Vertex Cover Might be Hard to Approximate to within $2 - \varepsilon$

Subhash Khot * Oded Regev †

Abstract

We show that vertex cover is hard to approximate within any constant factor better than 2 where the hardness is based on a conjecture regarding the power of unique 2-prover-1-round games presented in [15]. We actually show a stronger result, namely, based on the same conjecture, vertex cover on k-uniform hypergraphs is hard to approximate within any constant factor better than k.

*Department of Computer Science, Princeton University, Princeton, NJ 08544. E-Mail: khot@cs.princeton.edu. Supported by Prof. Sanjeev Arora's David and Lucile Packard Fellowship, NSF Grant CCR-0098180, and an NSF ITR grant

†Institute for Advanced Study, Princeton, NJ. E-Mail: odedr@ias.edu. Research supported by NSF grant CCR-9987845.

1 Introduction

Minimum vertex cover is the problem of finding the smallest set of vertices that touches all the edges in a given graph. This is one of the most fundamental NP-complete problems. A simple 2-approximation algorithm exists for this problem: construct a maximal matching by greedily adding edges and then let the vertex cover contain both endpoints of each edge in the matching. It can be seen that the resulting set of vertices indeed touches all the edges and that its size is at most twice that of the minimum vertex cover. However, despite considerable efforts, state of the art techniques can only achieve an approximation ratio of $2 - o(1)$ [10].

Given this state of affairs, one might strongly suspect that vertex cover is NP-hard to approximate within $2-\varepsilon$ for any $\varepsilon > 0$. This is one of the major open questions in the field of approximation algorithms. In [11], Håstad showed that approximating vertex cover within constant factors less than $\frac{7}{6}$ is NP-hard. This was recently improved by Dinur and Safra [5] to 1.36. In a related result, Arora et al. [1] considered algorithms based on linear programming. They showed an integrality gap of $2 - \epsilon$ for a large family of linear programs for vertex cover. This implies that many linear programming based algorithms cannot obtain an approximation ratio better than 2.

In this paper we consider the more general problem of vertex cover on k-uniform hypergraphs. A k-uniform hypergraph $H = (V, E)$ consists of a set of vertices V and a collection E of k-element subsets of V called hyperedges (or simply edges). A *vertex cover* of H is a subset of vertices $S \subseteq V$ such that every hyperedge in E intersects S, i.e., $e \cap S \neq \emptyset$ for each $e \in E$. An *independent set* in H is a subset whose complement is a vertex cover, or in other words a subset of vertices that contains no hyperedge entirely within it. The Ek-Vertex-Cover problem is the problem of finding a minimum size vertex cover in a k-uniform hypergraph. Notice that for $k = 2$, this problem is equivalent to the vertex cover problem on graphs. The simple algorithm presented before can be easily extended to k-uniform hypergraphs, achieving a factor k approximation. However, as before, the best approximation algorithms yield only a tiny improvement, achieving a $k - o(1)$ approximation ratio [10].

The first explicit hardness result shown for Ek-Vertex-Cover was due to Trevisan [17] who showed (among other results) an inapproximability factor of $k^{1/19}$. Holmerin [14] showed that E4-Vertex-Cover is NP-hard to approximate within $(2-\varepsilon)$, and more recently [13] that Ek-Vertex-Cover is NP-hard to approximate within $k^{1-\varepsilon}$, and also that it is NP-hard to approximate E3-Vertex-Cover within factor $(3/2-\varepsilon)$. Goldreich [8] showed a direct 'FGLSS'-type [6] reduction (involving no use of the long-code, a crucial component in most recent PCP constructions) attaining a hardness factor of $(2 - \varepsilon)$ for Ek-Vertex-Cover for some constant k. Dinur, Guruswami and Khot [3] gave a fairly simple proof of an $\Omega(k)$ hardness result for Ek-Vertex-Cover and a more complicated proof that shows a factor $(k - 3 - \varepsilon)$ hardness for Ek-Vertex-Cover. A recent paper by Dinur et al. [4] improves upon all previous results by showing a $(k - 1 - \varepsilon)$ hardness result.

With this recent progress on the Ek-Vertex-Cover problem, there is a strong reason to believe that it is NP-hard to approximate Ek-Vertex-Cover within $k - \epsilon$ for every $k \geq 2$. The current techniques, however, seem very inadequate to prove such a result. In [15], Khot presented the *unique games conjecture* as an approach to attack many fundamental open problems. The conjecture deals with 2-prover-1-round games where two (all-powerful) provers try to convince a probabilistic verifier that a certain NP-statement is true. The proof system is 1-round, meaning the verifier asks both the provers one question each and accepts or rejects depending on the provers' answers. The game is called *unique* if the answer of one prover completely determines the answer of the second prover and vice versa. The conjecture essentially states that it is NP-hard to distinguish whether the provers' optimal strategy in a unique 2-prover-1-round game has success probability very close to 1 or it is negligible. Assuming this conjecture, Khot was able to show several new hardness results including the hardness of Min-2SAT-Deletion and hardness of the Not-All-Equal predicate on 3 variables. He also observed that a variation of his conjecture would imply a $\sqrt{2} - \varepsilon$ hardness result for vertex cover.

In this paper, we continue this line of research and assuming the unique games conjecture, we prove a tight $k - \varepsilon$ hardness result for Ek-Vertex-Cover. In particular, this gives a $2 - \varepsilon$ hardness result for vertex cover, giving a further evidence that factor 2 may be the right answer for this problem.

Main techniques: Many of the recent hardness results are shown via constructions of new Probabilistically Checkable Proof systems (PCPs) (see, e.g., [2, 11, 12, 9]). These constructions typically involve two modules, the so-called *Outer PCP* and the so-called *Inner PCP*. The Outer PCP is essentially a 2-prover-1-round game and the Inner PCP is based on long codes and often, the Fourier analysis of long codes.

