Hardness of Set Cover with Intersection 1

V.S.Anil Kumar[1], Sunil Arya[2] and H.Ramesh[3]

[1] MPI für Informatik, Saarbrücken. kumar@mpi-sb.mpg.de
[2] Department of Computer Science, Hong Kong University of Science and Technology. arya@cs.ust.hk
[3] Department of Computer Science and Automation, Indian Institute of Science, Bangalore. ramesh@csa.iisc.ernet.in

Abstract. We consider a restricted version of the general Set Covering problem in which each set in the given set system intersects with any other set in at most 1 element. We show that the Set Covering problem with intersection 1 cannot be approximated within a $o(\log n)$ factor in random polynomial time unless $NP \subseteq ZTIME(n^{O(\log\log n)})$. We also observe that the main challenge in derandomizing this reduction lies in finding a hitting set for large volume combinatorial rectangles satisfying certain intersection properties. These properties are not satisfied by current methods of hitting set construction.
An example of a Set Covering problem with the intersection 1 property is the problem of covering a given set of points in two or higher dimensions using straight lines; any two straight lines intersect in at most one point. The best approximation algorithm currently known for this problem has an approximation factor of $\theta(\log n)$, and beating this bound seems hard. We observe that this problem is Max-SNP-Hard.

1 Introduction

The general Set Covering problem requires covering a given base set B of size n using the fewest number of sets from a given collection of subsets of B. This is a classical NP-Complete problem and its instances arise in numerous diverse settings. Thus approximation algorithms which run in polynomial time are of interest.

Johnson[12] showed that the greedy algorithm for Set Cover gives an $O(\log n)$ approximation factor. Much later, following advances in Probabilistically Checkable Proofs [4], Lund and Yannakakis [15] and Bellare et al. [7] showed that there exists a positive constant c such that the Set Covering problem cannot be approximated in polynomial time within a $c \log n$ factor unless $NP \subseteq DTIME(n^{O(\log\log n)})$. Feige [10] improved the approximation threshold to $(1 - o(1))\log n$, under the same assumption. Raz and Safra[19] and Arora and Sudan[5] then obtained improved Probabilistically Checkable Proof Systems with sub-constant error probability; their work implied that the Set Covering problem cannot be approximated within a $c \log n$ approximation factor (for some constant c) unless $NP = P$.

Note that all the above hardness results are for general instances of the Set Covering problem and do not hold for instances when the intersection of any pair of sets in the given collection is guaranteed to be at most 1. Our motivation for considering this restriction to intersection 1 arose from the following geometric instance of the Set Covering problem.

Given a collection of points and lines in a plane, consider the problem of covering the points with as few lines as possible. Megiddo and Tamir[16] showed that this problem is NP-Hard. Hassin and Megiddo[11] showed NP-Hardness even when the lines are axis-parallel but in 3D. The best approximation factor known for this problem is $\Theta(\log n)$. Improving this factor seems to be hard, and this motivated our study of inapproximability for Set Covering with intersection 1. Note that any two lines intersect in at most 1 point.

The problem of covering points with lines was in turn motivated by the problem of covering a rectilinear polygon with holes using rectangles [13]. This problem has applications in printing integrated circuits and image compression[9]. This problem is known to be Max-SNP-Hard even when the rectangles are constrained to be axis-parallel. For this case, an $O(\sqrt{\log n})$-factor approximation algorithm was obtained recently by Anil Kumar and Ramesh[2]. However, this algorithm does not extend to the case when the rectangles need not be axis-parallel. Getting a $o(\log n)$-factor approximation algorithm for this case seems to require solving the problem of covering points with arbitrary lines, though we are not sure of the exact nature of this relationship.

Our Result. We show that there exists a constant $c > 0$ such that approximating the Set Covering problem with intersection 1 to within a factor of $c \log n$ in random polynomial time is possible only if $NP \subseteq ZTIME(n^{O(\log \log n)})$ (where $ZTIME(t)$ denotes the class of languages that have a probabilistic algorithm running in expected time t with zero error). We also give a sub-exponential derandomization which shows that approximating the Set Covering problem with intersection 1 to within a factor of $c\frac{\log n}{\log \log n}$ in deterministic polynomial time is possible only if $NP \subseteq DTIME(2^{n^{1-\epsilon}})$, for any constant $\epsilon < 1/2$.

The starting point for our result above is the Lund-Yannakakis hardness proof[15] for the general Set Covering problem. This proof uses an auxiliary set system with certain properties. We show that this auxiliary set system necessarily leads to large intersection. We then replace this auxiliary set system by another carefully chosen set system with additional properties and modify the reduction appropriately to ensure that intersection sizes stay small. The key features of the new set system are partitions of the base set into several sets of smaller size (instead of just 2 sets as in the case of the Lund-Yannakakis system or a constant number of sets as in Feige's system; small sets will lead to small intersection) and several such partitions (so that sets which "access" the same partition in the Lund-Yannakakis system and therefore have large intersection now "access" distinct partitions).

