Abstract

A pair of clauses in a CNF formula constitutes a conflict if there is a variable that occurs positively in one clause and negatively in the other. A CNF formula without any conflicts is satisfiable. The Lovász Local Lemma implies that a $k$-CNF formula is satisfiable if each clause conflicts with at most $\frac{2k}{e} - 1$ clauses. It does not, however, give any good bound on how many conflicts an unsatisfiable formula has globally.

We show here that every unsatisfiable $k$-CNF formula requires $\Omega(2^{0.69k})$ conflicts and there exist unsatisfiable $k$-CNF formulas with $O(3.51^k)$ conflicts.

1 Introduction

A boolean formula in conjunctive normal form, a CNF formula for short, is a conjunction (AND) of clauses, which are disjunctions (OR) of literals. A literal is either a boolean variable $x$ or its negation $\bar{x}$. We assume that a clause does neither contain the same literal twice nor a variable and its negation. A CNF formula where each clause contains exactly $k$ literals is called a $k$-CNF formula. Satisfiability, the problem of deciding whether a CNF formula is satisfiable, plays a major role in computer science. How can a $k$-CNF formula be unsatisfiable? If $k$ is large, each clause is extremely easy to satisfy individually. However, it can be that there are conflicts between the clauses, making it impossible to satisfy all of them simultaneously. If a $k$-CNF formula is unsatisfiable, then we expect that there are many conflicts.

To give a formal setup, we say two clauses conflict if there is at least one variable that appears positively in one clause and negatively in the other. For example, the two clauses $(x \lor y)$ and $(\bar{x} \lor u)$ conflict, as well as $(x \lor y)$ and $(\bar{x} \lor \bar{y})$ do. Any CNF formula without the empty clause and without any conflicts is satisfiable. For a formula $F$ we define the conflict graph $CG(F)$, whose vertices are the clauses of $F$, and two clauses are connected by an edge if they conflict. $\Delta(F)$ denotes the maximum degree of $CG(F)$, and $e(F)$ the number of conflicts in $F$, i.e., the number of edges in $CG(F)$. In fact, any $k$-CNF formula is satisfiable unless $\Delta(F)$ and $e(F)$ are large. A quantitative result follows from the lopsided Lovász Local Lemma [4, 1, 8]: A $k$-CNF formula $F$ is satisfiable unless some clause conflicts with $\frac{2k}{e}$ or more clauses, i.e., unless $\Delta(F) \geq \frac{2k}{e}$. Up to a constant factor, this is tight: Consider the formula containing all $2^k$ clauses over the variables $x_1, \ldots, x_k$, the complete $k$-CNF formula which we denote by $K_k$. It is unsatisfiable, and $\Delta(K_k) = 2^k - 1$. 

*Research is supported by the SNF Grant 200021-118001/1
As its name suggests, the lopsided Lovász Local Lemma implies a local result. Our goal is to obtain a global result: \( F \) is satisfiable unless the total number of conflicts is very large. We define two functions

\[
\begin{align*}
\text{lc}(k) & := \max \{ d \in \mathbb{N}_0 \mid \text{every } k\text{-CNF formula } F \text{ with } \Delta(F) \leq d \text{ is satisfiable} \}, \\
\text{gc}(k) & := \max \{ e \in \mathbb{N}_0 \mid \text{every } k\text{-CNF formula } F \text{ with } e(F) \leq e \text{ is satisfiable} \}.
\end{align*}
\]

The abbreviations lc and gc stand for local conflicts and global conflicts, respectively. From the above discussion, \( 2^{k/2} - 1 \leq \text{lc}(k) \leq 2^k - 2 \), hence we know \( \text{lc}(k) \) up to a constant factor. In contrast, it does not seem to be easy to prove nontrivial upper and lower bounds on \( \text{gc}(k) \). Certainly, \( \text{gc}(k) \geq \text{lc}(k) \geq 2^{k/2} - 1 \) and \( \text{gc}(k) \leq e(K_k) - 1 = \left( \frac{2^k}{2} \right) - 1 \). Ignoring constant factors, \( \text{gc}(k) \) lies somewhere between \( 2^k \) and \( 4^k \). This leaves much space for improvement. In [10], we proved that \( \text{gc}(k) \in \Omega(2.27^k) \) and \( \text{gc}(k) \leq \frac{4^k}{\log k} k \). In this paper, we improve upon these bounds. Surprisingly, \( \text{gc}(k) \) is exponentially smaller than \( 4^k \).

**Theorem 1.** Any unsatisfiable \( k \)-CNF formula contains \( \Omega(2.69^k) \) conflicts. There are unsatisfiable \( k \)-CNF formulas with \( O(3.51^k) \) conflicts.

We obtain the lower bound by a more sophisticated application of the idea we used in [10]. The upper bound follows from a construction that is partially probabilistic, and inspired in parts by Erdős’ construction in [3] of sparse \( k \)-uniform hypergraphs that are not 2-colorable. To simplify notation, we view formulas as sets of clauses, and clauses as sets of literals. Hence, \( |F| \) denotes the number of clauses in \( F \). Still, we will sometimes find it convenient to use the more traditional logic notation.

