Maximum Subset Intersection

Raphaël Clifford and Alexandru Popa

University of Bristol, Department of Computer Science, Merchant Venturer's Building, Woodland Road Bristol, BS8 1UB, United Kingdom

Abstract

Consider the following problem. Given n sets of sets $A_1, \ldots, A_u$ with elements over a universe $\mathcal{E} = \{e_1, \ldots, e_n\}$, the goal is to select exactly one set from each of $A_1, \ldots, A_u$ in order to maximize the size of the intersection of the sets. In this paper we present two NP-Hardness proofs, the first via a direct reduction from 3-SAT. The second proof involves a gap-preserving reduction from MAX-CLIQUE which enables us to show that our problem cannot be approximated within an $n^{1-\epsilon}$ multiplicative factor, for any $\epsilon > 0$, unless $NP = ZPP$.

Keywords: approximation algorithms, combinatorial problems.

1. Introduction

The set cover problem was among the first problems for which approximation algorithms were analyzed. Set cover is a fundamental problem in the field of approximation algorithms and its study led to the development of many important techniques. Johnson [8] shows that there is a greedy algorithm that gives an approximation ratio of $O(\log n)$. This result is extended by Chvátal [1] to the weighted version of set cover. In [2] Feige shows that there are no $(1 - \epsilon) \ln n$-approxmation algorithms for set cover, for any $\epsilon > 0$, unless $NP \subset TIME(n^{O(\log \log n)})$.

Many variants of this problem are defined: minimum sum set cover [3], k-set cover [4], maximum coverage [7], budgeted maximum coverage [9] and hitting set [5] are just a few example from a very rich literature. These problems have a wide variety of applications. For example, the maximum coverage problem has applications in circuit layout, job scheduling, facility location, and other areas (see e.g. Chapter 3 of [7] and the references therein).

In this paper we define and study the maximum subset intersection problem which is a natural counterpart to the set cover problem.

Problem 1. *(*MAX-INTERSECT*) Given n sets of sets $A_1, \ldots, A_u$ with elements over a universe $\mathcal{E} = \{e_1, \ldots, e_n\}$, the goal is to select exactly one set from each of $A_1, \ldots, A_u$ in order to maximize the size of the intersection of the sets.*

An application for this problem is as follows. Consider a production line consisting of u stages which, ideally, should produce n types of devices. A

different machine is responsible for each one of the stages. The problem is that each machine can have one of a finite number of possible settings. For each setting, a machine can produce only a fixed subset of the possible devices. The manufacturer wants to maximize the total number of items which can be produced by this production line by choosing the appropriate settings.

In this application, the sets of sets $A_1, \ldots, A_u$ correspond to the machines. The setting of a machine i is associated with a choice of one of the sets contained in A_i. The elements in the universe $\mathcal{E}$ are in turn the devices that can be produced by the production line. In the set cover version of this problem we would assume that each setting only has to be covered by one of the machines and that we want to minimize the number of machines chosen.

Although at a first glance this new problem appears similar to the other variants of set cover mentioned above, we show that it is computationally harder to find approximate solutions. We present two NP-Hardness proofs, the first via a direct reduction from 3-SAT. The second proof involves a gap-preserving reduction from MAX-CLIQUE which enables us to show that our problem cannot be approximated within an $n^{1-\epsilon}$ multiplicative factor, for any $\epsilon > 0$, unless $NP = ZPP$ (Section 4).

2. Preliminaries

In this section we recall some background material useful for the results of this paper.

Definition. *(3-SAT) Given a Boolean formula in CNF (Conjunctive Normal Form) where each clause contains at most three literals (a literal is either variable or its negation), decide if there is a truth assignment of the variables that satisfies all the clauses.*

Definition. *(MAX-CLIQUE) Given an undirected graph G=(V,E) find a subset of nodes of maximum cardinality such that every pair of nodes in the set are adjacent.*

The two problems above are NP-Hard [5].

Definition. *An algorithm $\mathcal{A}$ is a c-approximation for problem Π, where c can be a constant or a function in the size of the input instance, if for all input instances x:*

- $\mathcal{A}(x) \leq c \cdot OPT_{\Pi}(x)$, *if* Π *is a minimization problem.*
- $\mathcal{A}(x) \geq \frac{OPT_{\Pi}(x)}{c}$, *if* Π *is a maximization problem.*

We give the definition of gap-introducing reduction between two maximization problems. A similar definition is presented in [10] for the case when the first problem is a minimization problem and the second is a maximization problem.