However, this *standard recipe* hasn't been very successful in attacking the vertex cover problem. Håstad's $\frac{7}{6}$ hardness remained the best known result for a long time. Dinur and Safra [5] were able to break this barrier by relying on techniques from extremal combinatorics. However, their approach still doesn't succeed in getting a hardness factor better than 1.36. Khot [15] observed that the bottleneck in getting hardness results for vertex cover and a number of other problems might be in the Outer PCP which has remained untouched so far. His conjecture basically states that a strong enough Outer PCP exists and then one can build Inner PCPs on top of it yielding the desired hardness of approximation results.

Khot's conjecture looks quite promising, at least in light of the lack of any other techniques. We think that it is worthwhile to investigate which problems could be solved via this conjecture and we show that vertex cover is one such problem. We combine this conjecture with the techniques from Dinur and Safra's paper [5], which include the biased long code, Friedgut's Theorem on sensitivity of Boolean functions and theorems in extremal set theory.

It turns out that one needs the unique games conjecture in a stronger form than what is stated in [15]. A significant contribution of this paper is to show that the stronger form follows from the original form. Roughly speaking, the original form states that in the *good case* the provers in the 2-prover-1-round game have a strategy that convinces the verifier with probability close to 1. The stronger form states that the provers in fact have a strategy such that with probability very close to 1, after fixing the question to the first prover, the verifier accepts for *every* question to the second prover.

We believe that this stronger form would be useful in our understanding of the unique games conjecture and an eventual resolution of this conjecture.

Overview of the paper: In Section 2, we introduce the tools from sensitivity analysis of set-families and extremal combinatorics. Section 3 explains the reduction to the stronger form of the conjecture and it is the crux of the paper. Section 4 explains the reduction to hypergraph vertex cover and shares many ideas with previous work such as [5] and [4].

2 Preliminaries

For a universe R, let $P(R)$ denote its power set, i.e., the family of all subsets of R. For a "bias parameter" $0 < p < 1$, the weight $\mu_p^R(F)$ of a set F is defined as

$$\mu_p^R(F) \stackrel{def}{=} p^{|F|}(1-p)^{|R \setminus F|}.$$

We will omit the superscript R when it is clear which universe we are talking about. The weight of a family $\mathcal{F} \subseteq P(R)$ is defined as

$$\mu_p(\mathcal{F}) \stackrel{def}{=} \sum_{F \in \mathcal{F}} \mu_p(F).$$

Note that the bias parameter defines a distribution on $P(R)$, where a subset is picked by independently picking every element in R with probability p. We denote this distribution by μ_p^R.

2.1 Friedgut's 'Core' Theorem

For a family $\mathcal{F}$, an element $\sigma \in R$ and a bias parameter p we define the "influence of the element on the family" as

$$\text{Influence}_p^R(\mathcal{F}, \sigma) \stackrel{def}{=} \Pr_{F \in \mu_p^R}[\text{exactly one of } F \cup \{\sigma\},\ F \setminus \{\sigma\} \text{ is in } \mathcal{F}].$$

As before, the superscript will often be omitted. The average sensitivity of a family is defined as sum of the influences of all elements.

$$\text{as}_p(\mathcal{F}) \stackrel{def}{=} \sum_{\sigma \in R} \text{Influence}_p(\mathcal{F}, \sigma).$$

Definition 2.1 *A family $\mathcal{F} \subseteq P(R)$ is called a core-family with a core $C \subseteq R$ if there exists a family $\mathcal{H} \subseteq P(C)$ such that*

$$\forall\ F \in P(R), \quad F \in \mathcal{F} \text{ iff } F \cap C \in \mathcal{H}.$$

A family $\mathcal{F} \subseteq P(R)$ is called *monotone* if $F \in \mathcal{F}$ and $F \subseteq F'$ implies $F' \in \mathcal{F}$. The following lemma is well-known (see, e.g., [5]).

Lemma 2.2 *If $\mathcal{F} \subseteq P(R)$ is monotone and $p \geq q$, then $\mu_p^R(\mathcal{F}) \geq \mu_q^R(\mathcal{F})$.*

We will use the following theorem that can be obtained by combining Russo's Theorem and Friedgut's Theorem. This theorem essentially says that a monotone family of subsets is well-approximated by a core-family with a "small" core.

Theorem 2.3 ([7, 16]) *Let p be a bias parameter, $\varepsilon, \delta > 0$ be constants and η be an "accuracy parameter". Let $\mathcal{F} \subseteq P(R)$ be a monotone family such that $\mu_p(\mathcal{F}) \geq \delta$. Then there exists $p' \in (p, p+\varepsilon)$ and a core family $\widehat{\mathcal{F}} \subseteq P(R)$ with a core $C \subseteq R$ such that*

- *The average sensitivity of the family $\mathcal{F}$ w.r.t the bias p' is at most $\frac{1}{\varepsilon}$, i.e., $as_{p'}(\mathcal{F}) \leq \frac{1}{\varepsilon}$.*
- *The size of C is a constant that depends only on $p, \delta, \varepsilon, \eta$.*
- *$\mu_{p'}(\mathcal{F} \Delta \widehat{\mathcal{F}}) < \eta$ where Δ denotes the symmetric difference of the two families.*

We will need the following two lemmas. The proof of the first lemma is in the appendix. The second lemma follows from a theorem of Frankl and can be found as Lemma A.4 in [3].

Lemma 2.4 *Let $\mathcal{F} \subseteq P(R)$ be a monotone family. Let T be a set of elements such that for every element $\sigma \in T$, $\text{Influence}_p(\mathcal{F}, \sigma) < \eta$ where η is any constant. Define a subfamily $\mathcal{F}'$ of the family $\mathcal{F}$ as*

$$\mathcal{F}' \stackrel{def}{=} \{F \mid F \in \mathcal{F},\ F \setminus T \in \mathcal{F}\}.$$

Then for any $0 < p < 1$ we have

$$\mu_p(\mathcal{F}') \geq \mu_p(\mathcal{F}) - \eta\ |T|\ (\min(p, 1-p))^{-|T|}.$$

Lemma 2.5 *Let $\varepsilon > 0$ be an arbitrarily small constant and define $p = 1 - \frac{1}{k} - \varepsilon$ to be the bias parameter. Then, for a sufficiently large universe R, the following holds. For any $\mathcal{F} \subseteq P(R)$ such that $\mu_p(\mathcal{F}) \geq 1 - \frac{1}{k}$ there exist k sets in the family $\mathcal{F}$ whose intersection is empty.*