We then show how the new set system above can be constructed in randomized polynomial time and also how this randomized algorithm can be derandomized using conditional probabilities and appropriate estimators in $O(2^{n^{1-\epsilon}})$ time, where ϵ is a positive constant, specified in Section 5. This leads to the two conditions above, namely, $NP \subseteq DTIME(2^{n^{1-\epsilon}})$ (but for a hardness of $O(\frac{\log n}{\log\log n})$) and $NP \subseteq ZTIME(n^{O(\log\log n)})$. A deterministic polynomial time construction of our new set system will lead to the quasi-NP-Hardness of approximating the Set Covering problem with intersection 1 to within a factor of $c \log n$, for some constant $c > 0$.

While the Lund-Yannakakis set system can be constructed in deterministic polynomial time using ϵ-biased limited independence sample spaces, this does not seem to be true of our set system. One of the main bottlenecks in constructing our set system in deterministic polynomial time is the task of obtaining a polynomial size hitting set for *Combinatorial Rectangles*, with the hitting set satisfying additional properties. One of these properties (the most important one) is the following: if a hitting set point has the elements i, j among its coordinates, then no other hitting set point can have both i, j among its coordinates. The only known construction of a polynomial size hitting set for combinatorial rectangles is by Linial, Luby, Saks, and Zuckerman [14] and is based on enumerating walks in a constant degree expander graph. In the full version of this paper, we show that the hitting set obtained by [14] does not satisfy the above property for reasons that seem intrinsic to the use of constant degree expander graphs.

In the full version, we also note that if the proof systems for NP obtained by Raz and Safra[19] or Arora and Sudan[5] have an additional property then the condition $NP \subseteq ZTIME(n^{O(\log\log n)})$ can be improved to $NP = ZPP$. Similarly, the statement that approximating the Set Covering problem with intersection 1 to within a factor of $c\frac{\log n}{\log\log n}$ in deterministic polynomial time is possible only if $NP \subseteq DTIME(2^{n^{1-\epsilon}})$ can be strengthened to approximation factor $c \log n$ instead of $c\frac{\log n}{\log\log n}$. The property needed of the proof systems is that the *degree*, i.e., the total number of random choices of the verifier for which a particular question is asked of a particular prover, be $O(n^\delta)$, for some small enough constant value δ. The degree influences the number of partitions in our auxiliary proof system and therefore needs to be small. It is not clear whether existing proof systems have this property [20].

The above proof of hardness for Set Covering with intersection 1 does not apply to the problem of covering points with lines, the original problem which motivated this paper; however, it does indicate that algorithms based on set cardinalities and small pairwise intersection alone are unlikely to give a $o(\log n)$ approximation factor for this problem.

Further, our result shows that constant VC-dimension alone does not help in getting a $o(\log n)$ approximation for the Set Covering problem. This is to be contrasted with the result of Brönnimann and Goodrich[8] which shows that if the VC-dimension is a constant and an $O(\frac{1}{\epsilon})$ sized (weighted) ϵ-net can be constructed in polynomial time, then a constant factor approximation can be obtained.

The paper is organized as follows. Section 2 will give an overview of the Lund-Yannakakis reduction. Section 3 shows why the Lund-Yannakakis proof does not show hardness of Set Covering when the intersection is constrained to be 1. Section 4 describes the reduction to Set Covering with intersection 1. This section describes a new set system we need to obtain in order to perform the reduction and shows hardness of approximation of its set cover, unless $NP \subseteq ZTIME(n^{O(\log\log n)})$. Section 5 will sketch the randomized construction of this set system. Section 6 sketches the sub-exponential time derandomization, which leads to a slightly different hardness result, unless $NP \subseteq DTIME(2^{n^{1-\epsilon}})$, $\epsilon < 1/2$. Section 7 enumerates several interesting open problems which arise from this paper.

2 Preliminaries: The Lund-Yannakakis Reduction

In this section, we sketch the version of the Lund-Yannakakis reduction described by Arora and Lund [3]. The reduction starts with a 2-Prover 1-Round proof system for Max-3SAT(5) which has inverse polylogarithmic error probability, uses $O(\log n \log\log n)$ randomness, and has $O(\log\log n)$ answer size. Here n is the size of the Max-3SAT(5) formula $\mathcal{F}$. Arora and Lund[3] abstract this proof system into the following *Label Cover* problem.

The Label Cover Problem. A bipartite graph G having $n' + n'$ vertices and edge set E is given, where $n' = n^{O(\log\log n)}$. All vertices have the same degree deg, which is polylogarithmic in n. For each edge $e \in E$, a partial function $f_e : [d] \to [d']$ is also given, where $d \geq d'$, and d, d' are polylogarithmic in n. The aim is to assign to each vertex on the left, a label in the range $1 \ldots d$, and to each vertex on the right, a label in the range $1 \ldots d'$, so as to maximize the number of edges $e = (u, v)$ satisfying $f_e(label(u)) = label(v)$. Edge $e = (u, v)$ is said to be *satisfied* by a labelling if the labelling satisfies $f_e(label(u)) = label(v)$.

