frac{\ln m}{np} \to 0 \text{ as } n \to \infty. \tag{4}$$

Then for any fixed $\varepsilon > 0$

$$\mathbf{P}\{(1-\varepsilon)L_0 \le C(A) \le (1+\varepsilon)L_0\} \to 1 \qquad (5)$$

as $n \to \infty$.

Corollary 1. Let $m = cn$, where c is some constant and p be a constant. Consider the problem (1) with a random A defined above. Then (4) holds.

Proof of Theorem 1. *Lower bound.* Let $X(l)$ be the random variable equal to the number of covers in A of size l. We have

$$\mathbf{E}X(l) = \binom{n}{l} P(l),$$

where $P(l)$ is the probability that fixed l columns form a cover in A. It is not difficult to see that

$$P(l) = \left(1-(1-p)^l\right)^m \le \exp\{-m(1-p)^l\}.$$

Thus, using the inequality $\binom{n}{k} \le n^k$, we have

$$\ln \mathbf{E}X(l) \le l \ln n - m(1-p)^l.$$

Taking $l = l_0 = -\lceil (1-\delta)\ln mp/\ln(1-p)\rceil$ we get

$$\begin{aligned}
\ln \mathbf{E}X(l_0) &\le -\ln mp/\ln(1-p)\ln n - -m\exp\{-(1-\delta)\ln(1-p)\frac{\ln mp}{\ln(1-p)}\} \\
&\le -(\ln mp/\ln(1-p))\ln n - mm^{-1}m^{\delta}p^{-1+\delta} \\
&= -(\ln mp/\ln(1-p))\ln n - (mp)^{\delta}\frac{1}{p}.
\end{aligned}$$

Considering two cases (p is a constant, and $p \to 0$) it is not difficult to see that for any fixed $0 < \delta < 1$ under the condition (3) the last expression tends to $-\infty$ as n tends to infinity.

Thus, the probability that there are no covers of size l_0 in a random $(0,1)$-matrix A tends to 1, because

$$\mathbf{P}\{X(l_0) \ge 1\} \le \mathbf{E}X(l_0) \to 0.$$

Clearly, if there are no covers of size l_0 in A then there are no covers of size smaller than l_0 as well. Therefore,

$$P\{C(A) \ge l_0\} \to 1.$$

Upper bound. We use the upper bound (2).

By Lemma 1, in a random (0,1)-matrix for any $\delta > 0$ with probability tending to 1 each row contains k 1's where $(1+\delta)pn \ge k \ge (1-\delta)pn$. Indeed, Lemma 1 implies that the probability that some fixed column contains k 1's with $(1+\delta)pn \ge k \ge (1-\delta)pn$ is

$$P_{bad} \le 2\exp\{-(\delta^2/3)np\},$$

and the expectation of the number of such rows is at most mP_{bad}. It is not difficult to see that

$$mP_{bad} \le 2\exp\{\ln m - (\delta^2/3)np\} = 2\exp\{\ln m - O(np)\} \to 0, \text{ as } n \to \infty,$$

by the condition $(\ln m)/np \to 0$ as $n \to \infty$. Hence, Markov's inequality $P\{X \ge 1\} \le \mathbf{E}X$ implies that the probability of the event 'each row contains k 1's where $(1+\delta)pn \ge k \ge (1-\delta)pn$ tends to 1.

Thus, we obtain

$$C(A) \le \frac{\ln \frac{mk}{n}}{\ln \frac{n}{n-k}} \le \frac{\ln(mp(1+\delta))}{\ln \frac{n}{n-np(1-\delta)}} \le -\frac{\ln(mp(1+\delta))}{\ln(1-p(1-\delta))}.$$

Simplifying we get

$$-\frac{\ln(mp(1+\delta))}{\ln(1-p(1-\delta))} = -\frac{\ln(mp)+\ln(1+\delta)}{\ln(1-p(1-\delta))}.$$

For any constant $\varepsilon > 0$ there exists a constant $\delta > 0$ such that the latter expression is at most

$$-(1+\varepsilon)\frac{\ln(mp)}{\ln(1-p)}.$$

Combining the inequality with the lower bound of $C(A)$ we arrive at the desired inequality. The proof of Theorem 1 is complete.

We can reformulate our result in other words. Let us define an asymptotic approximation ratio of an algorithm as the limit of approximation ratio when n goes to infinity. Then Theorem 1 gives the conditions guaranteeing the asymptotic approximation ratio of the greedy algorithm is equal to 1 in the average case.