3 Constructing the Strong Label Cover

For the ease of presentation, we will talk in terms of the Label Cover problem instead of 2-Prover-1-Round Games. Let R be some large enough label set. A Unique-Label-Cover (Unique-LC) Φ is a tuple (X, Y, Π, W) where X and Y are two sets of variables. In addition, for every $x \in X$ and $y \in Y$ there exists a bijection test $\pi_{xy} \in \Pi$, i.e., a one-to-one and onto function $\pi_{xy} : R \to R$ and a weight $w_{xy} \geq 0$. We denote by $w(\Phi, x)$ the sum $\sum_{y \in Y} w_{xy}$ and by $w(\Phi)$ the sum $\sum_{x \in X, y \in Y} w_{xy}$. An assignment is a function $A : X \cup Y \to R$. A test π_{xy} is satisfied by an assignment A if $\pi_{xy}(A(x)) = A(y)$. Also, for an assignment A, the weight of satisfied tests, denoted by $w_A(\Phi)$, is $\sum w_{xy}$ where the sum is taken over all $x \in X$ and $y \in Y$ such that π_{xy} is satisfied by A. Similarly, we define $w_A(\Phi, x)$ as $\sum w_{xy}$ where the sum is now taken over all $y \in Y$ such that π_{xy} is satisfied by A. The following conjecture was presented in [15]:

Conjecture 3.1 *For any $\zeta, \gamma > 0$ there exists a constant $|R|$ such that the following is NP-hard. Given a Unique-LC Φ with label set R and $w(\Phi) = 1$ distinguish between the case where there exists an assignment A such that $w_A(\Phi) \geq 1 - \zeta$ and the case where for any assignment A, $w_A(\Phi) \leq \gamma$.*

A Strong-Label-Cover (Strong-LC) $\Phi = (X, Y, E, \Pi)$ is defined as follows. We are given a bipartite graph (X, Y, E) possibly with parallel edges in which all left degrees are equal to some constant d. In addition, there exists a bijection test $\pi_e \in \Pi$ for each edge $e \in E$. An assignment and the tests satisfied by it are defined as before. In this section we prove the following theorem:

Theorem 3.2 *Assuming Conjecture 3.1, for any $\zeta, \gamma > 0$ there exist constants $|R|, d$ such that the following is NP-hard. Given a Strong-LC with label set R and left degrees d distinguish between the case where there exists an assignment in which at least $1 - \zeta$ fraction of the X vertices have all of their tests satisfied and the case where no assignment satisfies more than γ of the edges.*

We begin with two lemmas. The first modifies the Unique-LC so that all X variables have the same weight. The second lemma rounds the edge weights.

Lemma 3.3 *Assuming Conjecture 3.1, for any $\zeta, \gamma, \beta > 0$ there exists a constant $|R|$ such that the following is NP-hard. Given a Unique-LC Φ with label set R and the property that $\forall x \in X,\ w(\Phi, x) = 1$, distinguish between the case where there exists an assignment A such that $1 - \beta$ of the X variables have $w_A(\Phi, x) > 1 - \zeta$ and the case where for any assignment A at most β of the X variables have $w_A(\Phi, x) > \gamma$.*

Proof: Consider a Unique-LC $\Phi' = (X', Y', \Pi', W')$ as given by Conjecture 3.1 with parameters ζ', γ' which will be chosen later. Also, let l be a large enough constant. The Unique-LC $\Phi = (X, Y, \Pi, W)$ is defined as follows. The set Y is equal to Y'. The set X includes $k(x)$ copies of each $x \in X'$, $x^{(1)}, \ldots, x^{(k(x))}$ where $k(x)$ is defined as $\lfloor l \cdot |X'| \cdot w(\Phi', x) \rfloor$. For every $x \in X'$, $y \in Y$ and $i \in [k(x)]$ we define $\pi_{x^{(i)}y}$ as π'_{xy} and the weight $w_{x^{(i)}y}$ as $w'_{xy}/w(\Phi', x)$. Notice that $w(\Phi, x) = 1$ for all $x \in X$ and that $(l-1)|X'| \leq |X| \leq l|X'|$.

We first prove the completeness part. Given an assignment A' to Φ' that satisfies tests of weight at least $1 - \zeta'$, consider the assignment A defined as $A(x^{(i)}) = A'(x)$. The weight of tests satisfied by A is:

$$\sum_{x\in X} w_A(\Phi,x) = \sum_{x\in X'} k(x)\cdot w_{A'}(\Phi',x)/w(\Phi',x) \geq$$
$$\sum_{x\in X'} l|X'|\cdot w_{A'}(\Phi',x) - \sum_{x\in X'} w_{A'}(\Phi',x)/w(\Phi',x) \geq$$
$$l|X'|(1-\zeta') - |X'| = l|X'|(1-\zeta'-\frac{1}{l}) \geq l|X'|(1-2\zeta')$$

for large enough l. This implies that for at least $1-\sqrt{2\zeta'}$ of the X variables $w_A(\Phi,x)\geq 1-\sqrt{2\zeta'}$ (for otherwise, $\sum_{x\in X} w_A(\Phi,x) < (1-\sqrt{2\zeta'})|X| + \sqrt{2\zeta'}|X|(1-\sqrt{2\zeta'}) \leq (1-2\zeta')l|X'|$). Hence, by choosing a small enough ζ', we get that at least $1-\beta$ of the X variables satisfy $w_A(\Phi,x)\geq 1-\zeta$.