The 2-Prover 1-Round proof system mentioned above ensures that either all the edges in G are satisfied by some labelling or that no labelling satisfies more than a $\frac{1}{\log^3 n}$ fraction of the edges, depending upon whether or not the Max-3SAT(5) formula $\mathcal{F}$ is satisfiable. Next, in time polynomial in the size of G, an instance $\mathcal{SC}$ of the Set Covering problem is obtained from this Label Cover problem $\mathcal{LC}$ with the following properties: if there exists a labelling satisfying all edges in G then there is a set cover of size $2n'$, and if no labelling satisfies more than a $\frac{1}{\log^3 n}$ fraction of the edges then the smallest set cover has size $\Omega(2n' \log n')$. The base set in $\mathcal{SC}$ will have size polynomial in n'. It follows that the Set Covering problem cannot be approximated to a logarithmic factor of the base set size unless $NP \subseteq DTIME(n^{O(\log\log n)})$.

Improving this condition to $NP = P$ requires using a stronger multi-prover proof system [19, 5] which has a constant number of provers (more than 2), $O(\log n)$ randomness, $O(\log\log n)$ answer sizes, and inverse polylogarithmic error probability. The reduction from such a proof system to the Set Covering problem is similar to the reduction from the Label Cover to the Set Covering problem mentioned above, with a modification needed to handle more than 2 provers (this modification is described in [7]).

In this abstract, we will only describe the reduction from Label Cover to the Set Covering problem and show how we can modify this reduction to hold for the case of intersection 1. This will show that Set Covering problem with intersection 1 cannot be approximated to a logarithmic factor unless $NP \subseteq ZTIME(n^{O(\log\log n)})$. The multi-prover proof system of the previous paragraph with an additional condition can strengthen the latter condition to $NP = ZPP$; this is described in the full version.

We now briefly sketch the reduction from an instance $\mathcal{LC}$ of Label Cover to an instance $\mathcal{SC}$ of the Set Covering problem.

2.1 Label Cover to Set Cover

The following auxiliary set system given by a base set $N = \{1 \dots n'\}$ and its partitions is needed.

The Auxiliary System of Partitions. Consider d' distinct partitions of N into two sets each, with the partitions satisfying the following property: if at most $\frac{\log n'}{2}$ sets in all are chosen from the various partitions with no two sets coming from the same partition, then the union of these sets does not cover N. Partitions with the above properties can be constructed deterministically in polynomial time [1, 17]. Let P_i^1, P_i^2 respectively denote the first and second sets in the ith partition. We describe the construction of $\mathcal{SC}$ next.

Using P_i^js to construct $\mathcal{SC}$. The base set B for $\mathcal{SC}$ is defined to be $\{(e,i)|e \in E, 1 \leq i \leq n'\}$. The collection C of subsets of B contains a set $C(v,a)$, for each vertex v and each possible label a with which v can be labelled. If v is a vertex on the left, then for each a, $1 \leq a \leq d$, $C(v,a)$ is defined as $\{(e,i)|e \text{ incident on } v \wedge i \in P^1_{f_e(a)}\}$. And if v is a vertex on the right, then for each a, $1 \leq a \leq d'$, $C(v,a)$ is defined as $\{(e,i)|e \text{ incident on } v \wedge i \in P_a^2\}$.

That $\mathcal{SC}$ satisfies the required conditions can be seen from the following facts.

1. If there exists a vertex labelling which satisfies all the edges, then B can be covered by just the sets $C(v,a)$ where a is the label given to v. Thus the size of the optimum cover is $2n'$ in this case.
2. If the total number of sets in the optimum set cover is at most some suitable constant times $n' \log n'$, then at least a constant fraction of the edges $e = (u,v)$ have the property that the number of sets of the form $C(u,*)$ plus the number of sets of the form $C(v,*)$ in the optimum set cover is at most $\frac{\log n'}{2}$. Then, for each such edge e, there must exist a label a such that $C(u,a)$ and $C(v, f_e(a))$ are both in this optimum cover. It can be easily seen that choosing a label uniformly at random from these sets for each vertex implies that there exists a labelling of the vertices which satisfies an $\Omega(\frac{1}{\log^2 n'}) \geq \frac{1}{\log^3 n}$ fraction of the edges.

3 $\mathcal{SC}$ has Large Intersection

There are two reasons why sets in the collection C in $\mathcal{SC}$ have large intersections.

Parts in the Partitions are Large. The first and obvious reason is that the sets in each partition in the auxiliary system of partitions are large and could have size $\frac{n'}{2}$; therefore, two sets in distinct partitions could have $\Omega(n')$ intersection. This could lead to sets $C(v,a)$ and $C(v,b)$ having $\Omega(n')$ common elements of the form (e,i), for some e incident on v.