**Related Work**

Let \( F \) be a CNF formula and \( u \) be a literal. We define \( \text{occ}_F(u) := |\{ C \in F \mid u \in C \}| \). For a variable \( x \), we write \( d_F(x) = \text{occ}_F(x) + \text{occ}_F(\bar{x}) \) and call it the degree of \( x \). We write \( d(F) = \max_x d_F(x) \). It is easy to see that for a \( k \)-CNF formula, \( \Delta(F) \leq k(d(F) - 1) \). Define

\[
f(k) := \max \{ d \in \mathbb{N}_0 \mid \text{every } k\text{-CNF formula } F \text{ with } d(F) \leq d \text{ is satisfiable} \}.
\]

By an application of Hall’s Theorem, Tovey [11] showed that every \( k \)-CNF formula \( F \) with \( d(F) \leq k \) is satisfiable, hence \( f(k) \geq k \). Later, Kratochvíl, Savický and Tuza [7] showed that \( f(k) \geq \frac{2^k}{\log k} \) and \( f(k) \leq 2^{k-1} - 2^{k-4} - 1 \). The upper bound was improved by Savický and Sgall [9] to \( f(k) \in O(k^{-0.26} 2^k) \), by Hoory and Szeider [6] to \( f(k) \in O\left(\frac{\log(k)2^k}{k}\right)\), and recently by Gebauer [5] to \( f(k) \leq \frac{2^{k+2}}{k} - 1 \), closing the gap between lower and upper bound on \( f(k) \) up to a constant factor. Actually, we used the formulas constructed in [6] to prove the upper bound \( \text{gc}(k) \leq \frac{4^k}{\log^2 k} k \) in [10].

## 2 A First Attempt

We sketch a first attempt on proving a nontrivial lower bound on \( \text{gc}(k) \). Though this attempt does not succeed, it leads us to other interesting questions, results, and finally proof methods which can be used to prove a lower bound on \( \text{gc}(k) \). Let \( F \) be a \( k \)-CNF
formula and \( x \) a variable. Every clause containing \( x \) conflicts with every clause containing \( \overline{x} \), thus \( e(F) \geq \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}) \). Furthermore,
\[
e(F) \geq \frac{1}{k} \sum_x \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}),
\]
where the \( \frac{1}{k} \) comes from the fact that each conflict might be counted up to \( k \) times, if two clauses contain several complementary literals. Every unsatisfiable \( k \)-CNF formula \( F \) contains a variable \( x \) with \( d_F(x) \geq \frac{2^k}{ck} \). If this variable is balanced, i.e., \( \text{occ}_F(x) \) and \( \text{occ}_F(\overline{x}) \) differ only in a polynomial factor in \( k \), then \( e(F) \geq \frac{4^k}{\text{poly}(k)} \). Indeed, in the formulas constructed in [5], all variables are balanced. The same holds for the complete \( k \)-CNF formula \( K_k \). It follows that when trying to obtain an upper bound on \( gc(k) \) that is exponentially smaller than \( 4^k \), we should construct a very unbalanced formula. We ask the following question:

**Question:** Is there a number \( a > 1 \) such that for every unsatisfiable \( k \)-CNF formula \( F \) there is a variable with \( \text{occ}_F(x) \geq a^k \) and \( \text{occ}_F(\overline{x}) \geq a^k \)?

The answer is a very strong no: In [10] we gave a simple inductive construction of a \( k \)-CNF formula \( F \) with \( \text{occ}_F(\overline{x}) \leq 1 \) for every variable \( x \). However, in this formula one has \( \text{occ}_F(x) \approx k! \). Allowing \( \text{occ}_F(\overline{x}) \) to be a small exponential in \( k \), we have the following result:

**Theorem 2.** \( \textbf{(i)} \) For every \( a > 1 \), \( b \geq \frac{a}{a - 1} \) there is a constant \( c \) such that for all sufficiently large \( k \), there is an unsatisfiable \( k \)-CNF formula \( F \) with \( \text{occ}_F(\overline{x}) \leq \frac{ck^2a^k}{\text{poly}(k)} \) and \( \text{occ}_F(x) \leq \frac{ck^2b^k}{\text{poly}(k)} \), for all \( x \).

\( \textbf{(ii)} \) Let \( 1 < a < \sqrt{2} \) and \( b = \sqrt{\frac{a^2 - 1}{a^2}} \). Then every \( k \)-CNF formula \( F \) with \( \text{occ}_F(x) \leq \frac{b^k}{8k} \) and \( \text{occ}_F(\overline{x}) \leq \frac{a^k}{8k} \) is satisfiable.

Of course, we can interchange the roles of \( x \) and \( \overline{x} \), but it is convenient to assume that \( \text{occ}_F(\overline{x}) \leq \text{occ}_F(x) \) for every \( x \). In the spirit of these results, we might suspect that if \( F \) is unsatisfiable, then for some variable \( x \), the product \( \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}) \) is large.

**Question:** Is there a number \( a > 2 \) such that every unsatisfiable \( k \)-CNF formula contains a variable \( x \) with \( \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}) \geq a^k \)?

Clearly, \( gc(k) \geq a^k \) for any such number \( a \). The complete \( k \)-CNF formula witnesses that \( a \) cannot be greater than \( 4 \), and it is not at all easy to come up with an unsatisfiable \( k \)-CNF formula where \( \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}) \) is exponentially smaller than \( 4^k \) for every \( x \). We cannot answer the above question, but we suspect that the answer is yes. We prove an upper bound on the possible value of \( a \):

**Theorem 3.** There are unsatisfiable \( k \)-CNF formulas with \( \text{occ}_F(x) \cdot \text{occ}_F(\overline{x}) \in O(3.01^k) \) for all variables \( x \).

3 Proofs

For a truth assignment \( \alpha \) and a clause \( C \), we will write \( \alpha \models C \) if \( \alpha \) satisfies \( C \), and \( \alpha \not\models C \) if it does not. Similarly, if \( \alpha \) satisfies a formula \( F \), we write \( \alpha \models F \). We begin by stating a version of the Lopsided Lovász Local Lemma formulated in terms of satisfiability. See [10] for a derivation of this version.
Lemma 4 (SAT version of the Lopsided Lovász Local Lemma). Let $F$ be a CNF formula not containing the empty clause. Sample a truth assignment $\alpha$ by independently setting each variable $x$ to true with some probability $p(x) \in [0, 1]$. If for any clause $C \in F$, it holds that
\[
\sum_{D \in F: C \text{ and } D \text{ conflict}} \Pr[\alpha \not| D] \leq \frac{1}{4},
\]
then $F$ is satisfiable.