We now prove the soundness part. Assume we are given an assignment A to Φ for which β of the variables have $w_A(\Phi,x) > \gamma$. Without loss of generality we can assume that for every $x\in X'$, the assignment $A(x^{(i)})$ is the same for all i. That is because the tests between $x^{(i)}$ and the Y variables are the same for all $i\in[k(x)]$. We define the assignment A' as $A'(x) = A(x^{(1)})$. The weight of tests satisfied by A' is:

$$\sum_{x\in X'} w_{A'}(\Phi',x) \geq \frac{1}{l|X'|}\sum_{x\in X'} k(x)\cdot w_{A'}(\Phi',x)/w(\Phi',x) =$$
$$\frac{1}{l|X'|}\sum_{x\in X} w_A(\Phi,x) \geq \frac{1}{l|X'|}\beta|X|\gamma \geq \frac{l-1}{l}\beta\gamma > \gamma',$$

for small enough γ'. ■

Lemma 3.4 *Assuming Conjecture 3.1, for any $\zeta,\gamma,\beta>0$ there exists a constant $|R|$ such that the following is NP-hard. We are given a Unique-LC Φ with label set R and the properties that $\forall x\in X$, $w(\Phi,x)=1$ and that there exists an integer $\alpha = O(|Y|)$ such that the weight w_{xy} is divisible by $\frac{1}{\alpha}$ for all $x\in X$, $y\in Y$. The goal is to distinguish between the case where there exists an assignment A such that $1-\beta$ of the X variables have $w_A(\Phi,x) > 1-\zeta$ and the case where for any assignment A at most β of the X variables have $w_A(\Phi,x)>\gamma$.*

Proof: Let $\Phi' = (X',Y',\Pi',W')$ be a Unique-LC as in Lemma 3.3 with parameters $\frac{\zeta}{2},\frac{\gamma}{2},\beta$. Let $l\in\mathbb{Z}$ be a large enough constant and define $\alpha = l|Y|$. The Unique-LC $\Phi=(X,Y,\Pi,W)$ has $X=X'$, $Y=Y'$, $\Pi=\Pi'$ and W is defined as follows. Fix an arbitrary Y variable y_0. For every $x\in X$ we would like to round the weights w_{xy} down to the nearest multiple of $\frac{1}{\alpha}$. In order to maintain the property $w(\Phi,x)=1$, we slightly increase w_{xy_0}. More formally, for every $x\in X$ and $y\neq y_0$ the weight w_{xy} is defined as $\lfloor \alpha w'_{xy}\rfloor/\alpha$. Also, for every $x\in X$, w_{xy_0} in defined as $1-\sum_{y\in Y\setminus\{y_0\}} w_{xy}$. Notice that for $y\neq y_0$, $w_{xy}\leq w'_{xy}$ and that $w_{xy_0}\leq w'_{xy_0}+|Y|/\alpha = w'_{xy_0}+\frac{1}{l}$.

Assume that there exists an assignment A to Φ' in which $1-\beta$ of the X' variables have $w_A(\Phi',x) > 1-\frac{\zeta}{2}$. Then, in Φ, the same assignment satisfies that for $1-\beta$ of the X variables $w_A(\Phi,x) > 1-\frac{\zeta}{2} - |Y|\cdot\frac{1}{l|Y|} > 1-\zeta$ for large enough l. Also, an assignment A to Φ in which β of the X variables have $w_A(\Phi,x)>\gamma$ satisfies that β of the X' variables have $w_A(\Phi',x) > \gamma-\frac{1}{l} > \frac{\gamma}{2}$ for large enough l. ■

Proof of Theorem 3.2: Let $\Phi'=(X',Y',\Pi',W')$ be a Unique-LC as in Lemma 3.4 with parameters ζ',γ',β' which will be chosen later. We define the Strong-LC $\Phi=(X,Y,E,\Pi)$ as follows. The set of vertices Y equals Y'. Let $d\in\mathbb{Z}$ be a constant that will be determined later. For each $x\in X'$

and each sequence $(y_1, \ldots, y_d)$ of Y vertices we create $\Pi_{i=1}^d \alpha w'_{xy_i}$ new vertices in X (notice that this number is integral). Each of these vertices is connected to $y_1, \ldots, y_d$ with the tests $\pi'_{xy_1}, \ldots, \pi'_{xy_d}$. The total number of vertices created from each $x \in X'$ is

$$\sum_{(y_1,\ldots,y_d)\in Y^d} \Pi_{i=1}^d \alpha w'_{xy_i} = \alpha^d (\sum_{y\in Y} w'_{xy})^d = \alpha^d$$

since $w(\Phi', x) = 1$. Hence, $|X| = \alpha^d |X'|$. Also note that Φ might contain parallel edges.

We first prove the completeness part. Assume that A' is an assignment to Φ' such that $1 - \beta'$ of the X' vertices have $w_{A'}(\Phi', x) > 1 - \zeta'$. Let A be the assignment to Φ assigning to each of the vertices created from $x \in X'$ the value $A'(x)$ and for each $y \in Y$ the value $A'(y)$. Consider a variable $x \in X'$ such that $w_{A'}(\Phi', x) > 1 - \zeta'$ and let Y_x denote the set of variables $y \in Y$ such that π'_{xy} is satisfied. Then the number of vertices in X that are connected only to vertices in Y_x is

$$\sum_{(y_1,\ldots,y_d)\in (Y_x)^d} \Pi_{i=1}^d \alpha w'_{xy_i} = \alpha^d (\sum_{y\in Y_x} w'_{xy})^d \geq \alpha^d (1-\zeta')^d.$$

Therefore, the total number of vertices all of whose tests are satisfied by A is at least

$$\alpha^d (1-\zeta')^d (1-\beta')|X'| = (1-\zeta')^d(1-\beta')|X| > (1-\zeta)|X|$$

for small enough ζ' and β'.

