Vertex Cover on $k$-Uniform Hypergraphs is Hard to Approximate within Factor $(k - 3 - \varepsilon)$

Irit Dinur*  
Venkatesan Guruswami†  
Subhash Khot‡

April, 2002

Abstract

Given a $k$-uniform hypergraph, the $k$-Vertex-Cover problem is to find a minimum subset of vertices that "hits" every edge. We show that for every integer $k \geq 5$, $k$-Vertex-Cover is NP-hard to approximate within a factor of $(k - 3 - \varepsilon)$, for an arbitrarily small constant $\varepsilon > 0$.

This almost matches the upper bound of $k$ for this problem, which is attained by the straightforward greedy approximation algorithm. The best previously known hardness result was due to Holmerin [Hol02a], who showed the NP-hardness of approximating $k$-Vertex-Cover within a factor of $k!^{\frac{1}{k-2}}$.

We present two constructions: one with a simple purely combinatorial analysis, showing $k$-Vertex-Cover to be NP-hard to approximate to within a factor $\Omega(k)$, followed by a stronger construction that obtains the $(k - 3 - \varepsilon)$ inapproximability bound. The latter construction introduces a novel way of combining ideas from Dinur and Safra's paper [DS02] and the notion of covering complexity introduced by Guruswami, Håstad and Sudan [GHS00]. This also allows us to prove a hardness factor of $(k - 1 - \varepsilon)$ assuming the hardness of $O(\log n)$-coloring a $c$-colorable graph for some fixed $c \geq 3$.

1 Introduction

Given a $k$-uniform hypergraph, $G = (V, E)$ with vertices $V$ and hyperedges $E \subseteq \binom{V}{k} \overset{def}{=} \{e \subseteq V \mid |e| = k\}$, a vertex-cover in $G$ is a subset $S \subseteq V$ that intersects each edge. An independent set in $G$ is a subset whose complement is a vertex cover, or in other words is a subset of vertices that contains no edge. The $k$-Vertex-Cover problem consists of finding a minimum size vertex cover in a $k$-uniform hypergraph. This problem is alternatively called the minimum hitting set problem with sets of size $k$ (and is equivalent to the set cover problem where each element of the universe occurs in exactly $k$ sets).

We show that this problem is NP-hard to approximate within $(k - 3 - \varepsilon)$ for an arbitrarily small constant $\varepsilon > 0$. The result is almost tight as this problem is approximable to within $k$ by repeatedly selecting one arbitrary hyperedge, adding all its vertices into the cover and and removing all the "covered" hyperedges. The best known algorithm [Hal00] gives only a slight improvement on this greedy algorithm, achieving an approximation factor of $k - o(1)$.

*School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540. irit@ias.edu
†University of California at Berkeley, Miller Institute for Basic Research in Science, Berkeley, CA 94720. venkat@cs.mit.edu
‡Department of Computer Science, Princeton University, Princeton, NJ 08544. khot@cs.princeton.edu
Previous Hardness Results

This problem was suggested by Trevisan [Tre01] who initiated a study of bounded degree instances of certain combinatorial problems. There it was shown that this problem is hard to approximate within a factor of $k^{1/19}$. Holmerin [Hol02b] showed that E4-Vertex-Cover is NP-hard to approximate within $(2 - \varepsilon)$, and more recently [Hol02a] that Ek-Vertex-Cover is NP-hard to approximate within $k^{1-\varepsilon}$. Goldreich [Gol01] found a simple ‘FGLSS’-type [FGL+91] reduction (involving no use of the long-code, a crucial component in most recent PCP constructions) to obtain a hardness factor of $(2 - \varepsilon)$ for Ek-Vertex-Cover for some constant $k$.

Our Results

We present two constructions: one that attains a hardness factor of $\Omega(k)$ (already improving the best previously known result) with a simple, purely combinatorial analysis, and one that is stronger attaining the inapproximability factor of $(k - 3 - \varepsilon)$.

Our “simple” construction follows that of [Hol02b] who showed that it is NP-hard to approximate E4-Vertex-Cover to within a factor $(2 - \varepsilon)$. Taking a new “set-theoretic” viewpoint on that construction, we give a purely combinatorial proof of Holmerin’s theorem (in contrast to Holmerin’s use of Fourier analysis), relying solely on one new Erdős-Ko-Rado (EKR) type combinatorial lemma (Lemma 2.2) that bounds the maximal size of $t$-intersecting families of subsets (i.e. families in which every pair of subsets intersect on at least $t$ elements). Taking this new ‘set-theoretic’ viewpoint results in a direct extension of that construction to obtain an $\Omega(k)$ inapproximability factor.

Our “strong” construction, achieving a hardness factor of $k - 3 - \varepsilon$ involves a novel way of combining ideas from Dinur and Safra’s paper [DS02] and the notion of covering complexity introduced by Gururwami, Hästad and Sudan [GHS00].

A constraint satisfaction problem is said to have covering complexity $L$ (see [GHS00]) if it is NP-hard to distinguish between the case when the CSP has a satisfying assignment and the case when, given any $L$ assignments, there exists a constraint that is not satisfied by any of these $L$ assignments. Gururwami et al [GHS00] show hardness of coloring 2-colorable 4-uniform hypergraphs with constantly many colors. This result is equivalent to saying that the Not-All-Equal predicate on 4 binary variables has arbitrarily high covering complexity. In our paper, we need to use a (stronger) variant of the notion of covering complexity (see Definition 4.2).

We combine the covering complexity techniques with the methods from [DS02]. We borrow a powerful tool used in their paper, namely the Friedgut’s Theorem about influence of variables on Boolean functions (Theorem 4.1). Assuming that for some fixed $c \geq 3$, it is hard to color an $n$-vertex $c$-colorable graph using at most $b \log n$ for every integer $b$, we get a predicate on 2 variables with the necessary covering complexity and this lets us prove a factor $(k - 1 - \varepsilon)$ hardness for the Ek-Vertex-Cover problem.

Location of the gap: All our hardness results have the gap between sizes of the vertex cover at the “right” location. Specifically, to prove a factor $(B - \varepsilon)$ hardness we show that it is hard to distinguish between $k$-uniform hypergraphs that have a vertex cover of weight $\frac{1}{B} + \varepsilon$ from those whose minimum vertex cover has weight at least $(1 - \varepsilon)$. This result is stronger than a gap of about $B$ achieved, for example, between vertex covers of weight $1/B^2$ and $1/B$. Put another way, our result shows that for $k$-uniform hypergraphs, for $k \geq 3$, there is a fixed $\alpha$ such that for arbitrarily small $\varepsilon > 0$, it is NP-hard to find an independent set consisting of a fraction $\varepsilon$ of the vertices even if the hypergraph is promised to contain an independent set comprising a fraction $\alpha$ of vertices. We remark that such a result is not known for graphs and seems out of reach of current techniques. (The recent factor 4/3
hardness result for vertex cover on graphs due to Dinur and Safra [DS02], for example, shows that it is NP-hard to distinguish between cases when the graph has an independent set of size \( n/3 \) and when no independent set has more than \( n/9 \) vertices.)

**Organization of the Paper**

We begin with some preliminaries, including the starting-point PCP theorem, and some combinatorial lemmas that are used later in the analysis of the constructions. In Section 3 we present our first construction and prove a hardness of approximation factor of \( \Omega(k) \). In Section 4 we present our stronger construction proving a \( (k - 3 - \varepsilon) \) hardness factor. In Section 5 we explain how a hardness assumption for graph coloring implies a factor \( (k - 1 - \varepsilon) \) hardness.

**Subsequent Work**

Following this work, Dinur, Guruswami, Khot and Regev [DGKR02] were able to unconditionally show a hardness of approximation factor of \( (k - 1 - \varepsilon) \). The techniques in this paper are quite different from the ones in [DGKR02] and could be of independent interest. Moreover, the \( \Omega(k) \) hardness factor is probably still the best in terms of holding for super-constant values of \( k \).

**2 Preliminaries**

For a universe \( R \), let \( P(R) \) denote its power set, i.e. the family of all subsets of \( R \). A family \( \mathcal{F} \subseteq P(R) \) is called *monotone* if \( F \in \mathcal{F}, F' \subseteq F \) implies \( F' \in \mathcal{F} \).