It is not possible to apply Lemma 4 directly to a formula $F$ which we want to prove being satisfiable. Instead, we apply it to a formula $F'$ we obtain from $F$ in the following way:

Definition 5. Let $F$ be a CNF formula. A truncation of $F$ is a CNF formula $F'$ that is obtained from $F$ by deleting some literals from some clauses.

For example, $(x \lor y) \land (\bar{y} \lor z)$ is a truncation of $(x \lor y \lor \bar{z}) \land (\bar{x} \lor y \lor z)$. A truncation of a $k$-CNF formula is not necessarily a $k$-CNF formula anymore. Any truth assignment satisfying a truncation $F'$ of $F$ also satisfies $F$. In our proofs, we will often find it easier to apply Lemma 4 to a special truncation of $F$ than to $F$ itself. We need a technical lemma on the binomial coefficient.

Lemma 6. Let $a, b \in \mathbb{N}$ with $b/a \leq 0.75$. Then
\[
\frac{a^b}{b!} \geq \binom{a}{b} > \frac{a^b}{b!} e^{-b^2/a}.
\]

Proof. The upper bound is trivial and true for all $a, b$. The lower bound follows like this.
\[
\binom{a}{b} = \frac{a(a-1) \cdots 1}{b!} = \frac{a^b}{b!} \prod_{j=0}^{b-1} \frac{a-j}{a} > \frac{a^b}{b!} e^{-2a/\sum_{j=0}^{b-1} j} \geq \frac{a^b}{b!} e^{-b^2/a},
\]
where we used the fact that $1 - x > e^{-2x}$ for $0 \leq x \leq 0.75$.  

3.1 Proof of Theorem 2 and 3

As we have argued in Section 2, in order to improve significantly upon the upper bound $gc(k) \leq 4^k$, we must construct a formula that is very unbalanced, i.e. $\text{occ}_F(x)$ is exponentially larger than $\text{occ}_F(\bar{x})$. First, we will construct an unsatisfiable CNF formula with $k$-clauses and some smaller clauses. In a second step, we expand all clauses to size $k$.

Definition 7. Let $F$ be a CNF formula with clauses of size at most $k$. For each $k'$-clause $C$ with $k' < k$, construct a complete $(k-k')$-CNF formula $K_{k-k'}$ over $k-k'$ new variables $y_1, \ldots, y_{k-k'}$. We replace $C$ by $C \lor K_{k-k'}$. Using distributivity, we expand it into a $k$-CNF formula $G$ called a $k$-CNFification of $F$.

For example, a 3-CNFFication of $(x \lor y) \land (\bar{x} \lor y \lor z)$ is $(x \lor y \lor y_1) \land (x \lor y \lor \bar{y}_1) \land (\bar{x} \lor y \lor z)$. A truth assignment satisfies $F$ if and only if it satisfies its $k$-CNFification $G$.

Definition 8. Let $\ell, k \in \mathbb{N}_0$. An $(\ell, k)$-CNF formula is a formula consisting of $\ell$-clauses containing only positive literals, and $k$-clauses containing only negative literals.
If $F$ is an $(\ell, k)$-CNF formula, we write $F = F^+ \land F^-$, where $F^+$ consists of the positive $\ell$-clauses and $F^-$ of the negative $k$-clauses.

**Proposition 9.** Let $\ell \leq k$ and let $F = F^+ \land F^-$ be an $(\ell, k)$-CNF formula. Let $G$ be the $k$-CNFication of $F$. Then

(i) $e(G) \leq 4^{k-\ell}|F^+| + 2^{k-\ell}|F^+| \cdot |F^-|$, 

(ii) $\text{occ}_G(x) \cdot \text{occ}_G(\bar{x}) \leq \max\{4^{k-\ell}, 2^{k-\ell}|F^+| \cdot |F^-|\}$.

**Proof.** To prove (i), note that every edge in $CG(F)$ runs between a positive $\ell$-clause $C$ and a negative $k$-clause $D$. Thus, $e(F) \leq |F^+| \cdot |F^-|$. In $G$, this edge is replaced by $2^{k-\ell}$ edges, since $C$ is replaced by $2^{k-\ell}$ copies. This explains the term $2^{k-\ell}|F^+| \cdot |F^-|$. Replacing $C$ by $2^{k-\ell}$ many $k$-clauses introduces at most $4^{k-\ell}$ new conflicts. This explains the term $4^{k-\ell}|F^+|$, and proves (i). To prove (ii), there are two cases. First, if $x$ appears in $F$, then $\text{occ}_G(x) = \text{occ}_F(\bar{x})$ and $\text{occ}_G(x) = \text{occ}_F(x)2^{k-\ell}$, thus $\text{occ}_G(x)\text{occ}_G(\bar{x}) \leq 2^{k-\ell}|F^+| \cdot |F^-|$. Second, if $x$ appears in $G$, but not in $F$, then $\text{occ}_G(x) = \text{occ}_G(\bar{x}) = 2^{k-\ell}-1$, and $\text{occ}_G(x) \cdot \text{occ}_G(\bar{x}) \leq 4^{k-\ell}$.

We should explore for which values of $|F^+|$ and $|F^-|$ there are unsatisfiable $(\ell, k)$-CNF formulas. We can then use Proposition 9 to derive upper bounds.

**Lemma 10.** For any $\rho \in (0, 1)$, there is a constant $c$ such that for all $k \in \mathbb{N}_0$ and $\ell \leq k$, there exists an unsatisfiable $(\ell, k)$-CNF formula $F = F^+ \land F^-$ with $|F^-| \leq ck^2 \rho^{-k}$ and $|F^+| \leq c \rho^{2(1-\rho)^{\ell}}$.

