A Threshold of $\ln(n)$ for approximating set cover

October 20, 2009

1 The k-prover proof system

- There is a single verifier V and k provers $P_1, P_2, \ldots, P_k$.
- Binary code with k codewords, each of length l and weight $\frac{l}{2}$, with Hamming distance at least $\frac{l}{3}$.
- Each prover is associated with a codeword.
- The Protocol:
  - The verifier selects l clauses $C_1, \ldots, C_l$, then selects a variable from each clause to form a set of distinguished variables $x_1, \ldots x_l$.
  - Prover P_i receives C_j for those coordinates in its codeword that are 1, and x_j for the coordinates that are 0, and replies with $2l$ bits.
  - The answer of the prover induces an assignment to the distinguished variables.
  - Acceptance predicate:
    - Weak: at least one pair of provers is consistent.
    - Strong: every pair of provers is consistent.

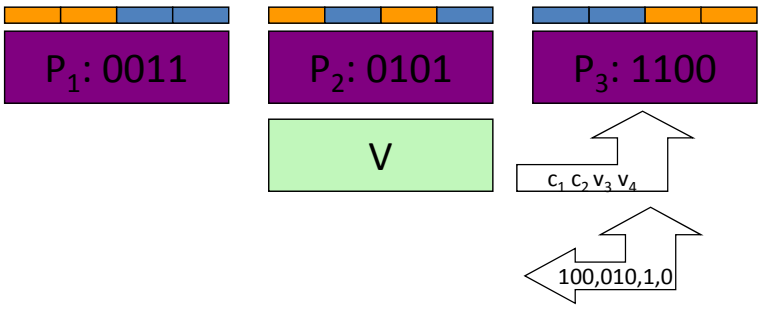

Figure 1: A k-prover proof system

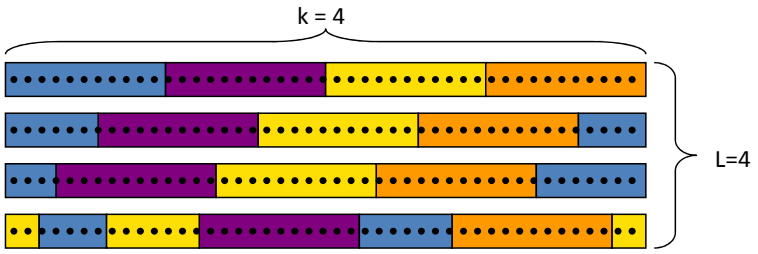

Figure 2: The partitioning system

Lemma 1 *Consider the k-prover proof system defined above and a 3CNF-5 formula ϕ. If ϕ is satisfiable, then the provers have a strategy that causes the verifier to always strongly accept. If at most a $(1-\epsilon)$-fraction of the clauses in ϕ are simultaneously satisfiable, then the verifier weakly accepts with probability at most $k^2.2^{-cl}$, where $c > 0$ is a constant that depends on ϵ.*

Proof:

- If ϕ is satisfiable, then provers can base their answers on a satisfying assignment.
- Assume $(1-\epsilon)$ clauses are satisfiable, and V weakly accepts with probability equal to δ, then with respect to P_i and P_j, V accepts with probability $\frac{\delta}{k^2}$.
- There are more than $l/6$ coordinates on which P_i receives a clause, and P_j receives a variable in that clause. Fix the question pairs in the other $5l/6$ coordinates in a way that maximizes the acceptance probability, which by averaging remains at least δ/k^2.
- The provers have a strategy that succeeds with $prob. = \delta/k^2$ on $l/6$ parallel repetitions of the original two-proof system.
- From the Theorem 1 stated below from the reference [30] it follows that $\frac{\delta}{k^2} < 2^{-cl}$, hence, $\delta \leq k^2.2^{-cl}$.

Theorem 1 *If a one-round two-prover proof system is repeated l times independently in parallel, then the error is 2^{-cl}, where $c > 0$ is a constant that depends only on the error of the original proof system (assuming this error was less than one) and on the length of the answers of the provers in the original proof system.*

2 Construction of a Partition System

Definition 1 *A partition system $B(m, L, k, d)$ has the following properties:*

- *There is a ground set B of m points.*
- *There is a collection of L distinct partitions $p_1, \ldots, p_L$.*