For a "bias parameter" \( 0 < p < 1 \), the weight \( \mu_p(F) \) of a set \( F \) is defined as

\[
\mu_p(F) \overset{\text{def}}{=} p^{|F|}(1-p)^{|R \setminus F|}.
\]

The weight of a family \( \mathcal{F} \subseteq P(R) \) is defined as

\[
\mu_p(\mathcal{F}) \overset{\text{def}}{=} \sum_{F \in \mathcal{F}} \mu_p(F).
\]

Note that the bias parameter defines a product distribution on \( P(R) \), where the probability of one subset \( F \in P(R) \) is determined by independently flipping a \( p \)-biased coin to determine the membership of each element of \( R \) in the subset. We denote this distribution by \( \mu_p \).

**2.1 Intersecting Families of Subsets**

In this section we present some combinatorial lemmas regarding intersecting families of subsets, that will be useful later.

**Definition 2.1** For a family of subsets \( \mathcal{F} \subseteq P(R) \), let

\[
\mathcal{F} \cap \mathcal{F} \overset{\text{def}}{=} \{ F_1 \cap F_2 \mid F_1, F_2 \in \mathcal{F} \}.
\]

**Lemma 2.2 (EKR-Core)** For every \( \varepsilon, \delta > 0 \), there exists some \( t = t(\varepsilon, \delta) > 0 \) such that for every finite \( R \) and set family \( \mathcal{F} \subseteq P(R) \), if \( \mu_{\frac{1}{t}+\delta}(\mathcal{F}) > \varepsilon \), then there exists some 'core' subset \( C \in \mathcal{F} \cap \mathcal{F} \) with \( |C| \leq t \).
**Proof Sketch:** A family of subsets is $t$-intersecting if for every $F_1, F_2 \in \mathcal{F}$, $|F_1 \cap F_2| \geq t$. The idea is that a family cannot be $t$-intersecting, for large (but constant) $t$ and still retain a non-negligible size. Thus, if $\mu_{1-\delta}(\mathcal{F}) > \varepsilon$, there exists some $t$ for which $\mathcal{F}$ is not $t+1$ intersecting, hence $\mathcal{F}$ has a 'core' subset of size $t$.

For the full proof (see Section A.1), we rely on the complete intersection theorem for finite sets of [AK97] that fully characterizes the maximal $t$-intersecting set families.

**Proposition 2.3** Let $p > 0$ and let $\mathcal{F} \subseteq P(R)$. Then $\mu_{p^2}(\mathcal{F} \cap \mathcal{F}) \geq (\mu_p(\mathcal{F}))^2$.

**Proof:**

$$\Pr_{F \in \mu_{p^2}^2} [F \in \mathcal{F} \cap \mathcal{F}] = \Pr_{F_1, F_2 \in \mu_p} [F_1 \cap F_2 \in \mathcal{F}] \geq \Pr_{F_1 \in \mu_p} [F_1 \in \mathcal{F}] \cdot \Pr_{F_2 \in \mu_p} [F_2 \in \mathcal{F}] = (\mu_p(\mathcal{F}))^2$$

Note that when $\mathcal{F}$ is monotone and defined by exactly one 'minterm', equality holds.

**Proposition 2.4** Let $p > 0$, $\mathcal{F} \subseteq P(R)$. Let $\mathcal{F}^k \overset{def}{=} \{F_1 \cap \cdots \cap F_k | F_i \in \mathcal{F}\}$. Then,

$$\mu_{p^k}(\mathcal{F}^k) \geq (\mu_p(\mathcal{F}))^k.$$**Proof:** By induction on $k$.}

### 2.2 Starting Point - PCP

**The Parallel Repetition Theorem**

As is the case for many inapproximability results, we begin our reduction from Raz’s parallel repetition theorem [Raz98] which is a version of the PCP theorem that is very powerful and convenient to work with. Let $\Phi = \{\varphi_1, ..., \varphi_n\}$ be a system of local-constraints over two sets of variables, denoted $X$ and $Y$. Let $R_X$ denote the range of the $X$-variables and $R_Y$ the range of the $Y$-variables $^1$. Assume each constraint $\varphi \in \Phi$ depends on exactly one $x \in X$ and one $y \in Y$, furthermore, for every value $a_x \in R_X$ assigned to $x$ there is exactly one value $a_y \in R_Y$ to $y$ such that the constraint $\varphi$ is satisfied. Therefore, we can write each local constraint $\varphi \in \Phi$ as a function from $R_X$ to $R_Y$, and use notation $\varphi_{x \rightarrow y} : R_X \rightarrow R_Y$ (this notation is borrowed from [DS02]). Furthermore, we assume that every $X$-variable appears in the same number of local-constraints in $\Phi$.

**Theorem 2.5 (PCP Theorem [AS98, ALM+98, Raz98])** Let $\Phi = \{\varphi_1, ..., \varphi_n\}$ be as above. There exists a universal constant $\gamma > 0$ such that for every constant $|R_X|$, it is $\text{NP}$-hard to distinguish between the following two cases:

- **YES**: There is an assignment $A : X \cup Y \rightarrow R_X \cup R_Y$ such that all $\varphi_1, ..., \varphi_n$ are satisfied by $A$, i.e. $\forall \varphi_{x \rightarrow y} \in \Phi$, $\varphi_{x \rightarrow y}(A(x)) = A(y)$.

- **NO**: No assignment can satisfy more than a fraction $\frac{1}{|R_X|^\gamma}$ of the constraints in $\Phi$.

$^1$Readers familiar with the Raz-verifier may prefer to think concretely of $R_X = [7^u]$ and $R_Y = [2^u]$ for some number $u$ of repetitions.
3 The “Simple” Construction

In this section, we prove the factor \( \Omega(k) \) hardness result for \( E_k \)-Vertex-Cover. Our construction follows that of [Hol02b] who showed that it is NP-hard to approximate \( E_4 \)-Vertex-Cover within factor \((2 - \varepsilon)\). Taking a new viewpoint on that construction we give a purely combinatorial proof of Holmerin’s theorem (in contrast to Holmerin’s use of Fourier analysis), relying solely on one new Erdős-Ko-Rado (EKR) type combinatorial lemma (Lemma 2.2) that bounds the maximal size of \( t \)-intersecting families of subsets (i.e., families in which every pair of subsets intersect on at least \( t \) elements). We then show a direct extension of that construction to obtain an \( \Omega(k) \) inapproximability result for \( E_k \)-Vertex-Cover.

The use of EKR-type bounds in the context of inapproximability results was initiated in [DS02] as part of a more complicated construction and analysis for proving a hardness result for approximating vertex-cover on graphs. The structure of the problem at hand allows a very modular use of EKR-type bounds, and perhaps provides a better intuition as to why they are useful. Since EKR-type bounds are known in many cases to be tight, we believe that similar such bounds may prove fruitful for obtaining improved inapproximability results for other approximation problems.

As a warmup, let us first prove

**Theorem 3.1** For any \( \delta > 0 \), it is NP-hard to approximate \( E_4 \)-Vertex-Cover to within \( 2 - \delta \).

This result is already known (see [Hol02b]), albeit using more complex analysis techniques.

**Proof:** Start with a PCP instance, as given in theorem 2.5, namely a set of local constraints \( \Phi = \{ \varphi_1, ..., \varphi_n \} \) over variables \( X \cup Y \), whose respective ranges are \( R_X, R_Y \). For parameters, fix \( t = \ell(\frac{1}{2k}, \delta) \), and take \( |R_X| > \left( \frac{2t}{\delta} \right)^{1/7} \) where \( \gamma > 0 \) is the universal constant from Theorem 2.5. From \( \Phi \), we now construct a 4-uniform hypergraph whose minimum vertex cover has weight \( \approx \frac{1}{2} \) or \( \approx 1 \) depending on whether \( \Phi \) is satisfiable or not.