**Proof.** We choose a set variables $V = \{x_1, \ldots, x_n\}$ of $n = k^2$ variables. There are $\binom{n}{k}$ $k$-clauses over $V$ containing only negative literals. We form $F^-$ by sampling $ck^2 \rho^{-k}$ of them, uniformly with replacement, and similarly, we form $F^+$ by sampling $ck^2(1-\rho)^\ell$ purely positive $\ell$-clauses, where $c$ is some suitable constant determined later. Set $F = F^- \land F^+$. We claim that with high probability, $F$ is unsatisfiable. Let $\alpha$ be any truth assignment. There are two cases.

*Ccase 1.* $\alpha$ sets at least $\rho n$ variables to true. For a random negative clause $C$,

$$\Pr[\alpha \not= C] \geq \frac{\binom{\rho n}{k}}{\binom{n}{k}} \geq c' \rho^k,$$

The last inequality follows from Lemma 6. Since we select the clauses of $F^-$ independently of each other, we obtain

$$\Pr[\alpha \mid F^-] \leq (1 - c' \rho^k)^{ck^2 \rho^{-k}} < e^{-ck^2} = e^{-k^2},$$

provided we chose $c$ large enough, i.e., $c \geq \frac{1}{\alpha}$.

*Ccase 2:* $\alpha$ sets at most $\frac{n}{\alpha}$ variables to true. Now a similar calculation shows that $\alpha$ satisfies $F^+$ with probability at most $e^{-k^2}$.

In any case, $\Pr[\alpha \mid F] \leq e^{-k^2}$. The expected number of satisfying assignments of $F$ is at most $2^{k^2}e^{-k^2} \ll 1$ and with high probability $F$ is unsatisfiable.

The bound in Lemma 10 is tight up to a polynomial factor in $k$:

**Lemma 11.** Let $F = F^+ \land F^-$ be an $(\ell, k)$-CNF formula. If there is a $\rho \in (0, 1)$ such that $|F^+| < \frac{1}{2}(1-\rho)^{-\ell}$ and $|F^-| < \frac{1}{2}\rho^{-k}$, then $F$ is satisfiable.
Proof. Sample a truth assignment \( \alpha \) by setting each variable independently to true with probability \( \rho \). For a negative \( k \)-clause \( C \), it holds that \( \Pr[\alpha \not\models C] = \rho^k \). Similarly, for a positive \( \ell \)-clause \( D \), \( \Pr[\alpha \not\models D] = (1 - \rho)^\ell \). Hence the expected number of clauses in \( F \) that are unsatisfied by \( \alpha \) is \( \rho^k |F^-| + (1 - \rho)^\ell |F^+| < \frac{1}{2} + \frac{1}{2} = 1 \). Therefore, with positive probability \( \alpha \) satisfies \( F \).

Proof of Theorem 2. (i) Apply Lemma 10 with \( \ell = k \) and \( \rho = \frac{1}{2} \).

(ii) We fix some probability \( p := \frac{1}{\alpha^2} \geq \frac{1}{2} \), and set every variable of \( F \) to true with probability \( p \), independent of each other. This gives a random truth assignment \( \alpha \). We define a truncation \( F' \) of \( F \) as follows: For each clause \( C \in F \), if at least half the literals of \( C \) are negative, we remove all positive literals from \( C \) and insert the truncated clause into \( F' \), otherwise we insert \( C \) into \( F' \) without truncating it. We write \( F' = F_k \wedge F^- \), where \( F^- \) consists of purely negative clauses of size at least \( \frac{k}{2} \), and \( F_k \) consists of \( k \)-clauses, each containing at least \( \frac{k}{2} \) positive literals. A clause in \( F^- \) is unsatisfied with probability at most \( p^k \), and a clause in \( F_k \) with probability at most \( p^k (1 - p)^\frac{k}{2} \). This is because in the worst case, half of all literals are negative: Since \( p \geq \frac{1}{2} \), negative literals are more likely to be unsatisfied than positive ones. Let \( C \in F' \) be any clause. A positive literal \( x \in C \) causes conflicts between \( C \) and the \( \text{occ}_{F'}(x) \leq \frac{k}{\text{occ}_F} \) clauses of \( F' \) containing \( \bar{x} \). Similarly, a negative literal \( \bar{y} \in C \) causes conflicts with the at most \( \frac{b k}{\text{occ}_F} \) clauses of \( F_k \) containing \( y \). Therefore

\[
\sum_{D \in F' \text{ and } C \text{ conflict}} \Pr[\alpha \not\models D] \leq \frac{a^k}{8} p^k + \frac{b^k}{8} (1 - p)^\frac{k}{2} = \frac{1}{4},
\]

since \( p = \frac{1}{\alpha^2} \) and \( b = \sqrt{\frac{a^k}{\alpha^{2\ell}} - 1} \). By Lemma 4, \( F' \) is satisfiable.

Part (ii) of Theorem 2 can easily be improved by defining a more careful truncation procedure: We remove all positive literals from a clause \( C \) if \( C \) contains less than \( \lambda k \) of them, for some \( \lambda \in (0, 1] \). Choosing \( \lambda \) and \( p \) optimally, we obtain a better result, but the calculations become messy, and it offers no additional insight. The crucial part of the proof is that by removing positive literals from a clause, we can use the fact that \( \text{occ}_F(x) \) is small to bound the number of clauses \( D \) that conflict with \( C \) and have a large probability of being unsatisfied. This is also the main idea in our proof of the lower bound of Theorem 1. It should be pointed out that for \( k = \ell \), an \((\ell, k)\)-CNF formula is just a monotone \( k \)-CNF formula. The size of a smallest unsatisfiable monotone \( k \)-CNF formula is the same—up to a factor of at most 2—as the minimum number of hyperedges in a \( k \)-uniform hypergraph that is not 2-colorable. In 1963, Erdős [2] raised the question what this number is, and proved lower bound of \( 2^{k-1} \) (this is easy, simple choose a random 2-coloring). One year later, he [3] gave a probabilistic construction of a non-2-colorable \( k \)-uniform hypergraph using \( c k^{2k} \) hyperedges. For \( k = \ell \) and \( \rho = \frac{1}{2} \), the statement and proof of Lemma 10 are basically the same in [3].