Definition. *[10] Assume Π_1 and Π_2 are some maximization problems. A gap-preserving reduction from Π_1 to Π_2 comes with four parameters (functions) f_1, α, f_2 and β. Given an instance x of Π_1, the reduction computes in polynomial time an instance y of Π_2 such that:*

$$OPT(x) \geq f_1(x) \Rightarrow OPT(y) \geq f_2(y)$$

$$OPT(x) < \alpha(|x|)f_1(x) \Rightarrow OPT(y) < \beta(|y|)f_2(y)$$

A gap-preserving reduction from Π_1 to Π_2 with the above parameters implies that if the problem Π_1 cannot have a α-approximation, then the problem Π_2 cannot have a β-approximation.

The next theorem was proved by Håstad in [6].

Theorem 1. *[6]* Max-Clique *does not have an $n^{1-\epsilon}$ approximation, for any $\epsilon > 0$, unless NP = ZPP .*

3. NP-Completeness

In this section we prove that Max-Intersect is NP-Hard using a reduction from 3-Sat. The decision version of Max-Intersect is the following.

Problem 2. *(*Set-Intersect*) Given an instance of* Max-Intersect *and an integer k, is it possible to select exactly one set from each of $A_1, \ldots, A_u$ so that their intersection has size at least k? An instance $(\{A_1, \ldots, A_u\}, k)$ is a YES instance if there is a collection of sets $\{S_1, \ldots, S_n\}$, $S_i \in A_i$ such that the cardinality of $\bigcap_{i=1}^{u} S_i$ is greater than or equal to k, otherwise it is a NO instance.*

We now describe the reduction used to prove the NP-Completeness of Set-Intersect. Consider a 3-Sat boolean formula with n variables and m clauses

$$\phi = (x_1(1) \vee x_1(2) \vee x_1(3)) \wedge \cdots \wedge (x_m(1) \vee x_m(2) \vee x_m(3))$$

where each x_{ij}, $1 \leq i \leq m$, $1 \leq j \leq 3$ is a literal (either a variable or the negation of a variable).

We construct an instance of Set-Intersect as follows. For each variable x we add two elements, x_1 and x_0, to the universe of elements. Intuitively, these correspond to an assignment of True or False, respectively, to the variable x. The resulting universe $\mathcal{E}$ has precisely $2n$ elements. To each clause and each variable we associate a set of sets of elements from the universe.

To simplify notation, in the rest of this section we write $\mathcal{E} \setminus x$ for $\mathcal{E} \setminus \{x\}$. To a clause $x_i(1) \vee x_i(2) \vee x_i(3)$ we associate the set:

$$C_i = \{\{\mathcal{E} \setminus x^{i1}\}, \{\mathcal{E} \setminus x^{i2}\}, \{\mathcal{E} \setminus x^{i3}\}\}$$

where for $1 \leq k \leq 3$, x^{ik} is defined as:

$$x^{ik} = \begin{cases} x_0 & \text{if } x_i(k) \text{ is a nonnegated literal corresponding to the variable } x \\ x_1 & \text{if } x_i(k) \text{ is a negated literal corresponding to the variable } x \end{cases}$$

To each variable x_i we associate the set of sets

$$V_i = \{\{\mathcal{E} \setminus x_{i1}\}, \{\mathcal{E} \setminus x_{i0}\}\}$$

The idea behind this construction is that a set of sets corresponding to a variable forces that variable to be assigned either true or false, and a set of sets corresponding to a clause ensures the satisfiability of that clause.

Example 1. *Consider the following boolean formula:*

$$\phi = (x \vee \bar{y} \vee z) \wedge (x \vee y \vee \bar{z})$$

Then the corresponding sets of sets will be:

$$C_1 = \{\{x_1, y_0, y_1, z_0, z_1\}, \{x_0, x_1, y_0, z_0, z_1\}, \{x_0, x_1, y_0, y_1, z_1\}\}$$

$$C_2 = \{\{x_1, y_0, y_1, z_0, z_1\}, \{x_0, x_1, y_1, z_0, z_1\}, \{x_0, x_1, y_0, y_1, z_0\}\}$$

$$V_x = \{\{x_0, y_0, y_1, z_0, z_1\}, \{x_1, y_0, y_1, z_0, z_1\}\}$$

$$V_y = \{\{x_0, x_1, y_0, z_0, z_1\}, \{x_0, x_1, y_1, z_0, z_1\}\}$$

$$V_z = \{\{x_0, x_1, y_0, y_1, z_0\}, \{x_0, x_1, y_0, y_1, z_1\}\}$$

A possible satisfying truth assignment for ϕ is $x = false$, $y = true$, $z = true$. The set of elements which corresponds to this assignment is $\{x_0, y_1, z_1\}$. We can find this set by selecting: third set from C_1, second set from C_2, first set from V_x and second set from V_y and V_z.