We present a construction of a weighted hypergraph \( G = (V, E, \Lambda) \), which can then be translated into an unweighted hypergraph via a standard duplication of vertices. The vertex set of \( G \) is

\[
V \overset{def}{=} X \times P(R_X)
\]

namely for each \( x \in X \) we construct a block of vertices denoted \( V[x] = \{ x \} \times P(R_X) \) corresponding to all possible subsets of \( R_X \). The weight of each vertex \( \langle x, F \rangle \in V \) is

\[
\Lambda(\langle x, F \rangle) \overset{def}{=} \frac{1}{|X|} \cdot \mu_{\frac{1}{2} - \delta}(F)
\]

The hyperedges are defined as follows. For every pair of local-constraints \( \varphi_{x_1 \rightarrow y}, \varphi_{x_2 \rightarrow y} \in \Phi \) sharing a mutual variable \( y \in Y \), we add the hyperedge \( \{ \langle x_1, F_1 \rangle, \langle x_1, F'_1 \rangle, \langle x_2, F_2 \rangle, \langle x_2, F'_2 \rangle \} \) if and only if there is no \( r_1 \in F_1 \cap F'_1 \) and \( r_2 \in F_2 \cap F'_2 \) such that \( \varphi_{x_1 \rightarrow y}(r_1) = \varphi_{x_2 \rightarrow y}(r_2) \):

\[
E \overset{def}{=} \bigcup_{\varphi_{x_1 \rightarrow y}, \varphi_{x_2 \rightarrow y} \in \Phi} \{ \langle x_1, F_1 \rangle, \langle x_1, F'_1 \rangle, \langle x_2, F_2 \rangle, \langle x_2, F'_2 \rangle \} \mid \varphi_{x_1 \rightarrow y}(F_1 \cap F'_1) \cap \varphi_{x_2 \rightarrow y}(F_2 \cap F'_2) = \phi \}
\]

where the union is taken over all pairs of local-constraints with a mutual variable \( y \).

**Lemma 3.2 (Completeness)** If \( \Phi \) is satisfiable, then \( G \) has a vertex cover whose weight is \( \leq \frac{1}{2} + \delta \).
Proof: Assume a satisfying assignment \( A : X \cup Y \to R_X \cup R_Y \) for \( \Phi \). The following set is a vertex cover of \( G \): 
\[
\{ \langle x, F \rangle \in V \mid x \in X, A(x) \not\in F \} \tag{1}
\]
For every hyperedge \( e = \{ \langle x_1, F_1 \rangle, \langle x_1, F'_1 \rangle, \langle x_2, F_2 \rangle, \langle x_2, F'_2 \rangle \} \) either \( A(x_1) \not\in F_1 \cap F'_1 \) or \( A(x_2) \not\in F_2 \cap F'_2 \), otherwise since \( A(x_1), A(x_2) \) agree on every mutual \( Y \)-variable, we have \( \varphi_{x_1 \to y}(F_1 \cap F'_1) \cap \varphi_{x_2 \to y}(F_2 \cap F'_2) \neq \phi \), and \( e \) would not have been a hyperedge.

Now note that the weight of the family \( \{ F \mid A(x) \not\in F \} \) w.r.t. the bias parameter \( (\frac{1}{2} - \delta) \) is \( (\frac{1}{2} + \delta) \).

Hence the weight of the vertex cover in (1) is \( \frac{1}{2} + \delta \).

Lemma 3.3 (Soundness) If \( G \) has a vertex cover whose weight is \( \leq 1 - \delta \), then \( \Phi \) is satisfiable.

Proof: Let \( S \subseteq V \) be such a vertex cover. By an averaging argument, there must be a set \( X' \subseteq X \), \( |X'| \geq \frac{\delta}{2} |X| \) such that for \( x \in X' \), \( \Pr_{v \in A \setminus V} [v \in S \mid v \in V[x]] \leq (1 - \frac{\delta}{2}) \). For each of these blocks, define 
\[
\mathcal{F}_x = \{ F \in P(R_X) \mid \langle x, F \rangle \not\in S \}
\]
It follows immediately that \( \forall x \in X' \), \( \mu_{\frac{1}{2} - \delta}(\mathcal{F}_x) \geq \frac{\delta}{2} \). The key observation is that due to Lemma 2.2 there exists some “core” subset \( C \) \( \subseteq \mathcal{F}_x \) \( \cap \mathcal{F}_x \) whose size is \( |C| \leq t = t(\frac{\delta}{2}, \delta) \). In other words, there are two subsets which we denote \( F_1, F_2 \) \( \subseteq \mathcal{F}_x \), such that \( |F_1 \cap F_2| \leq t \).

We next translate these “cores” into an assignment satisfying more than \( \frac{1}{|R_X|} \) fraction of \( \Phi \). Let \( x_1, x_2 \in X' \), and denote their cores respectively by \( C_{x_1} = F_{x_1} \cap F_{x_2} \), \( C_{x_2} = F_{x_2} \cap F_{x_1} \). The next observation is that for every \( \varphi_{x_1 \to y}, \varphi_{x_2 \to y} \in \Phi \) with \( x_1, x_2 \in X' \), there always exists some \( r_1 \in C_{x_1} \) and \( r_2 \in C_{x_2} \) such that \( \varphi_{x_1 \to y}(r_1) = \varphi_{x_2 \to y}(r_2) \), or in other words, we have 
\[
\varphi_{x_1 \to y}(C_{x_1}) \cap \varphi_{x_2 \to y}(C_{x_2}) \neq \phi \ . \tag{2}
\]
Indeed, if not, then the set 
\[
\{ \langle x_1, F_{x_1} \rangle, \langle x_1, F_{x_2} \rangle, \langle x_2, F_{x_1} \rangle, \langle x_2, F_{x_2} \rangle \}
\]
would be a hyperedge not hit by \( S \), contradicting the assumption that \( S \) is a vertex cover.

Let \( Y' \subseteq Y \) denote the set of all \( Y \)-variables that participate in some local-constraint with some \( x \in X' \), \( Y' \triangleq \{ y \mid \varphi_{x \to y} \in \Phi, x \in X' \} \). Associate each such \( y \in Y' \), with one arbitrary \( x \in X' \) with \( \varphi_{x \to y} \in \Phi \), and let \( C_y \triangleq \varphi_{x \to y}(C_x) \subseteq R_Y \). Now define a random assignment \( A \) by independently selecting for each \( x \in X', y \in Y' \) a random value from \( C_x, C_y \) respectively. Assign the rest of the variables \( (X \setminus X') \cup (Y \setminus Y') \) with any arbitrary value. To complete the proof, we prove:

Proposition 3.4

\[
E_A[\# \{ \varphi_{x \to y} \text{ is satisfied by } A \}] \geq \frac{\delta}{2t^2} \cdot |\Phi|
\]

Proof: We will show that for any \( x \in X' \), any \( \varphi_{x \to y} \in \Phi \) is satisfied by \( A \) with probability \( \geq \frac{1}{t^2} \), thus the expected number of local-constraints satisfied by \( A \) is \( \frac{|X'|}{|X|} \cdot \frac{1}{t^2} \geq \frac{\delta}{2t^2} \cdot |\Phi| \) (because every \( x \in X \) appears in the same number of local constraints). Assume \( C_y = \varphi_{x' \to y}(C_x') \) for some \( x' \in X' \) (note that \( |C_y| \leq t \)). By Equation (2), we have 
\[
C_y \cap \varphi_{x \to y}(C_x) = \varphi_{x \to y}(C_x) \cap \varphi_{x' \to y}(C_x') \neq \phi \ .
\]
Therefore, there is at least one value \( a_x \in C_x \) such that \( \varphi_{x \to y}(a_x) \in C_y \). Since for every \( x \in X' \), \( |C_x| \leq t \), there is at least a \( \frac{1}{t^2} \) probability of having \( \varphi_{x \to y}(A(x)) = A(y) \).
Thus, there exists some assignment \( A \) that meets the expectation, which means it satisfies \( \frac{\delta}{|T|} > \frac{1}{|X|} \) of the local constraints in \( \Phi \), hence \( \Phi \) is satisfiable.

Thus we have proven that distinguishing between the case where the minimum vertex cover of \( G \) has weight at most \((\frac{1}{4} + \delta)\) and the case where it has weight at least \((1 - \delta)\), enables deciding whether \( \Phi \) is satisfiable or not, and is therefore NP-hard.

3.1 \( k \)-Vertex-Cover is NP hard to Approximate Within \( k / 3 = \Omega(k) \).

We extend the construction above to work for any constant value of \( k \geq 4 \). We assume without loss of generality that \( k \) is divisible by 4, and for other values of \( k \) we can use the construction for the nearest \( k' = 4m \) and add \( k - k' \) distinct vertices to each edge.