We now prove the soundness part. Assume that no assignment A' to Φ' has more than β' of the X' vertices with $w_{A'}(\Phi', x) > \gamma'$. Let A be an assignment to Φ and define A' as follows. For each $y \in Y$ let $A'(y) = A(y)$. For $x \in X'$ define $A'(x)$ as the value in R that maximizes $w_{A'}(\Phi', x)$. For $x \in X'$ and $i \in R$, let $u_{x,i}$ be $\sum w'_{xy}$ where the sum is taken over all $y \in Y$ such that $A'(y) = \pi_{x,y}(i)$. Then notice that $w_{A'}(\Phi', x) = \max_i u_{x,i}$. Hence, for at least $1 - \beta'$ of the X' vertices, $u_{x,i} < \gamma'$ for all $i \in R$. Fix a vertex $x \in X'$ for which $u_{x,i} < \gamma'$ for all $i \in R$. Consider the subset $Z_x \subseteq Y^d$ of tuples $(y_1, \ldots, y_d)$ such that for all $i \in R$ there exists at most one j for which $\pi_{x,y_j}(i) = A'(y_j)$. A vertex in X created from such a tuple will have at most one satisfied edge. The number of such vertices created from x as above is

$$\sum_{(y_1,\ldots,y_d)\in Z_x} \Pi_{i=1}^d \alpha w'_{xy_i} = \alpha^d \sum_{(y_1,\ldots,y_d)\in Z_x} \Pi_{i=1}^d w'_{xy_i}.$$

Notice that since $\sum_{y\in Y} w'_{xy} = 1$, this defines a probability measure on Y. Also note that the sum above is exactly the probability that a tuple $(y_1, \ldots, y_d)$ is in Z_x where each element of the tuple is chosen according to this probability. Hence, the sum is at least $1\cdot(1-\gamma')\cdot(1-2\gamma')\ldots(1-(d-1)\gamma') \geq (1-d\gamma')^d > 1 - \frac{\gamma}{2}$ for small enough γ'. The number of satisfied edges in A is therefore at most

$$\alpha^d \cdot \beta' \cdot |X'| \cdot d + \alpha^d (1-\beta')|X'|[(1-\frac{\gamma}{2})\cdot 1 + \frac{\gamma}{2}\cdot d] =$$
$$|X|(\beta' d + (1-\beta')(1-\frac{\gamma}{2}) + (1-\beta')\frac{\gamma}{2}d) < \gamma|X|d = \gamma|E|$$

for a small enough β' and a large enough d. ■

4 Reduction to Vertex Cover in k-Uniform Hypergraphs

Let $\Phi = (X, Y, E, \Pi)$ be an instance of Strong-LC given by Theorem 3.2 with parameters ζ, γ which will be chosen later. We will reduce this instance to an Independent Set problem on k-uniform hypergraphs. The vertices of the hypergraph we construct are weighted. One can obtain an unweighted hypergraph by using standard techniques (see, e.g., [5]). The hypergraph will either contain an independent set of weight $1-\frac{1}{k}-2\varepsilon$ or no independent set of weight δ where ε, δ can be made arbitrarily small. In the following we fix ε and δ and let $p = 1-\frac{1}{k}-\varepsilon$ be the bias parameter.

4.1 Construction of the hypergraph

The set of vertices of the hypergraph will correspond to the bits of the long codes of labels assigned to variables in X. Namely, the set of vertices is defined to be $X \times P(R)$. A vertex is a pair (x, F) where $x \in X$ is a variable of the Strong-LC and $F \in P(R)$ is a subset of R. We define the *block* of a variable $x \in X$ as the set of vertices that correspond to x, i.e.,

$$B[x] \stackrel{def}{=} \{(x, F) \mid F \subseteq R\}.$$

The weight of a vertex (x, F) is defined to be

$$weight(x, F) \stackrel{def}{=} \frac{1}{|X|} \cdot \mu_p^R(F).$$

Thus the sum of the weights of all the vertices in the hypergraph equals 1.

Now we define the edges of the hypergraph. For any two tests $\pi_{x_1 y}$ and $\pi_{x_2 y}$ in Π that share the same y variable we define the following edges between the block $B[x_1]$ and the block $B[x_2]$:

$$\left\{ \{(x_1, G), (x_2, F_1), (x_2, F_2), \ldots, (x_2, F_{k-1})\} \;\middle|\; \pi_{x_1 y}(G) \cap \pi_{x_2 y}(\cap_{i=1}^{k-1} F_i) = \emptyset \right\}.$$

We say that these edges correspond to the pair of tests $\pi_{x_1 y}$ and $\pi_{x_2 y}$. Notice that every edge contains exactly k vertices, one vertex from the block $B[x_1]$ and $k-1$ vertices from the block $B[x_2]$. Also note that we can have edges between $B[x_1]$ and $B[x_2]$ that correspond to more than one pair of tests. This can be as a result of parallel edges in (X, Y, E) or as a result of several y variables to which both x_1 and x_2 are connected.

4.2 Completeness

Assume that the Strong-LC instance Φ has an assignment A in which at least a $1-\zeta$ fraction of the X-variables have all their tests satisfied. Let X_0 be the set of all such variables with $|X_0| \geq (1-\zeta)|X|$. We claim that the following is an independent set:

$$\mathcal{IS} = \{(x, F) \mid x \in X_0,\ A(x) \in F\}.$$

Consider any edge $\{(x_1, G), (x_2, F_1), \ldots, (x_2, F_{k-1})\}$ and let $\pi_{x_1 y}$ and $\pi_{x_2 y}$ be the pair of tests it corresponds to. Assume towards contradiction that all its vertices are in $\mathcal{IS}$. Clearly, this implies that $x_1 \in X_0$ and $x_2 \in X_0$. Also, since all tests incident to both x_1 and x_2 are satisfied by A, we have

$$\pi_{x_1 y}(A(x_1)) = A(y) = \pi_{x_2 y}(A(x_2)).$$

Therefore, $A(y) \in \pi_{x_1 y}(G) \cap \pi_{x_2 y}(\cap_{i=1}^{k-1} F_i)$. In particular, this implies that $\pi_{x_1 y}(G) \cap \pi_{x_2 y}(\cap_{i=1}^{k-1} F_i) \neq \phi$ and we reach a contradiction by recalling the construction of the edges.

Note that with the bias parameter $p = 1-\frac{1}{k}-\varepsilon$, for every $x \in X_0$ the weight of the set $\mathcal{IS} \cap B[x]$ is exactly p times the total weight of vertices in $B[x]$. Hence

$$weight(\mathcal{IS}) = (1-\zeta) \cdot (1-\frac{1}{k}-\varepsilon) \geq 1-\frac{1}{k}-2\varepsilon$$

since ζ can be chosen to be arbitrarily small.