Proof of Theorem 3. Combining Lemma 10 and Proposition 9, we conclude that for any \( \rho \in (0, 1) \) and \( 0 \leq \ell \leq k \), there is an unsatisfiable \( k \)-CNF formula \( F \) with

\[
\text{occ}_F(x) \cdot \text{occ}_F(\bar{x}) \leq \max\{4^{k-\ell}, 2^{k-\ell} c k^4 \rho^{-k}(1 - \rho)^{-\ell}\},
\]

for every variable \( x \). The constant \( c \) depends on \( \rho \), but not on \( k \) or \( \ell \). For fixed \( k, \ell > 1 \), the term \( \rho^{-k}(1 - \rho)^{-\ell} \) is minimized for \( \rho = \frac{k}{k+\ell} \). Choosing \( \ell = [0.2055k] \), we get \( \rho \approx 0.83 \) and \( \text{occ}_F(x) \cdot \text{occ}_F(\bar{x}) \in O(3.01^k) \).
3.2 Proof of the Main Theorem

Proof of the upper bound of Theorem 1. As in the previous proof, Proposition 9 together with Lemma 10 yield an unsatisfiable $k$-CNF formula $F$ with

$$e(F) \leq 4^{k-\ell}ck^2(1-\rho)^{-\ell} + 2^{k-\ell}c^2k^4\rho^{-k}(1-\rho)^{-\ell}.$$ 

For $\rho \approx 0.6298$ and $\ell = \lceil 0.333k \rceil$, we obtain $e(F) \in O(3.51^k)$.

Proof of the lower bound in Theorem 1. Let $F$ be an unsatisfiable $k$-CNF and let $e(F)$ be the number of conflicts in $F$. We will show that $e(F) \in \Omega(2.69^k)$. In the proof, $x$ denotes a variable and $u$ a positive or negative literal. We assume $\text{occ}_F(x) \leq \text{occ}_F(\bar{x})$ for all variables $x$. We can do so since otherwise we just replace $x$ by $\bar{x}$ and vice versa. This changes neither $e(F)$, nor satisfiability of $F$. Also we can assume that $\text{occ}_F(x)$ and $\text{occ}_F(\bar{x})$ are both at least 1, if $x$ occurs in $F$ at all. For $x$, we define

$$p(x) := \max \left\{ \frac{1}{2}, \sqrt[16]{\text{occ}_F(x) / 16e(F)} \right\},$$

and set $x$ to true with probability $p(x)$ independently of all other variables yielding a random assignment $\alpha$. Since $\text{occ}_F(u) \leq e(F)$, we have $p(x) \leq 1$. We set $p(\bar{x}) = 1 - p(x)$. By definition, $p(x) \geq p(\bar{x})$. Let us list some properties of this distribution. First, if $p(u) < \frac{1}{2}$ for some literal $u$, then $u$ is a negative literal $\bar{x}$, and $p(x) = \sqrt[16]{\frac{\text{occ}_F(x)}{16e(F)}} > \frac{1}{2}$.

Second, if $p(u) = \frac{1}{2}$, then both $\sqrt[16]{\frac{\text{occ}_F(u)}{16e(F)}} \leq \frac{1}{2}$ and $\sqrt[16]{\frac{\text{occ}_F(\bar{u})}{16e(F)}} \leq \frac{1}{2}$ hold. We distinguish two types of clauses: Bad clauses, which contain at least one literal $u$ with $p(u) < \frac{1}{2}$, and good clauses, which contain only literals $u$ with $p(u) \geq \frac{1}{2}$. Let $B \subseteq F$ denote the set of bad clauses and $G \subseteq F$ the set of good clauses.

Lemma 12. $\sum_{C \in B} \text{Pr}[\alpha \nmid C] \leq \frac{1}{8}$.

Proof. For each clause $C \in B$, let $u_C$ be the literal in $C$ minimizing $p(u)$, breaking ties arbitrarily. This means $\text{Pr}[\alpha \nmid C] \leq p(u_C)^k$. Since $C$ is a bad clause, $p(u_C) < \frac{1}{2}$, $u_C$ is a negative literal $\bar{x}_C$, and $p(x_C) = \sqrt[16]{\frac{\text{occ}_F(x_C)}{16e(F)}}$. Thus

$$\sum_{C \in B} \text{Pr}[\alpha \nmid C] \leq \sum_{C \in B} p(x_C)^k = \sum_{C \in B} \frac{\text{occ}_F(x_C)}{16e(F)}. \quad (3)$$

Since clause $C$ contains $\bar{x}_C$, it conflicts with all $\text{occ}_F(x_C)$ clauses containing $x_C$, thus $\sum_{C \in B} \text{occ}_F(x_C) \leq 2e(F)$. The factor 2 arises since we count each conflict possibly twice, once from each side. Combining this with (3) proves the lemma.

We cannot directly apply Lemma 4 to $F$. Therefore we apply the below sparsification process to $F$.

Lemma 13. Let $F'$ be the result of the sparsification process. If $F'$ does not contain the empty clause, then $F'$ is satisfiable.

