Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 105 (2007)

A Note on Set Cover Inapproximability Independent of Universe Size

Jelani Nelson*
MIT CSAIL
minilek@mit.edu

Abstract

In the set cover problem we are given a collection of m sets whose union covers $[n] = \{1, \ldots, n\}$ and must find a minimum-sized subcollection whose union still covers $[n]$. We investigate the approximability of set cover by an approximation ratio that depends only on m and observe that, for any constant $c < 1/2$, set cover cannot be approximated to within $O(2^{\log^{1-1/(\log\log m)^c} m})$ unless SAT can be decided in slightly subexponential time. The main ingredients in the observation are the $\Omega(\log n)$ hardness of approximation proof of Lund and Yannakakis and a hardness result for label cover due to Dinur and Safra.

1 Introduction

Set cover is one of the oldest known NP-complete problems, being listed as one of Karp's "original 21 NP-complete problems" [7]. In the set cover problem we are given a collection of m sets whose union covers $[n] = \{1, \ldots, n\}$ and must find a subcollection of minimum size still covering $[n]$. Its approximability in terms of n is well-understood. It is known that a simple greedy algorithm [6, 12] and a randomized rounding scheme of an LP relaxation [8, 13] both achieve an approximation ratio of $(1 - o(1)) \ln n$. On the negative side, it is known that set cover cannot be approximated to within $c \ln n$ for some constant $c < 1$ unless $\mathrm{P} = \mathrm{NP}$ [11], or to within better than $(1 - o(1)) \ln n$ unless $\mathrm{NP} \subseteq \mathrm{TIME}(n^{O(\log\log n)})$ [4].

This work makes a first attempt at understanding the (in)approximability of set cover to within an approximation ratio that only depends on m, independent of n. The only known result concerning approximability of set cover in terms of m that the author could find is the upper bound of [2], which gives a non-trivial improvement over $O(m)$-approximation only when the VC-dimension of the set system is sufficiently smaller than $\lg m - \lg\lg m$.[1] The original hardness proof for set cover by Lund and Yannakakis [1, 9] gives a reduction where n and m are polynomially related, as do subsequent proofs. Such results thus imply $\Omega(\log m)$-hardness of approximation for set cover. The question then becomes whether polynomial-time $O(\log m)$-approximation is possible regardless of the relationship between m and n. An instance with $m = O(\log n)$ can be solved exactly in polynomial time by brute force since set cover can be solved in time $O(\mathrm{poly}(n) \cdot 2^{O(m)})$, but what about m slightly superlogarithmic? By combining the proof of Lund and Yannakakis [9] with a result of Dinur and Safra [3], we observe that set cover cannot be approximated to within $2^{\log^{1-\delta_c(m)} m}$ in polynomial time for any constant $c < 1/2$ unless SAT can be decided in time $2^{O(2^{\log^{1-\delta_c(n)} n})}$, where $\delta_c(n) = 1/(\log\log n)^c$.

*Supported by an NDSEG fellowship.
[1]The abstract of [2] has a slight bug; the VC-dimension "d" they mention actually refers to the VC-dimension of the dual set system.

ISSN 1433-8092

2 A Motivating Example

Here we give an example of why one may want an approximation algorithm for set cover whose approximation ratio depends only on m. Consider the following *taxonomy labeling* problem introduced by Rabani, Schulman, and Swamy [10]. In this problem there is some finite alphabet Σ and a tree with n nodes. Each node is labeled with a string in $(\Sigma \cup \{0,1\}^*)^m$, and we must replace all occurrences of "*" in all labels with elements of Σ so as to minimize the maximum Hamming distance of labels of adjacent nodes. The authors show a reduction from taxonomy labeling to what they call the *multicut packing* problem on trees. In the multicut packing problem on trees we are given a tree on n nodes and a set of m multicut instances $M_i = \{(s_1^i, t_1^i), \ldots, (s_{r_i}^i, t_{r_i}^i)\}$. For each i we must output a multicut, i.e. a set of edges whose removal disconnects s_i from t_i for all i, and the objective is to minimize the maximum number of times any edge is used in a multicut. Rabani, Schulman, and Swamy obtain a $O(\log^2 m)$-approximation for multicut packing on trees, independent of n.