3. Integral and fractional covers

In this section we consider the same probabilistic model in which $A = (a_{ij})$ is a random (0,1)-matrix such that $\mathbf{P}\{a_{ij} = 1\} = p$ and $\mathbf{P}\{a_{ij} = 0\} = 1-p$ independently for all $i,\ j$. In the previous section we have estimated the value of

$C(A)$ for almost all matrices. In this section we find the value of the optimum of the linear relaxation of (1).

Recall that the linear relaxation of (1) is the same program where the restriction $\mathbf{x} \in \{0,1\}^n$ is replaced by $0 \le x_j \le 1$, $j = 1, \dots, n$. We denote the optimum value of the linear relaxation by $q(A)$.

Theorem 2. Let

$$\frac{\ln m}{np} \to 0, \qquad \frac{\ln n}{mp} \to 0 \text{ as } n \to \infty.$$

Then for any fixed $\varepsilon > 0$

$$\mathbf{P}\{(1-\varepsilon)/p \le q(A) \le (1+\varepsilon)/p\} \to 1$$

as $n \to \infty$.

Proof. We have already shown that the condition $(\ln m)/np \to 0$ as $n \to \infty$ implies that the probability of the event "each row contains k 1's" where $(1+\delta)pn \ge k \ge (1-\delta)pn$ tends to 1.

Similarly we can show that the condition $(\ln m)/np \to 0$ as $n \to \infty$ implies that the probability of the event "each column contains t 1's" where $(1+\delta)pm \ge t \ge (1-\delta)pm$ tends to 1.

Proof of Claim 1. Let $\mathbf{x} = (x_1, \dots, x_n)^T$ be an optimal solution of the linear relaxation of (1). We have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_j \ge \sum_{i=1}^{m} 1 = m.$$

On the other hand,

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_j = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}x_j = \sum_{j=1}^{n} x_j \sum_{i=1}^{m} a_{ij} \le \sum_{j=1}^{n} x_j(1+\delta)pm = q(1+\delta)mp.$$

Therefore, $q \ge ((1+\delta)p)^{-1}$. Furthermore,

$$\mathbf{x} = (\frac{1}{(1-\delta)np}, \dots, \frac{1}{(1-\delta)np})$$

is a feasible solution to the linear relaxation of (1) because

$$\sum_{j=1}^{n} a_{ij}x_j = \sum_{j=1}^{n} a_{ij}\frac{1}{(1-\delta)np} = \frac{1}{(1-\delta)np}\sum_{j=1}^{n} a_{ij} \ge$$

$$(\frac{1}{(1-\delta)np}) \cdot (1-\delta)pn = 1.$$

This implies $\sum_{j=1}^{n} \frac{1}{(1-\delta)np} \ge q$, that is, $1/(1-\delta)p \ge q$. We have

$$((1+\delta)p)^{-1} \le q \le ((1-\delta)p)^{-1}.$$

This implies the assertion of Theorem 2.

Corollary 2. Let all the conditions of Theorem 1 and 2 hold. Then for any fixed $\varepsilon > 0$

$$\mathbf{P}\{(1-\varepsilon)\ln mp \le \frac{C(A)}{q(A)} \le (1+\varepsilon)\ln mp\} \to 1$$

as $n \to \infty$.

4. Average case analysis: towards the general case

It seems interesting to extend our technique to the general distribution where $P\{a_{ij} = 1\} = p_{ij}$. The main ingredient of this technique was obtaining lower bounds for the size of minimum cover of random matrices.

In this section we do the first step towards this goal. We consider a probabilistic model in which $A = (a_{ij})$ is a random (0,1)-matrix such that $\mathbf{P}\{a_{ij} = 1\} = p_i$ and $\mathbf{P}\{a_{ij} = 0\} = 1 - p_i$ independently for all i, j. The difference between this model and the one from the previous section is that we allow here different probabilities for different rows.

Let

$$\overline{p} = m^{-1} \cdot \sum_{i=1}^{m} p_i, \qquad p_{max} = \max_i p_i.$$

Theorem 3. Let $p_{max} \to 0$ as $n \to \infty$, and

$$\frac{\ln \ln n}{\ln(m\overline{p})} \to 0 \text{ as } n \to \infty.$$

Then for any fixed $\varepsilon > 0$

$$\mathbf{P}\{(1-\varepsilon)\frac{\ln(m\overline{p})}{\overline{p}} \le C(A)\} \to 1$$

as $n \to \infty$.