The vertex set for our hypergraph is the same as in the case for \( k = 4 \), but the weights are different. We set

\[
\rho = \left( \frac{1}{2} - \delta \right)^{4/k},
\]

and

\[
\forall \rho = \langle x, F \rangle \in X \times P(R_X), \quad \Lambda(\rho) \overset{\text{def}}{=} \frac{1}{|X|} \cdot \mu_p(F)
\]

The hyperedges are as follows. For every \( \varphi_{x \rightarrow y}, \varphi_{x' \rightarrow y} \in \Phi \), and every \( \langle x, F_1 \rangle, \ldots, \langle x, F_{k'} \rangle \in V[x] \) and \( \langle x', F'_1 \rangle, \ldots, \langle x', F'_{k'} \rangle \in V[x'] \), we add the hyperedge \( \{ \langle x, F_1 \rangle, \ldots, \langle x, F_{k'} \rangle, \langle x', F'_1 \rangle, \ldots, \langle x', F'_{k'} \rangle \} \) to \( E \) if there is no \( r \in x \cap \cdots \cap x' \) and \( r' \in x' \cap \cdots \cap x' \) with \( \varphi_{x \rightarrow y}(r) = \varphi_{x' \rightarrow y}(r') \). That is,

\[
E \overset{\text{def}}{=} \bigcup_{\varphi_{x \rightarrow y}, \varphi_{x' \rightarrow y} \in \Phi} \{ \{ \langle x, F_1 \rangle, \ldots, \langle x, F_{k'} \rangle, \langle x', F'_1 \rangle, \ldots, \langle x', F'_{k'} \rangle \} \mid \varphi_{x \rightarrow y}(F_i) \cap \varphi_{x' \rightarrow y}(F'_i) = \emptyset \}
\]

As in the case of \( k = 4 \),

**Lemma 3.5 (Completeness)** If \( \Phi \) is satisfiable, then \( G \) has a hitting set whose weight is at most \( 1 - \rho \).

**Proof:** Again we take \( A \) to be a satisfying assignment, and set \( S = \bigcup_{x \in X} \{ \langle x, F \rangle \mid F \not\in A(x) \} \). The weight of \( S \) is \( 1 - \rho \).

The proof of soundness is also quite similar to that of Lemma 3.3, with one minor twist.

**Lemma 3.6 (Soundness)** If \( G \) has a hitting set whose weight is \( \leq 1 - \delta \), then \( \Phi \) is satisfiable.

**Proof:** Let \( S \subset V \) be such a hitting set. Again we consider the set \( X' \subseteq X \) for which \( \forall x \in X' \), \( \Pr_{v \in X} [v \in S \mid v \in X] \leq (1 - \frac{\delta}{2}). \) Again \( \frac{|X'|}{|X|} \geq \frac{\delta}{2} \). For each \( x \in X' \) let

\[
F_x = \{ F \in P(R_X) \mid \langle x, F \rangle \not\in S \}
\]

We cannot immediately apply Lemma 2.2 to find “weak-cores” for each \( F_x \) because lemma 2.2 works only when \( p < \frac{1}{4} - \delta \). Thus, we must first consider the family \( \mathcal{F}_x = \{ F_1 \cap \cdots \cap F_{k'/4} \mid F_i \in F_x \} \) whose size is guaranteed to be, by Proposition 2.4,

\[
\mu_{\frac{1}{4} - \delta}(F_x) = \mu_{\frac{1}{4} - \delta}(F_x^{k'/4}) \geq (\mu_p(F_x))^{k/4} \geq (\delta/2)^{k/4}.
\]
Now we can deduce the existence of a “core” subset \( C_x \subseteq R_x \), \(|C_x| \leq t = t((\delta/2)^{k/4}, \delta)\), such that there are some \( F^1_x, \ldots, F^k_x \in \mathcal{F}_x \) so that \( C_x = F^1_x \cap \cdots \cap F^k_x \). Again, observe that for every \( \varphi_{x \rightarrow y}, \varphi'_{x \rightarrow y} \in \Phi \), \( \varphi_{x \rightarrow y}(C_x) \cap \varphi'_{x \rightarrow y}(C_x') \neq \emptyset \). From here we can proceed as in the proof of Lemma 3.3 above to define a random assignment that satisfies an expected fraction \( \delta\) of the constraints in \( \Phi \).

All in all, we have proven

**Theorem 3.7** It is NP-hard to approximate \( \text{Ek-Vertex-Cover} \) to within a factor \( k/3 \).

**Proof:** For a \( k \)-uniform hypergraph, we have proven in Lemmas 3.5.3.6 that it is NP-hard to distinguish between a vertex-cover of size \( 1 - \delta \) and \( 1 - p = 1 - (\frac{1}{2} - \delta)^{\frac{k}{k}} \). The last inequality follows from

\[
(1 - \frac{3}{k})^{\frac{k}{k}} < \frac{1}{e^{\frac{k}{k}}} < \frac{1}{2} - \delta
\]

which implies that there is no factor \( \frac{k}{3} \) approximation algorithm for vertex-cover on \( k \)-uniform hypergraphs, unless \( \text{P=NP} \).

### 4 The “Stronger” Construction

In this section we present our stronger construction, achieving a hardness of approximation factor of \((k - 3 - \varepsilon)\) for an arbitrarily small \( \varepsilon > 0 \). This construction is quite different from the one in Section 3. Let us first introduce a couple of tools we use. First, Friedgut’s Theorem about influence of variables on Boolean functions (equivalently, influence of elements on set-families) and, second, a result that follows from Holmerin’s result [Hol02b] on the hardness of vertex cover in 4-uniform hypergraphs.

#### 4.1 Friedgut’s “Core” Theorem

Recall that a family \( \mathcal{F} \subseteq \mathcal{P}(R) \) is called monotone if \( F \in \mathcal{F} \) and \( F \subseteq F' \) implies \( F' \in \mathcal{F} \). 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

\[
\text{Influence}_p(\mathcal{F}, \sigma) \overset{\text{def}}{=} \mathbb{P}_{F \in \mu_p}[\text{exactly one of } F \cup \{\sigma\}, F \setminus \{\sigma\} \text{ is in } \mathcal{F}]
\]

The average sensitivity of a family is defined as the sum of the influences of all elements.

\[
\text{as}_p(\mathcal{F}) \overset{\text{def}}{=} \sum_{\sigma \in R} \text{Influence}_p(\mathcal{F}, \sigma)
\]

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, in some sense, determined by a ‘core’, see [DS02] for details.

**Theorem 4.1** ([Fri98, Rus82]) Let \( p \) be a bias parameter, \( \varepsilon, \delta > 0 \) be constants and \( \zeta \) be an “accuracy parameter”. Let \( \mathcal{F} \subseteq \mathcal{P}(R) \) be a monotone family such that \( \mu_p(\mathcal{F}) \geq \delta \). Then there exists \( p' \in (p, p + \varepsilon) \) and a set \( C \subseteq R \) called the “core” with the following properties:

- The average sensitivity of the family \( \mathcal{F} \) w.r.t the bias \( p' \) is at most \( \frac{1}{\zeta} \), i.e. \( \text{as}_{p'}(\mathcal{F}) \leq \frac{1}{\zeta} \).
- The size of \( C \) is a constant that depends only on \( p, \delta, \varepsilon, \zeta \).
• If a family $\mathcal{H} \subseteq P(R \setminus C)$ is defined as

$$\mathcal{H} \overset{\text{def}}{=} \{ H \mid H \subseteq R \setminus C, C \cup H \in \mathcal{F} \},$$

i.e. $\mathcal{H}$ consists of all possible extensions of the core $C$ into set that belongs to $\mathcal{F}$, then $\mu_p(\mathcal{H}) \geq 1 - \zeta$ where $\mu_p(\mathcal{H})$ is the weight of the family $\mathcal{H}$ under the $\mu_p$-distribution over the universe $R \setminus C$.

### 4.2 Hardness of 4CSP

We define a constraint satisfaction problem on 4 variables (4CSP) which captures a notion related to the notion of covering complexity introduced by [GHS00]. Our hypergraph construction will be based on the hardness of this 4CSP.

**Definition 4.2** A 4CSP $\mathcal{L} = (X, \Phi)$ over a domain $D$ is defined as follows: $X$ is a set of variables which take values from domain $D$. Every $\phi \in \Phi$ is a constraint on 4 variables (which is satisfied provided the values of the 4 variables belong to a specific subset of $D^4$). We define YES and NO instances of the CSP as follows.