Proof. We will show that (2) applies to $F'$. Fix a clause $C \in F'$. After the sparsification process, every literal $u$ fulfills $\sum_{D \in G', u \in D} \text{Pr}[\alpha \nmid D] \leq \frac{1}{8k}$. Therefore, the terms $\text{Pr}[\alpha \nmid D]$, for all good clauses $D$ conflicting with $C$, sum up to at most $\frac{1}{8}$. By Lemma 12, the terms $\text{Pr}[\alpha \nmid D]$ for all bad clauses $D$ also sum up to at most $\frac{1}{8}$. Hence (2) holds, and by Lemma 4, $F'$ is satisfiable, and clearly $F$ as well.
Algorithm: Sparsification Process
Let $G' = \{D \in F \mid p(u) \geq \frac{1}{2}, \forall u \in D\}$ be the set of good clauses in $F$.

while $\exists$ a literal $u : \sum_{D \in G' : u \in D} \Pr[\alpha \not\models D] > \frac{1}{8k}$ do

Let $C$ be some clause maximizing $\Pr[\alpha \not\models D]$ among all clauses $D \in G' : u \in D$.
$C' := C \setminus \{u\}$

$G' := (G' \setminus \{C\}) \cup \{C'\}$

end

return $F' := G' \cup B$

Contrary, if $F$ is unsatisfiable, the sparsification process produces the empty clause. We will show that $e(F)$ is large. There is some $C \in G$ all whose literals are being deleted during the sparsification process. Write $C = \{u_1, u_2, \ldots, u_k\}$, and order the $u_i$ such that $\text{occ}_F(u_1) \leq \text{occ}_F(u_2) \leq \cdots \leq \text{occ}_F(u_k)$. One checks that this implies that $p(u_1) \leq p(u_2) \leq \cdots \leq p(u_k)$. Fix any $\ell \in \{1, \ldots, k\}$ and let $u_j$ be the first literal among $u_1, \ldots, u_\ell$ that is deleted from $C$. Let $C'$ denote what is left of $C$ just before that deletion, and consider the set $G'$ at this point of time. Then $\{u_1, \ldots, u_\ell\} \subseteq C' \in G'$. By the definition of the process,

$$\frac{1}{8k} < \sum_{D \in G' : u_j \in D} \Pr[\alpha \not\models D] \leq \sum_{D \in G' : u_j \in D} \Pr[\alpha \not\models C'] \leq \text{occ}_F(u_j) \Pr[\alpha \not\models C'] \leq \text{occ}_F(u_\ell) \prod_{i=1}^{\ell} (1 - p(u_i)).$$

Since $p(u) \geq \frac{1}{128 \text{occ}_F(F) \text{ke}(F)}$ for all literals $u$ in a good clause, it follows that $\frac{1}{128 \text{ke}(F)} \leq p(u_\ell)^k \prod_{i=1}^{\ell} (1 - p(u_i))$, for every $1 \leq \ell \leq k$.

Let $(q_1, \ldots, q_k) \in [\frac{1}{2}, 1]^k$ be any sequence satisfying the $k$ inequalities $\frac{1}{128 \text{ke}(F)} \leq q_\ell \prod_{i=1}^{\ell} (1 - q_i)$ for all $1 \leq \ell \leq k$, for example, the $p(u_i)$ are such a sequence. We want to make the $q_\ell$ as small as possible: If (i) $q_\ell > \frac{1}{2}$ and (ii) $\frac{1}{128 \text{ke}(F)} < q_\ell^k \prod_{i=1}^{\ell} (1 - q_i)$, we can decrease $q_\ell$ until one of (i) and (ii) becomes an equality. The other $k - 1$ inequalities stay satisfied. In the end we get a sequence $q_1, \ldots, q_k$ satisfying $\frac{1}{128 \text{ke}(F)} = q_\ell^k \prod_{i=1}^{\ell} (1 - q_i)$ whenever $q_\ell > \frac{1}{2}$. This sequence is non-decreasing: If $q_\ell > q_{\ell+1}$, then $q_\ell > \frac{1}{2}$, and $\frac{1}{128 \text{ke}(F)} \leq q_{\ell+1}^k \prod_{i=1}^{\ell+1} (1 - q_i) < q_{\ell+1}^k \prod_{i=1}^{\ell} (1 - q_i)$, and $q_\ell = q_\ell + \sqrt{1 - q_{\ell+1}}$. We define

$$f_k(t) := t \sqrt{1 - t},$$

thus $q_j = f_k(q_{j+1})$. By $f_k^{(j)}(t)$ we denote $f_k(f_k(\ldots (f_k(t)) \ldots))$, the $j$-fold iterated application of $f_k(t)$, with $f_k^{(0)}(t) = t$. We obtain $q_j = f_k^{(k-j)}(q_k) > \frac{1}{2}$ for $\ell^* \leq j \leq k$. By Part (v) of Proposition 15, $f_k^{(k-1)}(q_k) \leq \frac{1}{2}$, thus $\ell^* \geq 2$. Therefore $q_1 = \cdots = q_{\ell^*-1} = \frac{1}{2}$, and
We obtain $\frac{1}{128ke(F)} \leq q_{\ell^* - 1}^{k} \prod_{i=1}^{\ell^* - 1} (1 - q_i) = 2^{-k - \ell^* + 1}$.

We obtain $e(F) \geq \frac{2^k + \ell^* - 1}{128k}$. How large is $\ell^*$? Define $S_k := \min\{\ell \in \mathbb{N}_0 \mid f_k^{(\ell)}(t) \leq \frac{1}{2} \forall t \in [0,1]\}$. By Part (v) of Proposition 15 (see appendix), $S_k$ is finite. Since $f_k^{(k+\ell^*)}(q_1) = q_{\ell^*} > \frac{1}{2}$, we conclude that $k - \ell^* \leq S_k - 1$, thus $e(F) \geq \frac{2^{2k-S_k}}{128k}$.

**Lemma 14.** The sequence $\frac{S_k}{k}$ converges to $\lim_{k \to \infty} \frac{S_k}{k} = -\int_{\frac{1}{2}}^{1} \frac{1}{x \ln(1-x)} \, dx < 0.572$.

The proof of this lemma is technical and not related to satisfiability. We prove it in the appendix. We conclude that $e(F) \geq \frac{\sqrt{2(2-0.572)^k}}{128k} \in \Omega(2.69^k)$.