4.3 Soundness

Now assume that there is no assignment to the Strong-LC instance Φ that satisfies even a γ fraction of the tests. We will show that the hypergraph contains no independent set of size δ.

Assume towards contradiction that the hypergraph contains an independent set $\mathcal{I}$ of size δ. For every variable $x \in X$, let

$$\mathcal{F}[x] = \{F \mid F \subseteq R,\ (x, F) \in \mathcal{I}\}.$$

Let X^* be the set of variables x such that $\mu_p^R(\mathcal{F}[x]) \geq \delta/2$, i.e., a weight of at least $\delta/2$ of the total weight in the block $B[x]$ belongs to the independent set $\mathcal{I}$. By an averaging argument, we have $|X^*| \geq \delta|X|/2$.

We will associate a "small" set of labels $L[x] \subseteq R$ for every $x \in X^*$ such that this "labeling" satisfies a weaker notion of consistency. More precisely we prove that

Lemma 4.1 *Given $\mathcal{I}$ and X^* as above, there exists a constant $h = h(k, \varepsilon, \delta)$ and non-empty sets of labels $L[x] \subseteq R$ for every $x \in X^*$ such that*

- $\forall\ x \in X^*,\ |L[x]| \leq h$
- *For every two tests $\pi_{x_1 y}$, $\pi_{x_2 y} \in \Pi$ sharing the same y variable, we have*

$$\pi_{x_1 y}(L[x_1]) \cap \pi_{x_2 y}(L[x_2]) \neq \emptyset.$$

This is the main technical lemma in the analysis and we prove it in the next subsection. Let us see how this lemma is sufficient to arrive at a contradiction. The idea is to define one label for every variable in $X \cup Y$ such that this labeling satisfies more than a γ fraction of the tests.

We will try to satisfy only those tests which are incident to X^*. This is a $\delta/2$ fraction of all the tests since $|X^*| \geq \delta|X|/2$ and the bipartite graph (X, Y, E) is left regular. Let Y^* be the set of variables in Y which have a test with some variable in X^*. For every $y \in Y^*$, fix $x(y)$ to be one variable in X^* with which y has a test. Define

$$L[y] \stackrel{def}{=} \pi_{x(y)y}(L[x(y)]).$$

We claim that for every test $\pi_{xy} \in \Pi$ with $x \in X^*$ and $y \in Y^*$ we have,

$$\pi_{xy}(L[x]) \cap L[y] \neq \emptyset.$$

When $x = x(y)$ this is clearly true and otherwise, it follows from Lemma 4.1.

Now consider the following probabilistic way of defining one label for every variable in $X^* \cup Y^*$. For a variable $x \in X^*$ (resp. $y \in Y^*$), define its label $A(x)$ (resp. $A(y)$) to be a randomly picked element of $L[x]$ (resp. $L[y]$). For each test $\pi_{xy} \in \Pi$ with $x \in X^*$ and $y \in Y^*$ the sets $\pi_{x,y}(L[x])$ and $L[y]$ intersect and both sets have size at most h. Therefore, with probability $1/h^2$, we have $\pi_{xy}(A(x)) = A(y)$ and the test is satisfied. Therefore, the expected fraction of satisfied tests is at least $\delta/(2h^2)$ and hence there must exist one assignment that satisfies these many tests. Choosing the parameter $\gamma < \delta/(2h^2)$ gives a contradiction.

4.4 Proof of Lemma 4.1

The set $L[x]$ for $x \in X^*$ will be constructed from the family $\mathcal{F}[x]$. Roughly speaking, the set $L[x]$ will be the core of the family $\mathcal{F}[x]$ along with all the elements which have non-negligible influence on the family $\mathcal{F}[x]$.[1]

[1] In [5], this set is referred to as the "extended core".

Let $\eta > 0$ be a sufficiently small "accuracy" parameter which will be fixed later. Applying Theorem 2.3, we get

Lemma 4.2 *For every variable $x \in X^*$, there exists a real number $p[x] \in (1 - \frac{1}{k} - \varepsilon, 1 - \frac{1}{k} - \frac{\varepsilon}{2})$ and a core-family $\widehat{\mathcal{F}}[x] \subseteq P(R)$ with core $C[x]$ such that*

- *The average sensitivity $\mathrm{as}_{p[x]}(\mathcal{F}[x]) \leq \frac{2}{\varepsilon}$.*
- *The size of $C[x]$ is at most h_0 which is a constant depending only on $k, \varepsilon, \eta, \delta$.*
- *$\mu^R_{p[x]}(\ \mathcal{F}[x]\ \Delta\ \widehat{\mathcal{F}}[x]) < \eta$, and in particular $\mu^R_{p[x]}(\widehat{\mathcal{F}}[x]) \geq \delta/4$ provided $\eta < \delta/4$.*

Let $\eta' > 0$ be a threshold parameter which will be chosen later. For every $x \in X^*$, we identify a set of elements $\mathrm{Infl}[x] \subseteq R \setminus C[x]$ that have non-negligible influence on the family $\mathcal{F}[x]$, i.e.,

$$\mathrm{Infl}[x] = \{\sigma \in R \setminus C[x] \mid \ \mathrm{Influence}_{p[x]}(\mathcal{F}[x], \sigma) \geq \eta'\}.$$

Since $\mathcal{F}[x]$ has average sensitivity at most $\frac{2}{\varepsilon}$ and the average sensitivity is simply the sum of influences of all the elements, it follows that the size of $\mathrm{Infl}[x]$ is at most $\frac{2}{\eta'\varepsilon}$ which is a constant. Finally, we define the set $L[x]$ as

$$L[x] \stackrel{def}{=} C[x]\ \cup\ \mathrm{Infl}[x]. \tag{1}$$

Clearly, $L[x]$ has size at most $h \stackrel{def}{=} h_0 + \frac{2}{\eta'\varepsilon}$.