- **YES**: There exists an assignment $f : X \mapsto D$ to the variables which satisfies every constraint $\phi \in \Phi$ (more formally, the values assigned by $f$ to the 4 variables in $\phi$ satisfy $\phi$).
- **NO**: For any subset of variables $Y \subseteq X$, $|Y| \geq \gamma |X|$ and for any $L$ assignments $f_1, f_2, \ldots, f_L : Y \mapsto D$, there exists a constraint $\phi \in \Phi$ such that all the 4 variables of the constraint $\phi$ are contained in $Y$ and every assignment $f_i, 1 \leq i \leq L$ fails to satisfy the constraint $\phi$.

For convenience, we say that $\phi$ is inside $Y$ if all the 4 variables of the constraint $\phi$ are in the set $Y$.

**Theorem 4.3** For every integer $L$ and every constant $\gamma > 0$, it is NP-hard to distinguish whether an instance $\mathcal{L}$ of a 4CSP over Boolean domain is a YES instance or a NO instance.

**Proof:** This follows immediately from a result of Holmerin [Hol02b]. He shows that for any constant $\gamma' > 0$, it is NP-hard to distinguish whether an $n$-vertex 4-uniform hypergraph is 2-colorable or it contains no independent set of size $\gamma'n$. Now let the vertices of the 4-uniform hypergraph be variables of a 4CSP. For every edge in the hypergraph, add a Not-All-Equal constraint on its 4 vertices. When the hypergraph is 2-colorable, it means that the 4CSP has a satisfying assignment. On the other hand, if there are $L$ assignments that satisfy every constraint inside a set of variables of size $\gamma n$, it means that this set of $\gamma n$ vertices can be colored properly with $2^L$ colors and hence there exists an independent set of size $\gamma' n = (\gamma/2^L)n$.

**Remark:** The notion of hardness between the YES and NO instances here is closely related to the notion of covering complexity introduced by [GHS00]. The notion of covering complexity requires that in the NO case, no $L$ assignments can together satisfy all the constraints. We require an even stronger condition that no $L$ assignments can satisfy every constraint inside a set of variables $Y$ whose size is $\gamma |X|$. 

9
4.3 The Construction of the Hypergraph

Let $B$ be the set of all $l$-tuples of variables of an instance $\mathcal{L}$ of a 4CSP given by Theorem 4.3. That is

$$B \overset{\text{def}}{=} \{(x_1, x_2, \ldots, x_l) \mid x_i \in X\}$$

An $l$-tuple $B \in B$ will be called a “block”. Let $R$ be the set of all possible “block assignments”, i.e. $R \overset{\text{def}}{=} D^l$ is the set of all strings of length $l$ over the domain $D$. Let $P(R)$ denote the family of all subsets of $R$, i.e.

$$P(R) \overset{\text{def}}{=} \{F \mid F \subseteq R\}$$

The vertex set $V$ of the hypergraph is defined to be

$$V \overset{\text{def}}{=} B \times P(R) = \{(B, F) \mid B \in B, F \in P(R)\}$$

The vertices will have weights. Let $p = 1 - \frac{1}{n^l} - \varepsilon$ be the “bias parameter”. The weight of a vertex $(B, F)$ is $\mu_p(F)$ where

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

To motivate the way we define the edges of the hypergraph, assume that $f : X \rightarrow D$ is an assignment that satisfies every constraint. Let $f[B]$ denote the restriction of this assignment to block $B$. Thus $f[B] \in R$. The edges of the hypergraph will be defined in such a way that the set of vertices $I_f$ is an independent set.

**Definition 4.4** We say that 4 blocks $(B_1, B_2, B_3, B_4)$ are “overlapping” if they agree on some $l - 1$ coordinates and the 4 variables on the remaining coordinate form a constraint in the 4CSP. More precisely, there exist variables $x_1, x_2, \ldots, x_{l-1}$ and $y_1, y_2, y_3, y_4$ and an index $t, 1 \leq t \leq l$ such that

1. $B_t = (x_1, x_2, \ldots, x_{l-1}, y_t, x_t, x_{t+1}, \ldots, x_{l-1})$ for $i = 1, 2, 3, 4$

2. There is a constraint $\phi \in \Phi$ on the variables $(y_1, y_2, y_3, y_4)$.

Note that the tuple $(\{x_j\}_{j=1}^{l-1}, \{y_j\}_{j=t}^{4}, t, \phi)$ completely characterizes the overlapping blocks.

For a block $B = (z_1, z_2, \ldots, z_l)$ and a block assignment $\sigma \in R$, let $\sigma(z_j)$ denote the value assigned by $\sigma$ to the variable $z_j$, which is just the $j^{th}$ coordinate of $\sigma$. For $1 \leq j \leq l$, let $\pi_j : D^l \rightarrow D^{l-1}$ be the projection operator that maps a string of length $l$ to its substring of length $l - 1$ obtained by dropping the $j^{th}$ coordinate.

**Definition 4.5** For any overlapping blocks $(B_1, B_2, B_3, B_4)$, characterized by $(\{x_j\}_{j=1}^{l-1}, \{y_j\}_{j=t}^{4}, t, \phi)$, and block assignments $\sigma^{(i)}$ to the blocks $B_j$s, we say that these block assignments are consistent if

1. $\pi_1(\sigma^{(1)}) = \pi_1(\sigma^{(2)}) = \pi_1(\sigma^{(3)}) = \pi_1(\sigma^{(4)})$

2. The values $\sigma^{(1)}(y_1), \sigma^{(2)}(y_2), \sigma^{(3)}(y_3), \sigma^{(4)}(y_4)$ satisfy the constraint $\phi$. 

10
In short, the first condition says that the assignments $\sigma^{(i)}$ must “project” down to a common assignment to the shared $(l - 1)$ coordinates, and the second condition says that the 4 values on the remaining coordinate must satisfy the constraint $\phi$.

Note that if $f : X \rightarrow D$ is an assignment that satisfies every constraint, and $f[B]$ is the restriction of this assignment to a block $B$, then for any overlapping blocks $(B_1, B_2, B_3, B_4)$, the block assignments $f[B_1], f[B_2], f[B_3], f[B_4]$ are consistent.

**Definition 4.6** For overlapping blocks $(B_1, B_2, B_3, B_4)$, and $k$ sets $F_1, F_2, \ldots, F_{k-3}, F^{(2)}, F^{(3)}, F^{(4)} \subseteq R$, we say that these $k$ sets are consistent if there exist block assignments $\sigma^{(i)}$ for block $B_i$, $1 \leq i \leq 4$, such that

1. $\sigma^{(1)} \in F_1 \cap F_2 \cap \ldots \cap F_{k-3}$
2. $\sigma^{(i)} \in F^{(i)}$ for $i = 2, 3, 4$.
3. The assignments $\sigma^{(i)}$ are consistent as per Definition 4.5.

**Remark:** Whenever we talk about consistency between sets $F_1, F_2, \ldots, F_{k-3}, F^{(2)}, F^{(3)}, F^{(4)}$, we have in mind a specific set of overlapping blocks $(B_1, B_2, B_3, B_4)$ which we will be clear from the context.

Now we are ready to define edges of the hypergraph. For overlapping blocks $(B_1, B_2, B_3, B_4)$, and sets $F_1, F_2, \ldots, F_{k-3}, F^{(2)}, F^{(3)}, F^{(4)}$ which are not consistent, we define

$$\{(B_1, F_j) | j = 1, 2, \ldots, k - 3\} \cup \{(B_i, F^{(i)}) | i = 2, 3, 4\}$$

to be an edge of the hypergraph. Thus every edge contains exactly $k$ vertices, i.e. this is a $k$-uniform hypergraph.