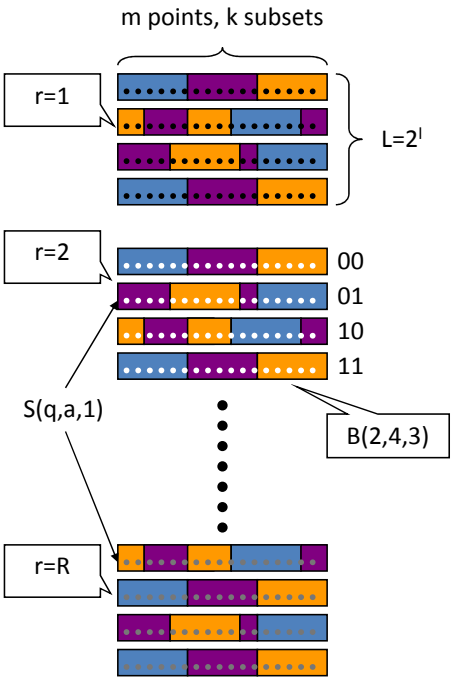

Figure 3: The Reduction to set cover

- *For $1 \leq i \leq L$, partition p_i is a collection of k disjoint subsets of B whose union is B.*
- *Any cover of the m points by subsets that appear in pairwise different partitions requires at least d subsets.*

There is a deterministic construction for such a partition system which is described in reference [27] of the paper.

3 The Reduction to Set Cover

- The verifier of the k-prover proof system uses its randomness,which is given in form of a random string r, to select l clauses and a distinguished variable in each clause. Call these distinguished variables the sequence of distinguished variables.
- The length of the random string is $(\log(5n/3) + \log(3))l = l\log(5n)$. Let $R = (5n)^l$, denote the number of possible random strings for the verifier.
- With each random string r, a distinct partition system $B_r(m, L, k, d)$ is associated, where $L = 2^l$, $m = n^{\Theta(l)}$ and $d = (1 - f(k))k \ln m$. (Altogether there are $N = mR$ points in the set cover problem.)

- Each of the L partitions is labeled by an l-bit string p, that corresponds to an assignment to the respective sequence of distinguished variables. Each subset in a partition is labeled by a unique prover i.
- We let $B(r, j, i)$ denote the i^{th} subset of partition j in partition system r.
- With each question-answer pair (q, a) of prover P_i, where $1 \leq i \leq k$, a subset $S(q, a, i)$ is associated as follows:
  - The notation $(q, i) \in r$ is used to say that on random string r, prover P_i receives question q.
  - For r such that $(q, i) \in r$, consider the induced sequence of distinguished variables, and extract from a on coordinate by coordinate basis an assignment a_r to this sequence of variables.
  - One of the partitions of partition system $B_r(m, L, k, d)$ has label a_r.
  - The subset $S(q, a, i)$ contains the points of subset $B(r, a_r, i)$, for all r with $(q, i) \in r$.
- Let Q denote the number of possible different questions that a prover may receive. A question to a single prover includes $l/2$ variables, for which there are $n^{l/2}$ possibilities (with repetition), and $l/2$ clauses, for which there are $(5n/3)^{l/2}$ possibilities. Hence, $Q = n^{l/2}.(5n/3)^{l/2}$. Observe that this number is the same for all provers.

Lemma 2 *If ϕ is satisfiable, then the above set of $N = mR$ points can be covered by kQ subsets. If only a $(1-\epsilon)$ fraction of the clauses in ϕ are simultaneously satisfiable, the above set requires $(1 - 2f(k))kQ\ln(m)$ subsets in order to be covered, where $f(k) \to 0$ as $k \to \infty$.*

Proof:

- **Completeness:** If all the clauses in ϕ are satisfiable then there is a set cover of size kQ.
- **Soundness:** If a fraction $(1 - \epsilon)$ of all the clauses of ϕ are satisfiable then more than $(1 - 2f(k))kQ\ln(m)$ subsets are required.

Proof of Completeness:

- If all the clauses in ϕ are satisfiable then, the provers answer consistently with the satisfying assignment.
- For any r, consider $S(q_1, a_1, r), \ldots, S(q_k, a_k, r)$ s.t. $(q_i, i) \in r$, and a_i is the appropriate answer.
- $B_r(m, L, k, d)$ is covered by these k sets.
- Similar for every r. The number of subsets is kQ.

Proof of Soundness:

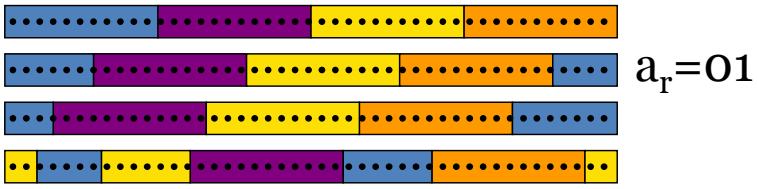

Figure 4: Proof of Completeness

- Assume a fraction $(1-\epsilon)$ of all the clauses of ϕ are satisfiable, and that there exists C that covers U, such that $|C| = (1-\delta)kQ\ln(m), \delta = 2f(k)$.
- With each question q to a prover P_i associate a weight $w_{q,i}$ equal to the number of answers a such that $S(q,a,i) \in C$. Hence, $\sum_{(q,i)} w_{q,i} = |C|$.
- With each random string r associate a weight $w_r = \sum_{(q,i)\in r} w_{q,i}$. This weight is equal to the number of subsets that participate in covering the m points of $B_r(m,L,k,d)$.
- Call r good if $w_r < (1-\delta/2)k\ln(m)$.