Clearly, the solution to this problem is to work with an auxiliary system of partitions where each partition is a partition into not just 2 large sets, but into several small sets. The problem remains if we form only a constant number of parts, as in [10]. We choose to partition into $(n')^{1-\epsilon}$ sets, where ϵ is some non-zero constant to be fixed later. This ensures that each set in each partition has size $\theta((n')^{\epsilon}\ \mathrm{polylog}(n))$ and that any two such sets have $O(1)$ intersection. However, smaller set size leads to other problems which we shall describe shortly.

Functions $f_e()$ are not 1-1. Suppose we work with smaller set sizes as above. Then consider the sets $C(v,a)$ and $C(v,b)$, where v is a vertex on the left and a,b are labels with the following property: for some edge e incident on v, $f_e(a) = f_e(b)$. Then each element $(e,*)$ which appears in $C(v,a)$ will also appear in $C(v,b)$, leading to an intersection size of up to $\Omega((n')^{\epsilon} * deg)$, where deg is the degree of v in G. This is a more serious problem. Our solution to this problem is to ensure that sets $C(v,a)$ and $C(v,b)$ are constructed using distinct partitions in the auxiliary system of partitions.

Next, we describe how to modify the auxiliary system of partitions and the construction of $\mathcal{SC}$ in accordance with the above.

4 $\mathcal{LC}$ to $\mathcal{SC}$ with Intersection 1

Our new auxiliary system of partitions $\mathcal{P}$ will have $d' * (deg+1) * d$ partitions, where deg is the degree of any vertex in G. Each partition has $m = (n')^{1-\epsilon}$ parts, for some $\epsilon > 0$ to be determined. These partitions are organized into d' *groups*, each containing $(deg+1) * d$ partitions. Each group is further organized into $deg + 1$ *subgroups*, each containing d partitions. The first $m/2$ sets in each partition comprise its *left half* and the last $m/2$ its *right half*.

Let $P_{g,s,p}$ denote the pth partition in the sth subgroup of the gth group and let $P_{g,s,p,k}$ denote the kth set (i.e., part) in this partition. Let B_k denote the set $\cup_{g,s,p} P_{g,s,p,k}$ if $1 \le k \le m/2$, and the set $\cup_{g,s} P_{g,s,1,k}$, if $m/2 < k \le m$. We also refer to B_k as the kth *column* of $\mathcal{P}$.

We need the following properties to be satisfied by the system of partitions $\mathcal{P}$.

1. The right sides of all partitions within a subgroup are identical, i.e., $P_{g,s,p,k} = P_{g,s,1,k}$, for every $k > m/2$.
2. $P(g,s,p,k) \cap P(g',s',p',k) = \phi$ unless either $g = g', s = s', p = p'$, or, $k > m/2$ and $g = g', s = s'$. In other words, no element appears twice within a column, modulo the fact that the right sides of partitions within a subgroup are identical.
3. $|B_k \cap B_{k'}| \le 1$ for all k, k', $1 \le k, k' \le m$, $k \ne k'$.

4. Suppose N is covered using at most $\beta m \log n'$ sets in all, disallowing sets on the right sides of those partitions which are not the first in their respective subgroups. Then there must be a partition in some subgroup s such that the number of sets chosen from the left side of this partition plus the number of sets chosen from right side of the first partition in s together sum to at least $\frac{3}{4}m$.

ϵ and β are constants which will be fixed later. Let $A_{p,k} = \cup_{g,s} P_{g,s,p,k}$, for each p, k, $1 \leq p \leq d, 1 \leq k \leq m/2$. Let $D_{g,k} = \cup_s P_{g,s,1,k}$, for each g, k, $1 \leq g \leq d'$, $m/2 + 1 \leq k \leq m$. Property 2 above implies that:

5. $|A_{p,k} \cap A_{p',k}| = 0$ for all $p \neq p'$, where $1 \leq p, p' \leq d$ and $k \leq m/2$.
6. $|D_{g,k} \cap D_{g',k}| = 0$ for all $g \neq g'$, where $1 \leq g, g' \leq d'$ and $k > m/2$.

We will describe how to obtain a system of partitions $\mathcal{P}$ satisfying these properties in Section 5 and Section 6. First, we show how a set system $\mathcal{SC}$ with intersection 1 can be constructed using $\mathcal{P}$.

4.1 Using $\mathcal{P}$ to construct $\mathcal{SC}$

The base set B for $\mathcal{SC}$ is defined to be $\{(e, i) | e \in E, 1 \leq i \leq n'\}$ as before. This set has size $(n')^2 * deg = O((n')^2 \text{ polylog}(n))$.