Now one may observe that multicut packing on trees is actually a special case of the following generalization of set cover. We are given N collections $\mathcal{C}_i = \{S_1^i, \ldots, S_{m_i}^i\}$ of subsets of $[n]$. We must choose a subcollection $\mathcal{C}_i'$ from each $\mathcal{C}_i$ so that $\bigcup_i \bigcup_{S \in \mathcal{C}_i'} S = [n]$, and the objective is to minimize $\max_i |\mathcal{C}_i'|$. In the case of multicut packing on trees, for each edge e we have a collection $\mathcal{C}_e$. The universe to be covered consists of all commodities in all multicut instances. Each m_i equals m, the number of multicut instances, and S_i^e is the set of commodities $(s,t) \in M_i$ such that e is on the unique path from s to t, i.e. the removal of edge e cuts (s,t) (recall the graph is a tree). As the approximation ratio for multicut packing on trees obtained in [10] is $O(\log^2 m)$, one may wonder whether such a result could be extended to all instances of this generalized set cover problem. Our observation implies that such a result is impossible unless SAT has subexponential time algorithms since $O(2^{\log^{1-\delta_c(m)} m})$-approximation is hard even when $N = 1$ (the usual set cover problem).

3 The Main Observation

Definition 1. LabelCover(c, s) *is the promise problem where we are given a bipartite graph that is both left-regular and right-regular with bipartition $V = V_1 \cup V_2$ ($|V| = n$), edge set E, label set $[L] = \{1, \ldots, L\}$, and a set of functions $f_e : [L] \to [L]$ indexed by edge (there is exactly one such function per edge in E). A labelling is a function $\ell : V \to [L]$, and an edge $e = (v_1, v_2)$ is said to be satisfied by ℓ if $\ell(v_2) = f_e(\ell(v_1))$. In the promise problem we are given an instance where either a labelling exists satisfying at least $c|E|$ edges, or no labelling satisfies more than $s|E|$ edges. We must decide which case holds.*

Theorem 2 ([3])**.** *For any constant $c < 1/2$ deciding LabelCover$(1, 2^{-\log^{1-\delta_c(n)} n})$ with a polynomial-size alphabet is NP-hard, where $\delta_c(n) = 1/(\log\log n)^c$.* □

The work of [3] actually defines LabelCover differently. In their definition one must assign a *set* of labels $\ell(v)$ to each vertex $v \in V$ so as to satisfy *all* edges. In this scenario an edge $e = (v_1, v_2)$, where $v_i \in V_i$, is said to be satisfied when for each label $\ell_2 \in \ell(v_2)$ there is a label $\ell_1 \in \ell(v_1)$ such that $f_e(\ell_1) = \ell_2$. The goal is then to minimize the l_p norm of the vector $(|l(v)|)_{v \in V}$. The work of [3] shows that polynomial-time $2^{\log^{1-\delta_c(n)} n}$-approximation is NP-hard for any $1 \le p \le \infty$, which implies Theorem 2 by using a known relationship [1] between the version of LabelCover defined in [3] with $p = 1$ to the version of LabelCover in Definition 1.

Theorem 3 ([1, 9])**.** *Suppose it is NP-hard to decide LabelCover$(1, \epsilon)$ with a label set of size $L = O(f(n))$. Then for any ℓ such that $2/\ell^2 < \epsilon$, set cover has no polynomial-time $O(\ell)$-approximation unless SAT can be decided in time $(nf(n)2^\ell)^{O(1)}$.*

Proof. For any ℓ satisfying $2/\ell^2 < \epsilon$, the work of [1, 9] reduces a LabelCover$(1, \epsilon)$ instance $\mathcal{I}$ with n vertices, m edges, and label size L to a set cover instance $\mathcal{S}$ with universe size $O(mL^2 2^{2\ell})$ and collection size nL such that approximating $\mathcal{S}$ to within $O(\ell)$ in time polynomial in $|\mathcal{S}|$ allows one to decide $\mathcal{I}$ in time polynomial in $|\mathcal{S}|$. Furthermore, the reduction takes time polynomial in $|\mathcal{S}|$. □

Corollary 4. *Set cover has no polynomial-time $2^{\log^{1-\delta_c(m)} m}$-approximation unless SAT can be decided in time $2^{O(2^{\log^{1-\delta_c(n)} n})}$.*

Proof. Combine Theorems 2 and 3 with $\ell = 2^{-\log^{1-\delta_c(m)} m}$. □

4 Conclusion

In the above discussion, to get better hardness for set cover in terms of m we showed $\Omega(\log n)$-hardness of approximation for set cover while decreasing m as a function of n. While this may be the right approach for reaching the limits of hardness in terms of m, such an approach may not be necessary to get more immediate improvements. For example, one may imagine being able to show $\Omega(\sqrt{\log n})$-hardness for set cover with a reduction where $m = O(\log^2 n)$, which would imply $\Omega(m^{1/4})$-hardness of approximation.