To finish the proof of Lemma 4.1, it remains to show that for every two tests $\pi_{x_1y}, \pi_{x_2y} \in \Pi$ sharing the same y variable, we have $\pi_{x_1y}(L[x_1]) \cap \pi_{x_2y}(L[x_2]) \neq \emptyset$. Note that π_{x_1y}, π_{x_2y} are bijections and w.l.o.g. we can assume them to be identity maps. Thus we need to show that $L[x_1] \cap L[x_2] \neq \emptyset$. It will be clear how the proof would work in the general case.

We will assume on the contrary that $L[x_1] \cap L[x_2] = \emptyset$ and exhibit an edge $\{(x_1, G), (x_2, F_i)_{i=1}^{k-1}\}$ all of whose vertices are in the supposed independent set $\mathcal{I}$, thus giving a contradiction. Let us begin with a lemma:

Lemma 4.3 *There exists $U_0 \subseteq C[x_1]$ such that defining $R' \stackrel{def}{=} R \setminus (C[x_1] \cup C[x_2])$ and $\mathcal{H}[x_1] \subseteq P(R')$ as*

$$\mathcal{H}[x_1] \stackrel{def}{=} \{H \mid H \in P(R'),\ U_0 \cup H \in \mathcal{F}[x_1]\ \}$$

we have $\mu^{R'}_{p[x_1]}(\mathcal{H}[x_1]) \geq 1 - 8\eta/\delta$.

Proof: The assumption $L[x_1] \cap L[x_2] = \emptyset$ along with Equation (1) gives

$$C[x_1] \cap C[x_2] = \emptyset, \quad C[x_2] \cap \mathrm{Infl}[x_1] = \emptyset.$$

This implies that every element of $C[x_2]$ has influence at most η' on the family $\mathcal{F}[x_1]$. Let $\mathcal{F}'[x_1] \subseteq \mathcal{F}[x_1]$ be a family defined as

$$\mathcal{F}'[x_1] \stackrel{def}{=} \{F \mid F \in \mathcal{F}[x_1], F \setminus C[x_2] \in \mathcal{F}[x_1]\}.$$

Applying Lemma 2.4, we get

$$\begin{aligned}\mu^R_{p[x_1]}(\mathcal{F}'[x_1]\ \Delta\ \mathcal{F}[x_1]) \quad &< \quad \eta'\ |C[x_2]|\ (\min(p[x_1], 1 - p[x_1]))^{-|C[x_2]|} \leq \\ &\quad\ \ \eta'\ h_0\ (\min(p[x_1], 1 - p[x_1]))^{-h_0} \leq \eta\end{aligned}$$

by choosing η' small enough. It follows that

$$\mu^R_{p[x_1]}(\widehat{\mathcal{F}}[x_1] \ \setminus \ \mathcal{F}'[x_1]) \leq \mu^R_{p[x_1]}(\mathcal{F}'[x_1] \ \Delta \ \widehat{\mathcal{F}}[x_1]) \ < 2\eta.$$

We would like to find a set $U_0 \subseteq C[x_1]$ in the core family $\widehat{\mathcal{F}}[x_1]$ such that the two families obtained by taking only the sets in $\widehat{\mathcal{F}}[x_1]$ and $\mathcal{F}'[x_1]$ whose intersection with $C[x_1]$ is U_0 are very close. So,

$$\begin{aligned} 2\eta \quad > \quad & \mu^R_{p[x_1]}(\widehat{\mathcal{F}}[x_1] \ \setminus \ \mathcal{F}'[x_1]) = \Pr_{D\in\mu^R_{p[x_1]}}[D \in \widehat{\mathcal{F}}[x_1] \ \setminus \ \mathcal{F}'[x_1]] = \\ & \sum_{U\subseteq C[x_1]} \Pr_{D\in\mu^R_{p[x_1]}}[D \cap C[x_1] = U \text{ and } D \in \widehat{\mathcal{F}}[x_1] \ \setminus \ \mathcal{F}'[x_1]] \stackrel{\{1\}}{=} \\ & \sum_{U\subseteq C[x_1],\ U\in\widehat{\mathcal{F}}[x_1]} \mu^{C[x_1]}_{p[x_1]}(U) \ \Pr_{D\in\mu^{R\setminus C[x_1]}_{p[x_1]}}[\ (U \cup D) \notin \mathcal{F}'[x_1] \ ] \end{aligned}$$

where $\{1\}$ holds since $\widehat{\mathcal{F}}[x_1]$ is a core family and hence depends only on $C[x_1]$. Since $\mu^{C[x_1]}_{p[x_1]}(\{U \subseteq C[x_1] \mid U \in \widehat{\mathcal{F}}[x_1]\}) \geq \delta/4$ this implies that there exists $U_0 \subseteq C[x_1]$, $U_0 \in \widehat{\mathcal{F}}[x_1]$ such that $\Pr_{D\in\mu^{R\setminus C[x_1]}_{p[x_1]}}[\ (U_0 \cup D) \notin \mathcal{F}'[x_1]\ ] < 2\eta/(\delta/4)$. In other words, if we define $\mathcal{G}$ as $\{D \subseteq R \setminus C[x_1] \mid D \cup U_0 \in \mathcal{F}'\}$, then $\mu^{R\setminus C[x_1]}_{p[x_1]}(\mathcal{G}) < 8\eta/\delta$ and $\mu^{R\setminus C[x_1]}_{p[x_1]}(P(R \setminus C[x_1]) \setminus \mathcal{G}) > 1 - 8\eta/\delta$.

Finally, notice that $\mathcal{G}$ (and $P(R \setminus C[x_1]) \setminus \mathcal{G}$) does not depend on $C[x_2]$. Hence, the family

$$\mathcal{H}[x_1] \stackrel{def}{=} \{H \mid H \subseteq R' = R \setminus (C[x_1] \cup C[x_2]), \quad H \notin \mathcal{G} \ \}$$

satisfies $\mu^{R'}_{p[x_1]}(\mathcal{H}[x_1]) = \mu^{R\setminus C[x_1]}_{p[x_1]}(P(R \setminus C[x_1]) \setminus \mathcal{G}) > 1 - 8\eta/\delta$, as required. ∎

Analogous to Lemma 4.3 we have by symmetry,

Lemma 4.4 *There exist $V_0 \subseteq C[x_2]$ such that defining $R' \stackrel{def}{=} R \setminus (C[x_1] \cup C[x_2])$ and $\mathcal{H}[x_2] \subseteq P(R')$ as*

$$\mathcal{H}[x_2] \stackrel{def}{=} \{H \mid H \in P(R'), \ V_0 \cup H \in \mathcal{F}[x_2] \ \}$$

we have $\mu^{R'}_{p[x_2]}(\mathcal{H}[x_2]) \geq 1 - 8\eta/\delta$.