The collection C of subsets of B contains $m/2$ sets $C_1(v, a) \ldots C_{m/2}(v, a)$, for each vertex v on the left (in graph G) and each possible label a with which v can be labelled. In addition, it contains $m/2$ sets $C_{m/2+1}(v, a) \ldots C_m(v, a)$, for each vertex v on the right in G and each possible label a with which v can be labelled. These sets are defined as follows.

Let E_v denote the set of edges incident on v in G. We edge-colour G using $deg + 1$ colours. Let $col(e)$ be the colour given to edge e in this edge colouring. For a vertex v on the left side, and any number k between 1 and $m/2$, $C_k(v, a) = \cup_{e \in E_v} \{(e, i) | i \in P_{f_e(a), col(e), a, k}\}$. For a vertex v on the right side, and any number k between $m/2 + 1$ and m, $C_k(v, a) = \cup_{e \in E_v} \{(e, i) | i \in P_{a, col(e), 1, k}\}$.

We now give the following lemmas which state that the set system $\mathcal{SC}$ has intersection 1 and that it has a set cover of small size if and only if there exists a way to label the vertices of G satisfying several edges simultaneously. The hardness of approximation of the set cover of $\mathcal{SC}$ is given in Corollary 1, whose proof will appear in the full version.

Lemma 1. *The intersection of any two distinct sets $C_k(v, a)$ and $C_{k'}(w, b)$ is at most 1.*

Proof. Note that for $|C_k(v, a) \cap C_{k'}(w, b)|$ to exceed 1, either v, w must be identical or there must be an edge between v and w. The reason for this is that each element in $C_k(v, a)$ has the form $(e, *)$ where e is an edge incident at v while each element in $C_{k'}(w, b)$ has the form $(e', *)$, where e' is an edge incident at w. We consider each case in turn.

Case 1. Suppose $v = w$. Then either $k \neq k'$ or $k = k', a \neq b$.

First, consider $C_k(v,a)$ and $C_{k'}(v,b)$ where $k \neq k'$ and v is a vertex in the left side. If $a = b$, observe that $C_k(v,a) \cap C_{k'}(v,a) = \phi$. So assume that $a \neq b$. The elements in the former set are of the form (e,i) where $i \in P_{f_e(a),col(e),a,k}$ and the elements of the latter set are of the form (e,j) where $j \in P_{f_e(b),col(e),b,k'}$. Note that $\cup_{e \in E_v} P_{f_e(a),col(e),a,k} \subseteq B_k$ and $\cup_{e \in E_v} P_{f_e(b),col(e),b,k'} \subseteq B_{k'}$. By Property 3 of $\mathcal{P}$, the intersection $B_k, B_{k'}$ is at most 1. However, this alone does not imply that $C_k(v,a)$ and $C_{k'}(v,b)$ have intersection at most 1, because there could be several tuples in both sets, all having identical second entries. This could happen if there are edges e_1, e_2 incident on v such that $f_{e_1}(a) = f_{e_2}(a), f_{e_1}(b) = f_{e_2}(b)$ and there had been no colouring on edges. Property 2 and the fact that $col(e_1) \neq col(e_2)$ for any two edges e_1, e_2 incident on v rule out this possibility, thus implying that $|C_k(v,a) \cap C_{k'}(v,b)| \leq 1$. The proof for the case where v is a vertex on the right is identical.

Second, consider $C_k(v,a)$ and $C_k(v,b)$, where v is a vertex on the left and $a \neq b$. Elements in the former set are of the form (e,i) where e is an edge incident on v and $i \in P_{f_e(a),col(e),a,k}$. Similarly, elements in the latter set are of the form (e,j) where $j \in P_{f_e(b),col(e),b,k}$. Note that $\cup_{e \in E_v} P_{f_e(a),col(e),a,k} \subseteq A_{a,k}$ and $\cup_{e \in E_v} P_{f_e(b),col(e),b,k} \subseteq A_{b,k}$. The claim follows from Property 5 in this case.

Third, consider $C_k(v,a)$ and $C_k(v,b)$, where v is a vertex on the right, $a \neq b$, and $k > m/2$. Elements in the former set are of the form (e,i) where e is an edge incident on v and $i \in P_{a,col(e),1,k}$. Similarly, elements in the latter set are of the form (e,j) where $j \in P_{b,col(e),1,k}$. Note that $\cup_{e \in E_v} P_{a,col(e),1,k} \subseteq D_{a,k}$ and $\cup_{e \in E_v} P_{b,col(e),1,k} \subseteq D_{b,k}$. The claim follows from Property 6 in this case.