### 4 Conclusion

We want to give some hindsight why a sparsification procedure is necessary in both lower bound proofs in this paper. The probability distribution we define is not a uniform one, but biased towards setting $x$ to `true` if $\text{occ}_F(x) \gg \text{occ}_F(\bar{x})$. Let $C$ be a clause containing $\bar{x}$. It conflicts with all clauses containing $\bar{x}$. It could happen that in all those clauses, $x$ is the only literal with $p(x) > \frac{1}{2}$. In this case, each such clause is satisfied with probability not much smaller than $2^{-k}$, and the sum (2) is greater than $\frac{1}{2}$. By removing $x$ from these clauses, we reduce the number of clauses conflicting with $C$, making the sum (2) much smaller. However, for other clauses $C'$, this sum might increase by removing $x$. We think that one will not be able to prove a tight lower bound using just a smarter sparsification process. We state some open problems and questions.

**Question:** Does $\lim_{k \to \infty} \sqrt[4]{\text{gc}(k)}$ exist?

If it does, it lies between 2.69 and 3.51. One way to prove existence would be to define “product” taking a $k$-CNF formula $F$ and an $\ell$-CNF formula $G$ to a $(k + \ell)$-CNF formula $F \circ G$ that is unsatisfiable if $F$ and $G$ are, and $e(F \circ G) = e(F)e(G)$. With 2 and 4 ruled out, there seems to be no obvious guess for the value of the limit. What about $\sqrt[4]{8} \approx 2.828$, the geometric mean of 2 and 4?

**Question:** Is there an $a > 2$ such that every unsatisfiable $k$-CNF formula contains a variable $x$ with $\text{occ}_F(x) \cdot \text{occ}_F(\bar{x}) \geq a^k$?

Where do our methods fail to prove this? The part in the proof of the lower bound of Theorem 1 that fails is Lemma 12. On the other hand, Lemma 12 proves more than we need for Theorem 1: It proves that $\Pr[\alpha \models D]$, summed up over all bad clauses gives at most $\frac{1}{5}$. We only need that the bad clauses conflicting with a specific clause sum up to at most $\frac{1}{5}$. Still, we do not see how to apply or extend our methods to prove that such an $a > 2$ exists.

We discussed lower and upper bounds on the minimum of several parameters of unsatisfiable $k$-CNF formulas. The following table lists them where bounds labeled with an asterisk are from this paper and unlabeled bounds are not attributed to any specific paper.
<table>
<thead>
<tr>
<th>parameter</th>
<th>notation</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>occurrences of a literal</td>
<td>$\text{occ}(x)$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>occurrences of a variable</td>
<td>$f(k)$</td>
<td>$\frac{2^k}{e^k}$ [7]</td>
<td>$\frac{2^{k+3}}{k}$ [5]</td>
</tr>
<tr>
<td>local conflict number</td>
<td>$\text{lc}(k)$</td>
<td>$\frac{2^k}{e^k}$ [7]</td>
<td>$2^k - 1$</td>
</tr>
<tr>
<td>conflicts caused by a variable</td>
<td>$\text{occ}(x)\text{occ}(\bar{x})$</td>
<td>$\frac{2^k}{e^k}$ [7]</td>
<td>$O(3.01^k)^*$</td>
</tr>
<tr>
<td>global conflict number</td>
<td>$\text{gc}(k)$</td>
<td>$\Omega(2.69^k)^*$</td>
<td>$O(3.51^k)^*$</td>
</tr>
</tbody>
</table>

References


Proof of Lemma 14

Proposition 15. Let $k \in \mathbb{N}$ and $f_k : [0,1] \to [0,1]$ with $f_k(t) = t \sqrt[2k]{1-t}$. For $t \in [0,1]$, the following statements hold.

(i) $f_k(t)$ attains its unique maximum at $t = t^*_k := \frac{k}{k+1}$.

(ii) $f_k(t) \leq t$, and $f_k(t) = t$ if and only if $t = 0$.

(iii) For $\ell \geq 1$, $f_k^{(\ell)}(t) \leq f_k^{(\ell)} \left( \frac{k}{k+1} \right)$.

(iv) For $\ell \geq 0$ and $t \in [0,1]$, $(1-t)^{\ell/k} t \leq f_k^{(\ell)}(t) \leq (1 - f_k^{(\ell)}(t))^{\ell/k} t$.

(v) For $k \geq 2$ and any $t \in [0,1]$, $f_k^{(k-1)}(t) \leq \frac{1}{2}$.

Proof. (i) follows from elementary calculus. (ii) holds since $\sqrt[2k]{1-t}$ is less than 1 for all $t > 0$. For $\ell = 1$, (iii) follows from (i), and for greater $\ell$, it follows from (ii) and induction on $\ell$. (iv) holds because each of the $\ell$ applications of $f_k$ multiplies its argument with a factor that is at least $\sqrt[2k]{1-t}$ and at most $1 - f_k^{(\ell)}(t)$. Suppose (v) does not hold. Then by (iii) we get $f_k^{(k-1)} \left( \frac{k}{k+1} \right) \geq f_k^{(k-1)}(t) > \frac{1}{2}$, and by (iv), we have

$$\frac{1}{2} < f_k^{(k-1)} \left( \frac{k}{k+1} \right) \leq \left( \frac{1}{2} \right)^{k-1} \frac{k}{k+1}.$$ 

An elementary calculation shows that this does not hold for any $k \geq 1$. $\square$