Let $p^* \stackrel{def}{=} 1 - \frac{1}{k} - \frac{\varepsilon}{2}$. Note that $\mathcal{H}[x_1]$ and $\mathcal{H}[x_2]$ are both monotone subfamilies of $P(R')$. Therefore, according to Lemma 2.2, $\mu^{R'}_{p^*}(\mathcal{H}[x_1]) \geq \mu^{R'}_{p[x_1]}(\mathcal{H}[x_1]) \geq 1 - 8\eta/\delta$ and similarly for x_2. Hence, the intersection of the families $\mathcal{H}[x_1]$ and $\mathcal{H}[x_2]$ satisfies

$$\mu^{R'}_{p^*}(\mathcal{H}[x_1] \cap \mathcal{H}[x_2]) \geq 1 - 16\eta/\delta > 1 - \frac{1}{k}$$

by picking η small enough. Hence, Lemma 2.5 implies that there exist sets $H_1, H_2, \ldots, H_k \in \mathcal{H}[x_1] \cap \mathcal{H}[x_2]$ such that

$$\cap_{i=1}^k H_i = \emptyset.$$

In particular, $H_1, H_2, \ldots, H_{k-1} \in \mathcal{H}[x_2]$ and $H_k \in \mathcal{H}[x_1]$.

Now define $G = U_0 \cup H_k$ and $F_i = V_0 \cup H_i$ for $1 \leq i \leq k-1$. By definition of the families $\mathcal{H}[x_1], \mathcal{H}[x_2]$, we have, $G \in \mathcal{F}[x_1], F_i \in \mathcal{F}[x_2]$ for $1 \leq i \leq k-1$. Thus $\{(x_1, G), (x_2, F_i)_{i=1}^{k-1}\}$ are vertices in the supposed independent set and they form an edge since

$$G \cap \left(\cap_{i=1}^{k-1} F_i \right) = \cap_{i=1}^k H_i = \emptyset.$$

This completes the proof.

References

[1] S. Arora, B. Bollobás, and L. Lovász. Proving integrality gaps without knowing the linear program. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (FOCS) 2002*, Vancouver, Canada, November 2002.

[2] M. Bellare, O. Goldreich, and M. Sudan. Free bits, PCPs, and nonapproximability—towards tight results. *SIAM Journal on Computing*, 27(3):804–915, June 1998.

[3] I. Dinur, V. Guruswami, and S. Khot. Vertex cover on k-uniform hypergraphs is hard to approximate within factor $(k-3-\varepsilon)$. In *Electronic Colloquium on Computational Complexity, Technical Report TR02-027*, 2002.

[4] I. Dinur, V. Guruswami, S. Khot, and O. Regev. A new multilayered PCP and the hardness of hypergraph vertex cover. 2002. Submitted.

[5] I. Dinur and S. Safra. On the importance of being biased. In *Proc. 34th ACM Symp. on Theory of Computing (STOC)*, 2002.

[6] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy. Approximating clique is almost NP-complete. In *Proc. 32nd IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 2–12, 1991.

[7] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–35, 1998.

[8] O. Goldreich. Using the FGLSS-reduction to prove inapproximability results for minimum vertex cover in hypergraphs. ECCC Technical Report TR01-102, 2001.

[9] V. Guruswami, J. Håstad, and M. Sudan. Hardness of approximate hypergraph coloring. In *Proc. 41st IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 149–158. IEEE Computer Society Press, 2000.

[10] E. Halperin. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs. *SIAM Journal on Computing*, 31(5):1608–1623, 2002.

[11] J. Håstad. Some optimal inapproximability results. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1–10, May 1997.

[12] J. Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Math.*, 182(1):105–142, 1999.

[13] J. Holmerin. Improved inapproximability results for vertex cover on k-regular hyper-graphs. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, 2002.

[14] J. Holmerin. Vertex cover on 4-regular hyper-graphs is hard to approximate within 2 - ϵ. In *Proc. 34th ACM Symp. on Theory of Computing (STOC)*, 2002.

[15] S. Khot. On the power of unique 2-Prover 1-Round games. In *Proc. 34th ACM Symp. on Theory of Computing (STOC)*, pages 767–775, May 2002.

[16] L. Russo. An approximate zero-one law. *Z. Wahrsch. Verw. Gebiete*, 61(1):129–139, 1982.

[17] L. Trevisan. Non-approximability results for optimization problems on bounded degree instances. In *Proc. 33rd ACM Symp. on Theory of Computing (STOC)*, 2001.

[18] V. V. Vazirani. *Approximation Algorithms*. Springer, 2001.

A Proofs

Proof of Lemma 2.4: A similar proof appears in [5]. Consider the family

$$\mathcal{F}'' \stackrel{def}{=} \{F \in P(R \setminus T) \mid F \cup T \in \mathcal{F},\ F \notin \mathcal{F}\}.$$

It can be seen that

$$\mu_p^R(\mathcal{F}) - \mu_p^R(\mathcal{F}') \leq \mu_p^{R\setminus T}(\mathcal{F}'').$$

For any set $F \in \mathcal{F}''$ there must exist some $D \subseteq T$ and an element $\sigma \in T$ such that $F \cup D \cup \{\sigma\} \in \mathcal{F}$ but $F \cup D \notin \mathcal{F}$. Hence, any set $F \in \mathcal{F}''$ contributes at least $\mu_p^{R\setminus T}(F) \cdot \min(p, 1-p)^{|T|}$ to the influence of one $\sigma \in T$. It remains to notice that the total influence of elements in T is at most $|T| \cdot \eta$. ■