Case 2. Finally consider sets $C_k(v,a)$ and $C_{k'}(w,b)$ where $e = (v,w)$ is an edge, v is on the left side, and w on the right. Then $C_k(v,a)$ contains elements of the form (e',i) where $i \in P_{f_{e'}(a),col(e'),a,k}$. $C_{k'}(w,b)$ contains elements of the form (e',j) where $j \in P_{b,col(e'),1,k'}$. The only possible elements in $C_k(v,a) \cap C_{k'}(w,b)$ are tuples with the first entry equal to e. Since $P_{f_e(a),col(e),a,k} \subseteq B_k$ and $P_{b,col(e),1,k'} \subseteq B_{k'}$ and $k \leq m/2, k' > m/2$, the claim follows from Properties 2 and 3 in this case. □

Lemma 2. *If there exists a way of labelling vertices of G satisfying all its edges then there exists a collection of $n'm$ sets in C which covers B.*

Proof. Let $label(v)$ denote the label given to vertex v by the above labelling. Consider the collection $C' \subset C$ comprising sets $C_1(v, label(v)) \ldots, C_{\frac{m}{2}}(v, label(v))$ for each vertex v on the left and sets $C_{\frac{m}{2}+1}(w, label(w)) \ldots, C_m(w, label(w))$ for each vertex w on the right. We show that these sets cover B. Since there are $m/2$ sets in C' per vertex, $|C'| = 2n' * \frac{m}{2} = n'm$.

Consider any edge $e = (v,w)$. It suffices to show that for every i, $1 \leq i \leq n'$, the tuple (e,i) in B is contained in either one of $C_1(v, label(v)) \ldots, C_{\frac{m}{2}}(v, label(v))$ or in one of $C_{\frac{m}{2}+1}(w, label(w)) \ldots, C_m(w, label(w))$. The key property we use is that $f_e(label(v)) = label(w)$.

Consider the partitions $P_{f_e(label(v)),col(e),label(v)}$ and $P_{label(w)),col(e),1}$. Since $f_e(label(v)) = label(w)$, the two partitions belong to the same group and subgroup. Since all partitions in a subgroup have the same right hand side, the element i must be present either in one of the sets $P_{label(w),col(e),label(v),k}$, where $k \leq m/2$, or in one of the sets $P_{label(w),col(e),1,k}$, where $k > m/2$. We consider each case in turn.

First, suppose $i \in P_{label(w),col(e),label(v),k}$, for some $k \leq m/2$. Then, from the definition of $C_k(v, label(v))$, $(e,i) \in C_k(v, label(v))$. Second, suppose $i \in P_{label(w),col(e),1,k}$, for some $k > m/2$. Then, from the definition of $C_k(w, label(w))$, $(e,i) \in C_k(w, label(w))$. The lemma follows. □

Lemma 3. *If the smallest collection C' of sets in C covering the base set B has size at most $\frac{\beta}{2} n' m \log n'$ then there exists a labelling of G which satisfies at least a $\frac{1}{32\beta^2 \log^2 n'}$ fraction of the edges. Recall that β was defined in Property 4 of $\mathcal{P}$.*

Proof. Given C', we need to demonstrate a labelling of G with the above property. For each vertex v, define $L(v)$ to be the collection of labels a such that $C_k(v,a) \in C'$ for some k. We think of $L(v)$ as the set of "suggested labels" for v given by C' and this will be a multiset in general. The labelling we obtain will ultimately choose a label for v from this set. It remains to show that there is a way of assigning each vertex v a label from $L(v)$ so as to satisfy sufficiently many edges.

We need some definitions. For an edge $e = (v,w)$, define $\#(e) = |L(v)| + |L(w)|$. Since the sum of the sizes of all $L(v)$s put together is at most $\frac{\beta}{2} n' m \log n'$ and since all vertices in G have identical degrees, the average value of $\#(e)$ is at most $\frac{\beta}{2} m \log n'$. Thus half the edges e have $\#(e) \leq \beta m \log n'$. We call these edges *good*.

We show how to determine a subset $L'(v)$ of $L(v)$ for each vertex v so that the following properties are satisfied. If v has a good edge incident on it then $L'(v)$ has size at most $4\beta \log n'$. Further, for each good edge $e = (v,w)$, there exists a label in $L'(v)$ and one in $L'(w)$ which together satisfy e. Clearly, random independent choices of labels from $L'(v)$ will satisfy a good edge with probability $\frac{1}{16\beta^2 \log^2 n'}$, implying a labelling which will satisfies at least a $\frac{1}{32\beta^2 \log^2 n'}$ fraction of the edges (since the total number of edges is at most twice the number of good edges), as required.

For each label $a \in L(v)$, include it in $L'(v)$ if and only if the number of sets of the form $C_*(v,a)$ in C' is at least $m/4$. Clearly, $|L'(v)| \leq \frac{\beta m \log n'}{m/4} = 4\beta \log n'$, for vertices v on which good edges are incident. It remains to show that for each good edge $e = (v,w)$, there exists a label in $L'(v)$ and one in $L'(w)$ which together satisfy e.