One might not expect to show hardness of approximation for set cover beyond $\Omega(\sqrt{m})$-hardness[2] due to an observation of Anupam Gupta and Danny Segev [5]. If $m < \ln n$ one can solve set cover in polynomial time by brute force. When $m > \ln^2 n$ the greedy $O(\log n)$-approximation algorithm also provides an $O(\sqrt{m})$ approximation. Thus, it suffices to give an $O(m/\log n)$-approximation algorithm for $\ln n < m < \ln^2 n$ to obtain an $O(\sqrt{m})$-approximation algorithm for all m. The algorithm is as follows, where the universe is $\{1, \ldots, n\}$ and the collection of sets is $\mathcal{C} = \{S_1, \ldots, S_m\}$.

1. Partition the S_j arbitrarily into $\ln n$ groups each of size at most $m/\ln n$. Let U_i be the union of all S_j in group i.

2. Solve a new set cover instance with $\mathcal{C}' = \{U_1, \ldots, U_{\ln n}\}$ optimally in polynomial time by brute force (note the optimal solution for $\mathcal{C}'$ is at least as cheap as that for $\mathcal{C}$).

3. Return the subcollection of $\mathcal{C}$ corresponding to the solution found in step 3. That is, if U_i is picked in step 3 then include all S_j in group i in the returned subcollection.

It would be interesting to pursue whether polynomial hardness is achievable, or perhaps whether one can find a relationship in the reverse direction (between LabelCover approximability and set cover approximability in terms of m) to show that improving the hardness presented here would require improving upon current hardness results for LabelCover.

Acknowledgments

I thank Erik Demaine for bringing [2] to my attention, Venkatesan Guruswami for bringing [1] to my attention, and Hervé Brönnimann, Swastik Kopparty, Yuval Rabani, Prasad Raghavendra, Madhu Sudan, and David Woodruff for useful comments and discussion. I also thank Anupam Gupta and Danny Segev for allowing me to include their $O(\sqrt{m})$-approximation algorithm.

References

[1] Sanjeev Arora, Carsten Lund. Hardness of Approximation. In *Approximation Algorithms for NP-Hard Problems*, Ed. D. Hochbaum, PWS Publishers, 1995, pp. 399–446.

[2] Hervé Brönnimann, Michael T. Goodrich. Almost Optimal Set Covers in Finite VC-Dimension. *Discrete & Computational Geometry*, 14(4):463–479, 1995.

[3] Irit Dinur, Shmuel Safra. On the Hardness of Approximating Label Cover. *Electronic Colloquium on Computational Complexity (ECCC)*, TR99-015, 1999.

[2] A previous version of this document conjectured too strongly that $\Omega(m^{1-\varepsilon})$-hardness should be achievable for any $\varepsilon > 0$.

[4] Uriel Feige. A Threshold of $\ln n$ for Approximating Set Cover. *J. ACM*, 45(4):634–652, 1998.

[5] Anupam Gupta, Danny Segev. Personal communication.

[6] David S. Johnson. Approximation Algorithms for Combinatorial Problems. *Journal of Computer and System Sciences*, 9(3):256–278, 1974.

[7] Richard M. Karp. Reducibility among Combinatorial Problems. *Complexity of Computer Computations*, R.E. Miller and J.W. Thatcher, eds. Plenum Press, New York, 1972.

[8] László Lovász. On the Ratio of the Optimal Integral and Fractional Covers. *Discrete Math.*, 13:383–390, 1975.

[9] Carsten Lund, Mihalis Yannakakis. On the Hardness of Approximating Minimization Problems. *J. ACM*, 41(5):960–981, 1994.

[10] Yuval Rabani, Leonard Schulman, Chaitanywa Swamy. Approximation Algorithms for Labeling Hierarchical Taxonomies. To appear in *Proceedings of the 19th Annual ACM Symposium on Discrete Algorithms*, 2008.

[11] Ran Raz, Shmuel Safra. A Sub-Constant Error-Probability Low-Degree Test, and Sub-Constant Error-Probability PCP Characterization of NP. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, pp. 474–484, 1997.

[12] Petr Slavík. A Tight Analysis of the Greedy Algorithm for Set Cover. In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, pp. 435–441, 1996.

[13] Aravind Srinivasan. Improved Approximation Guarantees for Packing and Covering Integer Programs. *SIAM J. Comput.*, 29(2):648–670, 1999.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/