Proposition 1 *The fraction of good r is at least $\delta/2$.*

Proof:

Assume otherwise, then $\sum_r w_r \geq (1-\delta/2)^2 kR\ln m > (1-\delta)kR\ln m$, where R denotes the number of possible random strings of the verifier. On the other hand we have:

$\sum_r w_r = \sum_r \sum_{(q,i)\in r} w_{q,i} = \sum_{q,i} \frac{R}{Q} w_{q,i} = \frac{R}{Q}|C|$,

where the middle equality follows from the fact that there are exactly $\frac{R}{Q}$ random strings that cause the verifier to send out question q. Hence, $|C| > (1-\delta)kQ\ln m$, contradiction.

Proposition 2 *Let C be a collection of subsets that covers S, where $|C| = (1-\delta)kQ\ln(m)$. Then for some strategy for the k-provers, the verifier accepts ϕ with probability at least $2\delta/(k\ln(m))^2$.*

This proves the soundness, since $2\delta/(k\ln(m))^2 > k^2 2^{-cl}$, for $l = \Theta(\log\log n)$.

Proof:

- Based on $|C|$, we describe a randomized strategy for the k provers. On question q addressed to prover P_i, prover P_i selects an answer a uniformly at random from the set of answers that satisfy $S(q,a,i) \in C$. We show that under this strategy for the provers, the verifier weakly accepts with probability at least $2d/(k\ln m)^2$.
- Observe that for a fixed r, there is a one to one correspondence between sets $B(r,p,i)$ that participate in the cover of $B_r(m,L,k,d)$ and sets $S(q,a,i)$ that belong to C. For this correspondence we need $(q,i) \in r$ and the projection of a on the sequence of distinguished variables to be p.

- Concentrate now only on good r, and compute a lower bound on the probability that the verifier accepts when he chooses a good r. Observe that by Property (4) of partition systems, and by the fact that for good r the respective $B_r(m, L, k, d)$ is covered by at most $(1 - \delta/2)k \ln(m)$ subsets, the cover C must have used two subsets from the same partition p in the cover of $B_r(m, L, k, d)$.

- Denote these two subsets by $B(r, p, i)$ and $B(r, p, j)$, where $i \neq j$, and their corresponding subsets in C by $S(q_i, a_i, i)$ and $S(q_j, a_j, j)$, respectively.

- Consider what happens when the verifier chooses random string r. The Prover P_i then receives question q_i and prover P_j receives question q_j. Let $A_{r,i}$ denote the set of answers satisfying $a \in A_{r,i}$ if and only if $S(q_i, a, i) \in C$. By the strategy described above, prover P_i selects an answer $a \in A_{r,i}$ at random (and P_j selects $a \in A_{r,j}$). Observe that for a_i and a_j above, $a_i \in A_{r,i}$ and $a_j \in A_{r,j}$, and furthermore, for good r, $|A_{r,i}| + |A_{r,j}| < k \ln(m)$. Hence, the joint probability that the provers choose to answer with a_i and a_j is at least $4/(k \ln m)^2$. Since both these answers are consistent with the label p of the same partition, the verifier weakly accepts. To complete the proof, use previous Proposition , which shows that the probability that a verifier chooses a good r is at least $\delta/2$.

Theorem 2 *If there is some $\epsilon > 0$ such that a polynomial time algorithm can approximate set cover in $(1 - \epsilon) \ln(n)$, then $NP \subset TIME(n^{O(\log \log(n))})$.*

Proof:

- Lets Assume that there is a polynomial time algorithm A that approximates set cover within $(1 - \epsilon) \ln n$.

- Now follow the reduction to set cover described above, with k sufficiently large so that $f(k)$ is smaller than $\epsilon/4$, and with $m = (5n)^{2l/\epsilon}$.

- Using the deterministic construction of partition systems described in reference [27], and observing that m, R and Q are bounded by $n^{O(\log \log n)}$, the time to perform this reduction is $n^{O(\log \log n)}$.

- Recall that the number of points in the set cover problem is $N = mR$ where $R = (5n)^l$, and observe that for m as above, $\ln m > (1 - \epsilon/2) \ln N$.

- By Lemma 2, if the original NP instance was satisfiable, all points can be covered by kQ subsets, and if the original NP instance was not satisfiable, all points cannot be covered by $(1 - 2f(k))kQ \ln m$. For our choice of k and m, the ratio between the two cases is $(1 - 2f(k)) \ln m > (1 - \epsilon) \ln N$.