Consider a good edge $e = (v,w)$. Using Property 4 of $\mathcal{P}$, it follows that there exists a label $a \in L(v)$ and a label $b \in L(w)$ such that the $f_e(a) = b$ and the number of sets of the form $C_*(v,a)$ or $C_*(w,b)$ in C' is at least $3m/4$. The latter implies that the number of sets of the form $C_*(v,a)$ in C' must be at least $m/4$, and likewise for $C_*(w,b)$. Thus $a \in L'(v)$ and $b \in L'(w)$. Since $f_e(a) = b$, the claim follows. □

Corollary 1. *Set Cover with intersection 1 cannot be approximated within a factor of $\frac{\beta \log n'}{2}$ in random polynomial time, for some constant β, $0 < \beta \leq \frac{1}{6}$, unless $NP \subseteq ZTIME(n^{O(\log \log n)})$. Further, if the auxiliary system of partitions*

$\mathcal{P}$ can be constructed in deterministic polynomial (in n') time, then approximating to within a $\frac{\beta \log n'}{2}$ factor is possible only if $NP = DTIME(n^{O(\log\log n)})$.

5 Randomized Construction of the Auxiliary System $\mathcal{P}$

The obvious randomized construction is the following. Ignore the division into groups and just view $\mathcal{P}$ as a collection of subgroups. For each partition which is the first in its subgroup, throw each element i independently and uniformly at random into one of the m sets in that partition. For each partition P which is not the first in its subgroup, throw each element i which is not present in any of the sets on the right side of the first partition Q in this subgroup, into one of the first $m/2$ sets in P. Property 1 is thus satisfied directly. We need to show that Properties 2,3,4 are together satisfied with non-zero probability.

It can be shown quite easily that Property 4 holds with probability at least $1 - (\frac{1}{e})^{n'^{1-23\beta}}$, provided $\epsilon > 22\beta$. Slightly weak versions of Properties 2 and 3 (intersection bounds of 2 instead of 1) also follow immediately. This can be improved in the case of intersection 1 using the Lovasz Local Lemma, but this does not give a constant success probability and also leads to problems in derandomization. The details of these calculations appear in the full version.

To obtain a high probability of success, we need to change the randomized construction above to respect the following additional restriction (we call this Property 7): each set $P_{g,s,p,k}$ has size at most $\frac{d' * (deg+1) * dn'}{m}$, for all g, s, p, k, $1 \le g \le d', 1 \le s \le deg + 1, 1 \le p \le d, 1 \le k \le m$.

The new randomized construction proceeds as in the previous random experiment, fixing partitions in the same order as before, except that any choice of throwing an element $i \in N$ which violates Properties 2,3,7 is precluded. Property 7 enables us to show that not too many choices are precluded for each element, and therefore, this experiment stays close in behaviour to the previous one (provided $22\beta < \epsilon < 1/2$), except that Properties 2,3,7 are all automatically satisfied. The details appear in the full version.

6 Derandomization in $O(2^{n^{1-\epsilon}})$ Time

The main hurdle in derandomizing the above randomized construction in polynomial time is Property 4. There could be up to $O(2^{m \times polylog(n)}) = O(2^{(n')^{1-\epsilon'}})$ ways of choosing $\beta m \log n'$ sets from the various partitions in $\mathcal{P}$ for a constant ϵ' slightly smaller than ϵ, and we need that each of these choices fails to cover N for Property 4 to be satisfied.

For the Lund-Yannakakis system of partitions described in Section 2.1, each partition was into 2 sets and the corresponding property could be obtained deterministically using small-bias $\log n$-wise independent sample space constructions. This is no longer true in our case. Feige's [10] system of partitions, where each partition is into several but still a constant number of parts, can be obtained deterministically using anti-universal sets [17]. However, it is not clear how to apply either Feige's modified proof system or his system of partitions to get intersection 1.

We show in the full version that enforcing Property 4 in polynomial time corresponds to constructing hitting combinatorial rectangles with certain restricted kinds of sets, though we do not know any efficient constructions for them. In this paper, we take the slower approach of using Conditional Probabilities and enforcing Property 4 by checking each of the above choices explicitly. However, note that the number of choices is superexponential in n (even though it is sub-exponential in n'). To obtain a derandomization which is sub-exponential in n, we make the following change in $\mathcal{P}$: the base set is taken to be of size n instead of n'. We use an appropriate pessimistic estimator and conditional probabilities to construct $\mathcal{P}$ with parameter n instead of n' (details will be given in the full version). This will give a gap of $\Theta(\log n)$ (instead of $\Theta(\log n')$) in the set cover instance $\mathcal{SC}$). But since the base set size in $\mathcal{SC}$ is now $O((n' * n)\ \mathrm{polylog}(n))$, we get a hardness of only $\Theta(\log n) = \Theta(\frac{\log n'}{\log\log n'})$ (note that the approximation factor must be with respect to the base set size) unless $NP \subset DTIME(2^{n^{1-\epsilon}})$, for any constant ϵ such that $22\beta < \epsilon < 1/2$.