Let’s verify that this way of defining edges makes sense. Suppose $f : X \rightarrow D$ is an assignment that satisfies every constraint. We will show that the set $I_f$ (see Equation (4) ) is an independent set. As observed before, for any overlapping blocks $(B_1, B_2, B_3, B_4)$, the block assignments $f[B_i]$ are consistent. Let

$$\{(B_1, F_j) | j = 1, 2, \ldots, k - 3\} \cup \{(B_i, F^{(i)}) | i = 2, 3, 4\}$$

be any $k$ vertices in the set $I_f$. We will show that the sets

$$F_1, F_2, \ldots, F_{k-3}, F^{(2)}, F^{(3)}, F^{(4)}$$

are consistent and hence these $k$ vertices cannot form an edge, thus proving that $I_f$ is indeed an independent set. By definition of the set $I_f$, we have $f[B_1] \in F_1 \cap F_2 \cap \ldots \cap F_{k-3}$ and $f[B_i] \in F^{(i)}$ for $i = 2, 3, 4$. Since the assignments $f[B_i]$ are consistent, taking $\sigma^{(i)} = f[B_i]$ in Definition 4.6 proves the claim.

### 4.4 Completeness

**Lemma 4.7** If the instance $(X, \Phi)$ as in Definition 4.2 is an YES instance, then there exists an independent set of weight $p|B|$ in the hypergraph constructed above.

**Proof:** We will show that if there is a global assignment $f : X \rightarrow D$ that satisfies every constraint, then the hypergraph constructed in Section 4.3 has a “large” independent set. As observed in the last section, the set

$$I_f = \{(B, F) | B \in B, f[B] \in F\}$$
is an independent set. The weight of this set is

\[ \sum_{B \in \mathcal{B}} \sum_{F : F \subseteq R, f[B] \in F} \mu_p(F) = \sum_{B \in \mathcal{B}} p = p|B| \]

Thus in the completeness case, there exists an independent set of weight \( p|B| \).

4.5 Soundness

**Lemma 4.8** If the instance \((X, \Phi)\) as in Definition 4.2 is a NO instance with parameters \( L = |D|^{1-\gamma} \) and \( \gamma = \delta/4 \), then the hypergraph constructed above has no independent set of weight \( \delta|B| \).

Before we prove the above lemma, let us first verify that together with Lemma 4.7 it proves our main hardness result:

**Theorem 4.9** For every integer \( k \geq 5 \) and every \( \varepsilon > 0 \), the vertex cover problem on \( k \)-uniform hypergraphs is \( NP \)-hard to approximate within a factor of \((k - 3 - \varepsilon)\).

**Proof:** By Lemmas 4.7 and 4.8, together with Theorem 4.3, we have a gap of \((p|B|, \delta|B|)\) in the size of the independent set which corresponds to a gap \((1 - p)|B|, (1 - \delta)|B|\) in the size of the vertex cover. This is a factor \( \frac{1 - \delta}{p} = \frac{1 - \delta}{\varepsilon/\delta + \varepsilon} = k - 3 - \varepsilon' \) gap where \( \varepsilon' \to 0 \) as \( \varepsilon, \delta \to 0 \).

It remains to prove the soundness lemma.

**Proof of Lemma 4.8:** We will show that if the 4CSP instance \( L \) is a NO instance, then the hypergraph we constructed has no independent set of size \( \delta|B| \). Assume on the contrary that the hypergraph has an independent set of size \( \delta|B| \). Call this independent set \( I \). We will construct a collection of \( |D|^{1-\gamma} \) assignments to a set of variables \( Y, |Y| \geq \delta|X|/4 \) in the 4CSP such that every constraint inside \( Y \) is satisfied by some assignment.

For every block \( B \), define

\[ \mathcal{F}[B] \overset{\text{def}}{=} \{ F \mid F \subseteq R, (B, F) \in I \} \]

A simple averaging argument shows that for at least \( \delta/2 \) fraction of the blocks \( B \), we have \( \mu_p(\mathcal{F}[B]) \geq \delta/2 \). Defining

\[ B' \overset{\text{def}}{=} \{ B \mid B \in B, \mu_p(\mathcal{F}[B]) \geq \delta/2 \} \]

we have \( |B'| \geq \delta|B|/2 \).

**Lemma 4.10** For every \( B \in \mathcal{B} \), the family \( \mathcal{F}[B] \) can be assumed to be a monotone family of subsets of \( R \).

**Proof:** The way we define the edges of the hypergraph, it is easy to see that if \((B, F)\) is a vertex of an independent set then we can also add \((B, F')\) to the independent set provided \( F \subseteq F' \). Thus when the independent set is maximal, every family \( \mathcal{F}[B] \) is monotone.

Using this lemma, for every \( B \in \mathcal{B}' \), the family \( \mathcal{F}[B] \) is a monotone family with \( \mu_p(\mathcal{F}[B]) \geq \delta/2 \)

Let \( \zeta > 0 \) be a sufficiently small “accuracy” parameter which will be fixed later. Applying Theorem 4.1, we get

**Lemma 4.11** For every block \( B \in \mathcal{B}' \), there exists a real number \( p[B] \in (p, p + \zeta) \) and a set \( C[B] \subseteq R \) called the “core” with the following properties:
• $\text{as}_{p[B]}(\mathcal{F}[B]) \leq \frac{2}{\delta}$.
• The size of $C[B]$ is at most $\Delta_0$ which is a constant depending only on $k, \varepsilon, \zeta, \delta$.

Let $\mathcal{H}[B] \subseteq R \setminus C[B]$ defined as

$$
\mathcal{H}[B] \stackrel{\text{def}}{=} \{ H \mid H \subseteq R \setminus C[B], \ C[B] \cup H \in \mathcal{F}[B] \}
$$

Then we have $\mu_{p[B]}(\mathcal{H}[B]) \geq 1 - \zeta$, where the weight of the family $\mathcal{H}[B]$ is measured w.r.t. the $p[B]$-distribution on the universe $R \setminus C[B]$.

4.5.1 Incorporating All Elements With Some Influence: the Extended Core

Let $\eta > 0$ be a threshold parameter which will be chosen later. For every $B \in B'$, we identify a set of elements $\text{Infl}[B] \subseteq R$ that have significant influence on the family $\mathcal{F}[B]$, i.e.

$$
\text{Infl}[B] = \{ \sigma \in R \mid \text{Influence}_{p[B]}(\mathcal{F}[B], \sigma) \geq \eta \}
$$

Since $\mathcal{F}[B]$ has average sensitivity at most $\frac{2}{\delta}$ and the average sensitivity is simply the sum of influences of all the elements, it follows that the size of $\text{Infl}[B]$ is at most $\frac{2}{\delta \eta}$ which is a constant. Finally define the “extended core” $\text{Ecore}[B]$ as

$$
\text{Ecore}[B] \stackrel{\text{def}}{=} C[B] \cup \text{Infl}[B]
$$

Clearly, the extended core has size at most $\Delta = \Delta_0 + \frac{2}{\delta \eta}$.

4.5.2 The Preservation Property

Given two block assignments $\sigma, \sigma'$, and a projection $\pi_j : D^j \mapsto D^{j-1}, 1 \leq j \leq k$, we say that the two assignments are “preserved” if $\pi_j(\sigma) \neq \pi_j(\sigma')$. Since $\sigma, \sigma'$ differ in at least one coordinate, they will be preserved with probability $1 - \frac{1}{k}$ when a projection $\pi_j, 1 \leq j \leq l$ is picked at random.

For a block $B \in B'$, say that its extended core is preserved under projection $\pi_j$ if every pair of elements in the extended core is preserved. In other words, the projection operator is one-to-one on the extended core.

The extended core has size at most $\Delta$. Choosing $l = \Delta^2$, the probability that the extended core is preserved under a random projection $\pi_j$, is at least $1 - \frac{(\Delta^2)}{1} \geq \frac{1}{2}$. Hence there exists an index $j_0, 1 \leq j_0 \leq l$, such that for at least half of the the blocks in $B'$, their extended core is preserved. Assume w.l.o.g. that $j_0 = l$ and $\pi = \pi_l$ denote the projection operator which acts simply by dropping the last coordinate.

$$
B'' \stackrel{\text{def}}{=} \{ B \mid B \in B', \text{Ecore}[B] \text{ is preserved by } \pi \}
$$

As noted, $|B''| \geq |B'|/2 \geq \delta |B|/4$. A simple averaging argument shows that we can fix variables $x_1, x_2, \ldots, x_{l-1} \in X$ such that for at least $\delta/4$ fraction of variables $y$, we have $(x_1, x_2, \ldots, x_{l-1}, y) \in B''$. Define

$$
Y \stackrel{\text{def}}{=} \{ y \mid y \in X, (x_1, x_2, \ldots, x_{l-1}, y) \in B'' \}
$$

Thus we have $|Y| \geq \delta |X|/4$. Denote by $B_y$ the block $(x_1, x_2, \ldots, x_{l-1}, y)$.
4.5.3 Defining Assignments

Now we are ready to define assignments to the variables in set $Y$ so that every constraint inside $Y$ is satisfied by some assignment. There will be one assignment $f_\tau : Y \mapsto D$ for every $\tau \in D^{l-1}$. For $\tau \in D^{l-1}$ and $\alpha \in D$, let $\tau \alpha \in R = D^l$ be the concatenated string.