To prove Lemma 14, we compute $\lim_{k \to \infty} \frac{S_k}{k}$ (and show that the limit exists). Recall the definition

$$S_k = \min \{ \ell \in \mathbb{N}_0 \mid f_k^{(\ell)}(t) \leq \frac{1}{2} \forall t \in [0,1] \},$$

where $f_k(t) = t \sqrt[2k]{1-t}$. By Part (iii) of Proposition 15, $S_k = \min \{ \ell \mid f_k^{(\ell)}(t^*_k) \leq \frac{1}{2} \}$, for $t^*_k := \frac{k}{k+1}$. We generalize the definition of $S_k$ by defining for $t \in (0,1]$,

$$S_k(t) := \min \{ \ell \mid f_k^{(\ell)}(t^*_k) \leq t \}.$$ 

Further, we set $s_k(t) := \frac{S_k(t)}{k}$. Let $0 < t_2 < t_1 < t^*_k$. We want to estimate $s_k(t_2) - s_k(t_1)$. This should be small if $|t_1 - t_2|$ is small. For brevity, we write $a := S_k(t_1)$, $b := S_k(t_2)$. Clearly $a \leq b$. We calculate

$$t_2 \geq f_k^{(b)}(t^*_k) = f_k^{(b-a+1)}(f_k^{(a-1)}(t^*_k)) \geq f_k^{(b-a+1)}(t_1) \geq (1-t_1)^{((b-a+1)/k)t_1},$$

$$t_2 < f_k^{(b-1)}(t^*_k) = f_k^{(b-a-1)}(f_k^{(a)}(t^*_k)) \leq f_k^{(b-a-1)}(t_1) \leq (1-t_2)^{((b-a-1)/k)t_1}.$$ 

Where we used part (iv) of Proposition 15. In fact, these inequalities also hold if $t_1 \geq t^*_k$, when $a = 0$:

$$t_2 \geq f_k^{(b)}(t^*_k) \geq (1-t^*_k)^{b/k} t_1 \geq (1-t_1)^{(b+1)/k} t_1,$$

$$t_2 < f_k^{(b-1)}(t^*_k) = (1-t_2)^{(b-1)/k} t_1.$$
One checks that the inequalities even hold if $t^* \leq t_2 < t_1 \leq 1$. Note that $\frac{b-a}{k} = s_k(t_2) - s_k(t_1)$. Solving for $\frac{b-a}{k}$, the above inequalities yield

$$\frac{\log t_2 - \log t_1}{\log(1 - t_1)} - \frac{1}{k} \leq s_k(t_2) - s_k(t_1) \leq \frac{\log t_2 - \log t_1}{\log(1 - t_2)} + \frac{1}{k},$$ (4)

for all $0 < t_2 < t_1 < 1$. The right inequality also holds for $0 < t_2 < t_1 \leq 1$. Multiplying with $-1$, we see that it also holds if $t_2 > t_1$. If $t_2 = t_1$, it is trivially true. Hence this inequality is true for all $t_1, t_2 \in (0, 1)$.

Suppose $s(t) = \lim_{k \to \infty} s_k(t)$ exists, for every fixed $t$. Inequality (4) also holds in the limit. Writing $t_1 = t$ and $t_2 = t + h$ and dividing (4) by $h$ gives

$$\frac{\log(t + h) - \log t}{h \log(1 - t)} \leq \frac{s(t + h) - s(t)}{h} \leq \frac{\log(t + h) - \log t}{h \log(1 - t - h)}.$$

Letting $h$ go to 0, we obtain $s'(t) = \frac{1}{t \log(1 - t)}$, thus $s(t) = s(1) - \int_t^1 \frac{1}{x \log(1 - x)} dx$. Observing that $\frac{s_k}{k} = s_k(\frac{1}{2})$ and $s_k(1) = 0$ for all $k$ proves the Lemma.

The above argument shows that if $s_k(t)$ converges pointwise, then it converges to a continuous function $s(t)$ on $(0, 1)$. We have to show that $\lim_{k \to \infty} s_k(t)$ does in fact exist. First plug in $t_1 = 1$ into the right inequality of (4) to observe that for each fixed $t_2$, the sequence $(s_k(t_2))_{k \in \mathbb{N}}$ is bounded from above. Clearly it is bounded from below by 0. Hence there exist $\overline{s}(t) := \limsup s_k(t)$ and similarly $\underline{s}(t) := \liminf s_k(t)$. We write shorthand $L(t_1, t_2) := \frac{\log t_2 - \log t_1}{\log(1 - t_1)}$ and $U(t_1, t_2) := \frac{\log t_2 - \log t_1}{\log(1 - t_2)}$. Now (4) reads as

$L(t_1, t_2) - \frac{1}{k} \leq s_k(t_2) - s_k(t_1) \leq U(t_1, t_2) + \frac{1}{k}$. We claim that

$$L(t_1, t_2) \leq \overline{s}(t_2) - \overline{s}(t_1) \leq U(t_1, t_2),$$ (5)

$$L(t_1, t_2) \leq \underline{s}(t_2) - \underline{s}(t_1) \leq U(t_1, t_2).$$ (6)

For sequences $(a_k)_{k \in \mathbb{N}}, (b_k)_{k \in \mathbb{N}}$, $\limsup a_k - \limsup b_k = \limsup(a_k - b_k)$ does not hold in general, hence the claim is now completely trivial. We will proof that $\overline{s}(t_2) - \overline{s}(t_1) \leq U(t_1, t_2)$. This will prove one claimed inequality. The other three inequalities can be proven similarly. Fix some small $\epsilon > 0$. For all sufficiently large $k$, $\frac{1}{k} \leq \epsilon$. We have $s_k(t_2) \geq \overline{s}(t_2) - \epsilon$ for infinitely many $k$, thus $s_k(t_1) \geq s_k(t_2) - U(t_1, t_2) - \frac{1}{k} \geq \overline{s}(t_2) - U(t_1, t_2) - 2\epsilon$ for infinitely many $k$. Therefore $\overline{s}(t_1) \geq s(t_2) - U(t_1, t_2) - 2\epsilon$. By making $\epsilon$ arbitrarily small, the claimed inequality follows.

We can now apply our non-rigorous argument from above, this time rigorously. Write $t = t_1, t_2 = t + h$, and divide (5) and (6) by $h$, send $h$ to 0, and we obtain $\overline{s}'(t) = \underline{s}'(t) = \frac{1}{t \log(1 - t)}$. Since $\overline{s}(1) = \underline{s}(1) = 0$, we obtain

$$\overline{s}(t) = \underline{s}(t) = \int_t^1 \frac{-1}{x \log(1 - x)} dx.$$