7 Open Problems

A significant contribution of this paper is that it leads to several open problems.

1. Is there a polynomial time algorithm for constructing the partition system in Section 4? In the full version, we show its relation to the question of construction of hitting sets for combinatorial rectangles with certain constraints. Can a hitting set for large volume combinatorial rectangles, with the property that any two hitting set points agree in at most one coordinate, be constructed in polynomial time? Alternatively, can a different proof system be obtained, as in [10], which will require a set system with weaker hitting properties?

2. Are there instances of the problem of covering points by lines, with an integrality gap of $\Theta(\log n)$? In the full version, we show that the an integrality gap of 2 and we describe a promising construction, which might have a larger gap.

3. Are there such explicit constructions for the the Set Covering problem with intersection 1? Randomized constructions are easy for this but we do not know how to do an explicit construction.

4. Is there a polynomial time algorithm for the problem of covering points with lines which has an $o(\log n)$ approximation factor, or can super-constant hardness (or even a hardness of factor 2) be proved? In the final version, we observe that the NP-Hardness proof of Megiddo and Tamir[16] can be easily extended to a Max-SNP-Hardness proof.

References

1. N. Alon, O. Goldreich, J. Hastad, R. Perralta. Simple Constructions of Almost k-Wise Independent Random Variables. *Random Structures and Algo-*

rithms, 3, 1992.
2. V.S. Anil Kumar and H. Ramesh. Covering Rectilinear Polygons with Axis-Parallel Rectangles. Proceedings of *31st ACM-SIAM Symposium in Theory of Computing*, 1999.
3. S. Arora, C. Lund. Hardness of Approximation. In *Approximation Algorithms for NP-Hard Problems*, Ed. D. Hochbaum, PWS Publishers, 1995, pp. 399-446.
4. S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy. Proof Verification and Intractability of Approximation Problems. Proceedings of *33rd IEEE Symposium on Foundations of Computer Science*, 1992, pp. 13-22.
5. S. Arora, M. Sudan. Improved Low Degree Testing and Applications. Proceedings of the *ACM Symposium on Theory of Computing*, 1997, pp. 485–495.
6. J. Beck. An Algorithmic Approach to the Lovasz Local Lemma I, *Random Structures and Algorithms*, 2, 1991, pp. 343-365.
7. M. Bellare, S. Goldwasser, C. Lund, A. Russell. Efficient Probabilistically Checkable Proofs and Applications to Approximation, Proceedings of *25th ACM Symposium on Theory of Computing*, 1993, pp. 294-303.
8. H. Brönnimann, M. Goodrich. Almost Optimal Set Covers in Finite VC-Dimension. *Discrete Comput. Geom.*, 14, 1995, pp. 263–279.
9. Y. Cheng, S.S. Iyengar and R.L. Kashyap. A New Method for Image compression using Irreducible Covers of Maximal Rectangles. *IEEE Transactions on Software Engineering*, Vol. 14, 5, 1988, pp. 651–658.
10. U. Feige. A threshold of $\ln n$ for Approximating Set Cover. *Journal of the ACM*, 45, 4, 1998, pp. 634–652.
11. R. Hassin and N. Megiddo. Approximation Algorithms for Hitting Objects with Straight Lines. *Discrete Applied Mathematics*, 30, 1991, pp. 29-42.
12. D.S. Johnson. Approximation Algorithms for Combinatorial Problems. *Journal of Computing and Systems Sciences*, 9, 1974, pp. 256–278.
13. C. Levcopoulos. Improved Bounds for Covering General Polygons by Rectangles. Proceedings of *6th Foundations of Software Tech. and Theoretical Comp. Sc.*, LNCS 287, 1987.
14. N. Linial, M. Luby, M. Saks, D. Zuckerman. Hitting Sets for Combinatorial Rectangles. Proceedings of *25 ACM Symposium on Theory of Computing*, 1993, pp. 286–293.
15. C. Lund, M. Yannakakis. On the Hardness of Approximating Minimization Problems. Proceedings of *25th ACM Symposium on Theory of Computing*, 1993, pp. 286–293.
16. N. Megiddo and A. Tamir, On the complexity of locating linear facilities in the plane, *Oper. Res. Let*, 1, 1982, pp. 194-197.
17. M. Naor, L. Schulman, A. Srinivasan. Splitters and Near-Optimal Derandomization. Proceedings of the *36th IEEE Symposium on Foundations of Computer Science*, 1995, pp. 182–191.
18. R. Raz. A Parallel Repetition Theorem. Proceedings of the *27th ACM Symposium on Theory of Computing*, 1995, pp. 447–456.
19. R. Raz and S. Safra. A Sub-Constant Error-Probability Low-Degree test and a Sub-Constant Error-Probability PCP Characterization of NP. Proceedings of the *ACM Symposium on Theory of Computing*, 1997, pp. 475–484.
20. Madhu Sudan. Personal communication.