The assignment $f_\tau : Y \mapsto D$ is defined as

$$f_\tau(y) = \begin{cases} \alpha & \text{if } \exists \alpha \in D \text{ s.t. } \tau \alpha \in \text{Ecore}[B_y] \\ \text{undefined} & \text{otherwise} \end{cases}$$

There are two things to note here. Firstly, since the extended core is preserved, there exists at most one $\alpha \in D$ such that $\tau \alpha \in \text{Ecore}[B_y]$. Thus the definition of $f_\tau$ is unambiguous. Secondly, though the assignment $f_\tau$ is undefined for some (or even all) of the variables in $Y$, we will still show that for every constraint $\phi$ inside $Y$, there exists an assignment $f_\tau$ such that it satisfies the constraint $\phi$. We prove this in the next section.

4.5.4 Finishing the Proof

In this section, we will show that for every constraint $\phi$ inside the set of variables $Y$, there exists an assignment $f_\tau$ that satisfies this constraint. Let $\phi$ be a constraint on the variables $\{y_1, y_2, y_3, y_4\}$ and consider the blocks

$$B_i = B_{y_i} = \{x_1, x_2, \ldots, x_{l-1}, y_i\}$$

Clearly, the blocks $(B_1, B_2, B_3, B_4)$ are overlapping. We prove our claim in several steps.

**Lemma 4.12** There exist sets $F_1', F_2', \ldots, F_{k-3}' \subseteq R \setminus C[B_1]$ such that

- $\bigcap_{j=1}^{k-3} F_j' = \phi$
- $F_j' \overset{\text{def}}{=} C[B_1] \cup F_j' \in F[B_1]$ for $1 \leq j \leq k-3$.

In particular, $\bigcap_{j=1}^{k-3} F_j = C[B_1]$.

**Proof:** From Lemma 4.11, the weight of the family $\mathcal{H}[B_1]$ w.r.t. the bias parameter $p[B]$ is at least $1 - \zeta$. Noting that $p[B] \leq 1 - \frac{1}{k-3} - \frac{\xi}{\Delta}$ and applying Lemma A.4, there exist sets $F_j' \subseteq R \setminus C[B_1], 1 \leq j \leq k-3$ whose intersection is empty. By definition of the family $\mathcal{H}[B_1]$, the sets $F_j' \overset{\text{def}}{=} C[B_1] \cup F_j' \in F[B_1]$. 

Define

$$S \overset{\text{def}}{=} \{\sigma \in R \mid \text{there exists } \sigma' \in C[B_1] \text{ such that } \pi(\sigma) = \pi(\sigma')\}$$

That is, $S$ is the set of all strings which share a common prefix of length $l - 1$ with some string in $C[B_1]$. Clearly $|S| = |D| \cdot |C[B_1]| \leq |D| \cdot \Delta_0$. For $i = 2, 3, 4$ define

$$T_i \overset{\text{def}}{=} S \setminus \text{Ecore}[B_i]$$

Thus $T_i$ is a set of size at most $|D| \cdot \Delta_0$. By definition of the extended core (at the end of the Section 4.5), all elements of the set $T_i$ have influence at most $\eta$ on the family $\mathcal{F}[B_i]$ w.r.t. bias $p[B_i]$. Applying Lemma A.5, if $\eta$ is small enough, there exists a set $F^{(i)} \in \mathcal{F}[B_i]$ such that $F^{(i)} \cap T_i = \phi$.

Now consider the following vertices of the hypergraph:

$$\{(B_1, F_j) \mid 1 \leq j \leq k-3\} \cup \{(B_i, F^{(i)}) \mid i = 2, 3, 4\}$$
There vertices are in the independent set $I$. Hence there cannot be an edge on these vertices. Therefore the sets $F_1, \ldots, F_j, F^{[2]}, F^{[3]}, F^{[4]}$ are consistent (see Definition 4.6). This means that there exist block assignments $\sigma^{(i)} \in R$ such that

- $\sigma^{(1)} \in \bigcap_{j=1}^{k-3} F_j = C[B_1]$ (by Lemma 4.12).
- For $i = 2, 3, 4$, $\sigma^{(i)} \in F^{(i)}$.
- The block assignments $\sigma^{(i)}$ have the same prefix of length $l-1$, i.e. there is a string $\tau \in D^{l-1}$, and values $\alpha_4 \in D$ such that $\sigma^{(i)} = \tau \alpha_4$.
- The values $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ satisfy the constraint $\phi$.

**Lemma 4.13** $\sigma^{(i)} \in E_{\text{core}}[B_i]$ for $i = 1, 2, 3, 4$.

**Proof:** We have $\sigma^{(1)} \in C[B_1] \subseteq E_{\text{core}}[B_1]$. Now consider $i = 2, 3, 4$. Since $\sigma^{(i)}$ has the same $(l-1)$-prefix with $\sigma^{(1)}$, by definition of the set $S$, $\sigma^{(i)} \in S$. Also $\sigma^{(i)} \in F^{(i)}$ and $F^{(i)} \cap T_i = \phi$. Therefore $\sigma^{(i)} \in E_{\text{core}}[B_i]$.

From this lemma, and the way the assignment $f_\tau$ is defined, we have $f_\tau(y_i) = \alpha_i$ for $i = 1, 2, 3, 4$. Since the values $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ satisfy the constraint $\phi$, it follows that the assignment $f_\tau$ satisfies the constraint $\phi$. This finishes the proof of Lemma 4.8.

## 5 Improved Result Assuming Hardness of Graph Coloring

A closer inspection of our proof in the previous section shows that the slack of 3 in our hardness result comes from the fact that CSP we started with had constraints that depend on 4 variables. Consequently, we needed to split the $k$-sets (that comprise a hyperedge) amongst four blocks, and had to get a small core as the intersection of $(k-3)$ sets belonging to any set family with substantial weight under $\mu_p$. This in turn limits the bias parameter $p$ to be at most $(1 - \frac{1}{k-3})$, leading to a factor $(k-3 - \varepsilon)$ hardness. If we had a 2CSP (where each constraint depends only on two variables) for which a hardness similar to Theorem 4.3 holds, then we will be able to get a hardness of approximation factor of $(k-1 - \varepsilon)$. This is because we will only need a small core as the intersection of $(k-1)$ sets, and can therefore pick the bias parameter $p$ to be $(1 - \frac{1}{k-1} - \varepsilon)$. The rest of the analysis remains unchanged.

Unfortunately, we still seem to be quite far from proving a result like Theorem 4.3 for 2CSPs. However, such a result follows if a strong hardness assumption on approximate graph coloring holds. The following lemma makes formal this connection. The proof is straightforward and uses ideas similar to that of Theorem 4.3.

**Lemma 5.1** Suppose that there exists a $c \geq 3$ such that for every positive integer $b$ it is NP-hard to $(\log n)$-color a $c$-colorable graph on $n$-vertices. Then there exists $d \geq 2$ such that for every integer $L$ and every constant $\gamma > 0$, given a 2CSP over domain size $d$, no polynomial time algorithm can tell if it is an YES instance or a NO instance unless $P = NP$ (where YES and NO instances are defined for a particular $L, \gamma$ as in Definition 4.2).

We note that the above hardness follows by a Turing reduction (unlike the many-one reductions presented in the rest of the paper). We therefore have a strong hardness result for $k$-Vertex-Cover based on the hardness assumption for graph coloring.
Theorem 5.2 Suppose that there exists a $c \geq 3$ such that for every positive integer $b$ it is NP-hard to $(b \log n)$-color a $c$-colorable graph on $n$-vertices. Then, for every integer $k \geq 3$ and $\varepsilon > 0$, there is no polynomial time factor $(k - 1 - \varepsilon)$ algorithm for the vertex cover problem on $k$-uniform hypergraphs unless $P = NP$.