We have to prove that if we can solve the Set-Intersect problem in polynomial time, then we can decide in polynomial time if a 3-Sat formula is satisfiable or not. This is formally stated in the following lemma.

Lemma 1. *Let ϕ be a boolean formula and $Q = \{V_1, \ldots, V_n, C_1, \ldots, C_m\}$ be the set of sets associated with this formula by the variable and clause gadgets. ϕ has a satisfiable truth assignment if and only if (Q, n) is a YES instance to* Set-Intersect *problem.*

Proof. For $1 \leq i \leq n$, let $T_i \in V_i$. Then,

$$\left| \bigcap_{i=1}^{n} T_i \right| = n$$

since the set T_i selected from V_i constrains us to pick only one of the two elements x_{i0} and x_{i1} that were associated with the set V_i. Since $\{V_1, \ldots, V_n\} \subset Q$, the size of the intersection of any sets selected from the sets of Q is less than or equal to n. This follows as we must select exactly one set from each of the sets contained in Q. Therefore, we only have to prove that the maximum size of the intersection of sets selected from sets of Q is greater than or equal to n.

We prove the "if" part of the lemma. Consider an assignment to the variables $x_1, \ldots, x_n$ which satisfies ϕ. Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be the corresponding set of elements:

$$s_i = \begin{cases} x_{i1} & \text{if } x_i \text{ is set to true} \\ x_{i0} & \text{if } x_i \text{ is set to false} \end{cases}$$

We select from every set of sets $(V_i)_{1 \leq i \leq n}$ and $(C_j)_{1 \leq j \leq m}$ the set that includes $\mathcal{S}$. Now we argue that in every $(V_i)_{1 \leq i \leq n}$ and $(C_j)_{1 \leq j \leq m}$ there is a such a set. From a set V_i corresponding to a variable x_i we select $\{\mathcal{E} \setminus x_{i1}\}$ if x_i is set to false, or $\{\mathcal{E} \setminus x_{i0}\}$ if x_i is set to true. Every C_i must also contain S, since all the clauses are satisfied. Let x be a variable that satisfies this clause: if x is set to true then we select $\{\mathcal{E} \setminus x_0\}$, otherwise we select $\{\mathcal{E} \setminus x_1\}$.

We prove the "only if" part. Suppose the intersection of sets is $\mathcal{S}$ and has size n. From this set we can find a satisfying truth assignment for the formula ϕ in the following way: if $x_{i1} \in \mathcal{S}$ then we set x_i to true, otherwise we set x_i to false. The resulting assignment is consistent: by the way we define sets V_i only one of the elements x_{i0} and x_{i1} can be part of the solution. All the clauses of ϕ are satisfied: the sets from C_i define all the three possible ways that clause can be satisfied (by satisfying the first literal, the second one, or the third one), therefore selecting one set from C_i constrains the associated assignment to satisfy clause i. □

Theorem 2. Set-Intersect *problem is NP-complete.*

Proof. The total number of elements in all the sets of sets constructed using the above reduction is $3(2n-1)m + 2(2n-1)n$, which is a polynomial in n and m (the number of variables, respectively the number of clauses of the boolean formula ϕ). Therefore, the reduction presented is polynomial time. Set-Intersect problem is also in NP, since the intersection of a given set of sets can be computed in polynomial time. These observations and Lemma 1 prove the theorem. □

4. Inapproximability results

In the previous section we show that the decision version of Max-Intersect is NP-complete. In this section we investigate the hardness of approximation of the optimization version.

We present the following gap-preserving reduction from Max-Clique. Let $G = (V, E)$ be a graph with n vertices and m edges. For every vertex i we add two elements in the universe, called i and i' (so there are $2n$ elements). To each vertex i of the graph we associate a set $C_i = \{A_i, B_i\}$, where the sets A_i, and B_i are defined as follows:

$$A_i = \{x \in V |\ i = x \text{ or } (i, x) \in E\}$$

$$B_i = \{i'\} \cup \{x \in V |\ i \neq x\}$$

The first set contains i and all the adjacent vertices, and the second set contains i' and $V - \{i\}$.

Example 2. *Suppose we have a graph $G = (V, E)$ with $V = \{1, 2, 3, 4\}$ and $E = \{(1, 2), (1, 3), (1, 4), (3, 4)\}$. Then the set of sets will be:*

$$\{\{1, 2, 3, 4\}, \{1', 2, 3, 4\}\}$$
$$\{\{2, 1\}, \{2', 1, 3, 4\}\}$$
$$\{\{3, 1, 4\}, \{3', 1, 2, 4\}\}$$
$$\{\{4, 1, 3\}, \{4', 1, 2, 3\}\}$$

The maximum intersection set and the maximum clique are $\{1, 3, 4\}$.