Proof of Theorem 3. Let $X(L)$ be the random variable equal to the number of covers in A of size L.

$$\mathbf{E}X(L) = \binom{n}{L} P(L),$$

where $P(L)$ is the probability that arbitrary fixed L columns form a cover in A. Using the inequality $1 - x < e^{-x}$, it is not difficult to see that

$$P(L) = \prod_{i=1}^{m} \left(1 - (1-p_i)^L\right) \le \exp\{-\sum_{i=1}^{m}(1-p_i)^L\}.$$

Thus, taking into account that $\binom{n}{k} \le n^k$, we get

$$\ln \mathbf{E}X(L) \le L \ln n - m \sum_{i=1}^{m} (1-p_i)^L.$$

Using the fact that the arithmetic mean is always at least as the geometric mean we can estimate the sum as follows:

$$\sum_{i=1}^{m} (1-p_i)^L = m \left(\frac{1}{m} \sum_{i=1}^{m} (1-p_i)^L \right) \ge \left(\prod_{i=1}^{m} (1-p_i)^L \right)^{1/m} = \prod_{i=1}^{m} (1-p_i)^{L/m}.$$

Taking this into account we get

$$\ln \mathbf{E}X(L) \le L \ln n - m \prod_{i=1}^{m} (1-p_i)^{L/m}.$$

Using the inequality

$$1 - x > e^{-\frac{x}{1-x}}, \qquad 0 < x < 1,$$

we have

$$\ln \mathbf{E}X(L) \le L \ln n - m \exp\{-\sum_{i=1}^{m} \frac{L}{m} \frac{p_i}{1-p_i}\} \le L \ln n - m \exp\{-L\overline{p}(1+o(1))\}.$$

Let

$$L_1 = (1-\varepsilon) \frac{\ln(m\overline{p})}{\overline{p}}.$$

The inequality above implies

$$\begin{aligned} \ln \mathbf{E}X(L_1) &\le \frac{\ln(m\overline{p})}{\overline{p}} \ln n - m \exp\{-(1-\varepsilon) \frac{\ln(m\overline{p})}{\overline{p}} \overline{p}(1+o(1))\} \\ &\le \frac{\ln m}{\overline{p}} \ln n - m \exp\{-(1-\varepsilon) \ln(m\overline{p})(1+o(1))\} \\ &= \frac{1}{\overline{p}} \ln m \ln n - m\overline{p}^{-(1-\varepsilon)(1+o(1))} m^{-(1-\varepsilon)(1+o(1))} \\ &= \frac{1}{\overline{p}} \left(\ln m \ln n - m(m\overline{p})^{\varepsilon(1+o(1))-o(1)} \right). \end{aligned}$$

For any fixed $\varepsilon > 0$ this expression tends to $-\infty$ when n goes to infinity in view of the conditions of Theorem 3.

Thus, the probability that there are no covers of size L_1 in a random $(0,1)$–matrix A tends to 1, because by Markov's inequality

$$\mathbf{P}\{X(L_1) \ge 1\} \le \mathbf{E}X(L_1) \to 0.$$

Clearly, if there are no covers of size L_1 in A then there are also no covers of size smaller than L_1. Therefore, with probability tending to 1

$$C(A) \ge L_1 = (1-\varepsilon) \frac{\ln(m\overline{p})}{\overline{p}}.$$

References

[1] N. Alon and J.H. Spencer, The Probabilistic Method, Wiley, 1992.

[2] V. Chvatal, A greedy heuristic for the set-covering problem, Mathematics of Operations Research, **4** (1979) 233–235.

[3] U. Feige, A threshold of $\ln n$ for the approximating set cover, Proceedings of the ACM Symposium on Theory of Computing, 1996, pp. 314-318.

[4] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-completeness, Freeman, New York, 1979.

[5] G.P. Gavrilov, A.A. Sapozhenko, Problems and exercises in discrete mathematics, Kluwer Texts in Math. Sci., v.14, Kluwer Academic Publishers, 1996.

[6] D.S. Johnson, Approximation algorithms for combinatorial problems, J. Comput. System Sci., **9** (1974) 256-278.

[7] R.M. Karp, Reducibility among combinatorial problems, in, Complexity of Computer Computations (R.E. Miller and J.W. Tatcher, Eds.), Plenum, New York, 1972, 85–103.

[8] L. Lovasz, On the ratio of optimal integral and fractional covers, Discrete Math. **13** (1975) 383-390.

[9] R.G. Nigmatullin, An algorithm of steepest descent in the set cover problem (Russian), Proceedings of Symposium on approximation algorithms. Kiev, May 17-22, 1969, p. 36.

[10] A.A. Sapozhenko, On the size of disjunctive normal forms obtained by the gradient algorithm (Russian), Discrete analysis. Novosibirsk, 1972, N 5, p. 111-116.

[11] P. Slavik, A tight analysis of the greedy algorithm for set cover, J. Algorithms, **25** (1997) 237-254.