6 Future Work

The vertex cover in every $k$-uniform hypergraph can be approximated to within factor $k - o(1)$, [Hal00]. An obvious open problem is that of improving our $(k - 3 - \varepsilon)$ bound and obtaining an 'optimal' inapproximability factor of $(k - \varepsilon)$, i.e. proving NP-hardness of approximating Ek-Vertex-Cover within a factor of $(k - \varepsilon)$ for any constant $\varepsilon > 0$. This problem is especially interesting for small values of $k$, as the $k = 2$ case (i.e. Vertex-Cover on graphs) has received a good deal of attention yet leaving the factor $(2 - \varepsilon)$ hardness result still out of reach.

Following this work, Dinur, Guruswami, Khot and Regev [DGKR02] were able to improve our work and show a hardness-of-approximation factor of $(k - 1 - \varepsilon)$.

Another possible direction is to extend these results for larger values of $k$. The largest plausible value of $k$ is $\ln n$ since the greedy set-cover algorithm can always be used to achieve a $(\ln n + 1)$ approximation on any hypergraph (here $n$ is the number of edges in the hypergraph). Our hardness result from Section 3 gives an $\Omega(k)$ inapproximability factor (assuming $NP \not\subseteq \text{DTIME}(2^{\log^{o(1)} n})$) for $k$ up to $\log^2 n$ for some absolute constant $\gamma > 0$. We conclude with the following conjecture:

Conjecture 6.1 It is NP-hard to approximate Ek-Vertex-Cover to within $k \cdot (1 - \varepsilon)$ for any $k \leq \ln n$ and any constant $\varepsilon > 0$.

References


A Some Useful Lemmas

A.1 Combinatorial Core

Lemma 2.2 (Combinatorial Core) For every ε, δ > 0, there exists some t = t(ε, δ) > 0 such that for every finite R and $\mathcal{F} \subset P(R)$, if $\mu_{\frac{1}{2} - \delta}(\mathcal{F}) > \varepsilon$, then there exists some 'core' subset $C \subset \mathcal{F}$ with $|C| \leq t$.

Proof: We begin by stating a continuous variant of the complete intersection theorem of Ahlswede and Khachatrian. This was already proven in [DS02] for $t = 2$ and the extension for larger $t$ is straightforward.

Define for every $i \geq 0$, $t > 0$ and $n \geq 2i + t$,

$$\mathcal{A}_{i,t} \overset{\text{def}}{=} \{ F \in P([n]) \mid |F \cap [1, t + 2i]| \geq t + i \}.$$

Clearly, for any $n' > n \geq 2i + t$, $\mu_p(\mathcal{A}_{i,t}^{n'}) = \mu_p(\mathcal{A}_{i,t}^n)$. Denoting $\binom{n}{k} \overset{\text{def}}{=} \{ F \subset [n] \mid |F| = k \}$, the complete intersection theorem of Ahlswede and Khachatrian states that

Theorem A.1 ([AK97]) Let $\mathcal{F} \subset \binom{[n]}{k}$ be t-intersecting (i.e. for every $F_1, F_2 \in \mathcal{F}$, $|F_1 \cap F_2| \geq t$). Then,

$$|\mathcal{F}| \leq \max_{0 \leq i \leq \frac{n}{2k}} |\mathcal{A}_{i,t} \cap \binom{[n]}{k}|.$$
The following lemma is a continuous variant of the above theorem,

**Lemma A.2** [DS02] Let \( \mathcal{F} \subseteq P([n]) \) be \( t \)-intersecting. For any \( p < \frac{1}{2} \),

\[
\mu_p(\mathcal{F}) \leq \max_i \{ \mu_p(A^n_{i,t}) \}
\]

We define for every \( t > 0 \) and \( p < \frac{1}{2} \), let \( a_{p,t} \) \( \overset{df}{=} \max_i (\mu_p(A^n_{i,t})) \). In order to prove Lemma 2.2 it suffices to prove that for a fixed \( p < \frac{1}{2} \), \( \limsup_{t \to \infty} a_{p,t} = 0 \).

Note that \( A^n_{i,t} \subseteq \{ F \in P([n]) \mid |F \cap [1, t+2i]| \geq (t+2i)/2 \} \). Define \( \mathcal{F}_{i,t} \) to be the family

\[
\{ F \in P([t+2i]) \mid |F| \geq (t+2i)/2 \}.
\]

We then have

\[
\mu_p(A^n_{i,t}) \leq \mu_p(\mathcal{F}_{i,t}).
\]

Now sets in \( \mathcal{F}_{i,t} \) contain at least a fraction 1/2 of the universe \([t+2i]\), while a random set drawn according to the product distribution \( \mu_p \) has an expected fraction \( p < 1/2 \) of elements. By standard Chernoff bounds the probability of a set picked according to \( \mu_p \) landing in \( \mathcal{F}_{i,t} \) is exponentially small in \( t \) and thus tends to zero as \( t \to \infty \). Hence \( \mu_p(\mathcal{F}_{i,t}) \) tends to zero as \( t \to \infty \) (for every \( i \)). Together with (5) this shows that for each fixed \( p < 1/2 \), \( \limsup_{t \to \infty} a_{p,t} = 0 \).

---

**A.2 \( k \)-wise Intersecting Families**

We will use the following theorem of Frankl.

**Theorem A.3** Let \( \mathcal{F} \subseteq P(R) \) where \( |R| = n \) and every set in the family \( \mathcal{F} \) has size \( m \). Assume that every \( k \) sets in the family have nonempty intersection and \( n > mk/(k-1) \). Then

\[
|\mathcal{F}| \leq \binom{n-1}{m-1}
\]

Note that a family of sets of size \( m \) containing one fixed element has size \( \binom{n-1}{m-1} \). We will use the above theorem to prove:

**Lemma A.4** Let \( \varepsilon > 0 \) be an arbitrarily small constant, \( k \geq 2 \) an integer and \( p = 1 - \frac{1}{k} - \varepsilon \). Let \( \mathcal{F} \subseteq P(R) \) be a family such that every \( k \) sets in this family have a nonempty intersection. Then

\[
\mu_p(\mathcal{F}) < p + \varepsilon
\]

provided the universe \( R \) is sufficiently large.

**Proof:** Let \( n = |R| \) be the size of the universe. Partition the family \( \mathcal{F} \) according to different set-sizes.

\[
\mathcal{F}_i \overset{df}{=} \{ F \mid F \in \mathcal{F}, \ |F| = i \}
\]

With the bias parameter \( p \), the total weight of all sets of size more than \( (p + \frac{\varepsilon}{2})n \) is at most \( \frac{\varepsilon}{2} \) when the universe is large enough. Hence

\[
\mu_p(\mathcal{F}) \leq \frac{\varepsilon}{2} + \sum_{m \leq (p + \frac{\varepsilon}{2})n} \mu_p(\mathcal{F}_m)
\]

18
For $m \leq (p + \frac{\varepsilon}{2})n$, we have $n > mk/(k - 1)$. Since every $k$ sets in the family $\mathcal{F}_m$ have a nonempty intersection, applying Frankl’s Theorem, we get

$$|\mathcal{F}_m| \leq \binom{n-1}{m-1}$$

Noting that every set in $\mathcal{F}_m$ has weight $p^m(1 - p)^{n-m}$ we have

$$\mu_p(\mathcal{F}) \leq \frac{\varepsilon}{2} + \sum_{m \leq (p + \frac{\varepsilon}{2})n} \binom{n-1}{m-1} p^m (1 - p)^{n-m}$$

$$\leq \frac{\varepsilon}{2} + p \left( \sum_m \binom{n-1}{m-1} p^{m-1} (1 - p)^{(n-1)-(m-1)} \right)$$

$$= \frac{\varepsilon}{2} + p$$

\[ \blacksquare \]

### A.3 Very Small Influence

The following lemma can be found in [DS02].

**Lemma A.5** Let $\mathcal{F} \subseteq P(R)$ be a monotone family. Let $T$ be a set of elements such that for every element $\sigma \in T$, $\text{Influence}_p(\mathcal{F}, \sigma) < \eta$. Assume $\eta$ is small enough so that

$$|T| \cdot \eta \cdot p^{-|T|} < \mu_p(\mathcal{F})$$

Then there exists a set $F \in \mathcal{F}$ such that $F \cap T = \emptyset$. 

---

19