We prove that this is a gap-preserving reduction.

Lemma 2. *Let A be an instance of the* Max-Clique *problem and B be the corresponding instance of the* Max-Intersect *problem. Let $\epsilon > 0$.*

$$OPT(A) = n \Rightarrow OPT(B) = n$$
$$OPT(A) < n^{1-\epsilon} \Rightarrow OPT(B) < n^{1-\epsilon}$$

Proof. The maximum clique in $G = (V, E)$ has the same as size as

$$\max_{P_i \in C_i} \left| \bigcap_{i=1}^{n} P_i \right|$$

Let the maximum clique be $S = \{i_1, \ldots, i_k\}$. The intersection of the sets $\mathcal{I} = \{I_1, \ldots, I_n\}$ is exactly S, where

$$I_i = \begin{cases} A_i & \text{if } i \in S \\ B_i & \text{otherwise} \end{cases}$$

This is true since any two nodes i and j in the clique are adjacent and therefore $i \in I_j$ and $j \in I_i$.

On the other hand, if the maximum intersection set is $S = \{i_1, \ldots, i_k\}$, then it is also a maximum clique in the graph $G = (V, E)$. First of all, notice that elements i' cannot be present in the final solution (an element i' is present only in the set C_i). Then, any two elements $i \in S$ and $j \in S$ correspond to adjacent vertices in the graph (i is in the neighbors list of j and j is in the neighbors list of i). Thus, set S is a maximum clique in $G = (V, E)$.

Therefore, the reduction presented above is a gap-preserving reduction. □

The inapproximability result is formally stated in the following theorem.

Theorem 3. *For any constant $\epsilon > 0$ the* Max-Intersect *problem does not admit a $n^{1-\epsilon}$-approximation unless NP = ZPP.*

Proof. The proof of the theorem is given by the reduction presented, Lemma 2 and Theorem 1. □

5. Conclusions and open problems

In this paper we present the maximum intersection set problem, we prove that it is NP-complete and we show that for any constant $\epsilon > 0$ the problem is inapproximable to $n^{1-\epsilon}$, unless $NP = ZPP$. In the end, we state the following open problem which might have a better approximation ratio.

Problem 3. *Given n sets $A_1, \ldots, A_u$ over a finite universe $\mathcal{E} = \{e_1, \ldots, e_n\}$ and an integer k, the goal is to select exactly k sets so their intersection is maximized.*

It worth noticing that a similar reduction from MAX-CLIQUE does not work here. Suppose n sets are constructed, one for each vertex, each set containing the neighbors of the vertex and the vertex itself. On one hand it is true that if the maximum size of a clique is k then the maximum intersection of k sets has to be larger than k. But, on the other hand, if the maximum size of a clique is small then the maximum intersection can still be large.

It is also interesting that the following similar problem admits a better approximation factor.

Problem 4. *Given n sets $A_1, \ldots, A_u$ over a finite universe $\mathcal{E} = \{e_1, \ldots, e_n\}$ and an integer k, the goal is to select exactly k sets so their intersection is minimized.*

This problem is actually equivalent to max-coverage problem [7]. Construct the complement of the given sets, and find the maximum coverage of k of them. Then the minimum intersection set is given by the entire universe minus the elements from maximum coverage.

Acknowledgements. I would like to thank Andrew Moss and Raphaël Clifford for giving me the problem. I would also like to thank to Anna Adamaszek for pointing me the open problem discussed in the end.

[1] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

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

[3] Uriel Feige, László Lovász, and Prasad Tetali. Approximating min sum set cover. *Algorithmica*, 40(4):219–234, 2004.

[4] Rajiv Gandhi, Samir Khuller, and Aravind Srinivasan. Approximation algorithms for partial covering problems. *Journal of Algorithms*, 53(1):55–84, 2004.

[5] M. R. Garey and D. S. Johnson. *Computers and intractability. A guide to the theory of NP-completeness.* W. H. Freeman, 1979.

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

[7] Dorit S. Hochbaum. *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company, 1997.

[8] David S. Johnson. Approximation algorithms for combinatorial problems. In *STOC*, pages 38–49, 1973.

[9] Samir Khuller, Anna Moss, and Joseph Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[10] Vijay V. Vazirani. *Approximation Algorithms*. Springer, 2004.