Unsatisfiable CNF Formulas Contain Many Conflicts

Dominik Scheder⋆ ⋆⋆ ⋆⋆⋆

Aarhus University

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 CNF formula with clauses of size exactly k (a *k-CNF formula*), is satisfiable unless some clause conflicts with at least $\frac{2^k}{e}$ 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 $\Omega(2.69^k)$ conflicts and there exist unsatisfiable k-CNF formulas with $O(3.51^k)$ conflicts.

1 Introduction

A boolean formula in conjunctive normal form (short a *CNF formula*) is a conjunction (AND) of *clauses*, which are disjunctions of literals. A *literal* is either a boolean variable x or its negation $\bar{x}$. SAT, the problem of deciding whether a CNF formula is satisfiable is a central problem in theoretical computer science, and was one of the first problems to be proven NP-complete. How can a CNF formula become unsatisfiable? Roughly speaking, there are two possibilities: Either some clause itself is impossible to satisfy – this is only the case for the empty clause. Or, each clause is individually satisfiable, but there are conflicts between the clauses, making it impossible to satisfy all of them simultaneously. When we consider k-CNF formulas, where each clause consists of exactly k literals (we require that literals in a clause do not repeat), then each clause is extremely easy to satisfy: Of the 2^k possible truth assignments to its variables, all but one satisfy it. If a k-CNF formula is unsatisfiable, we expect it to have 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 $(x \vee y)$ and $(\bar{x} \vee u)$ conflict. Similarly, $(x \vee y)$ and

⋆ The author acknowledges support from the Danish National Research Foundation and The National Science Foundation of China (under the grant 61061130540) for the Sino-Danish Center for the Theory of Interactive Computation, within which this work was performed.

⋆⋆ Research was supported by the SNF Grant 200021-118001/1.

⋆⋆⋆ The author acknowledges support from the Simons Institute for the Theory of Computing, UC Berkeley.

$(\bar{x} \vee \bar{y})$ do. Suppose F is a CNF formula without the empty clause, and without any conflicts. Then clearly F 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 $CG(F)$. Our above observation now reads as follows: If F does not contain the empty clause, and $e(F) = 0$, then F is satisfiable. In fact, any k-CNF formula is satisfiable unless $\Delta(F)$ and $e(F)$ are large. How large? A quantitative result follows from the Lopsided Lovász Local Lemma [1–3]: A k-CNF formula F is satisfiable unless some clause conflicts with $\frac{2^k}{e}$ or more clauses, i.e., unless $\Delta(F) \geq \frac{2^k}{e}$. Up to a constant factor, this is tight: Consider the formula containing all 2^k clauses over the variables $x_1, \ldots, x_k$. We call this a *complete k-CNF formula* and denote it by $\mathcal{K}_k$. It is unsatisfiable, and $\Delta(\mathcal{K}_k) = 2^k - 1$.

As its name suggests, the Lopsided Lovász Local Lemma implies a *local* result: A k-CNF formula F is satisfiable, unless *somewhere* in F there are many conflicts. We want to obtain a *global* result: F is satisfiable unless the total number of conflicts is *very* large. We define two functions:

$$lc(k) := \max\{d \in \mathbb{N}_0 \mid \text{ every } k\text{-CNF formula } F \text{ with } \Delta(F) \leq d \text{ is satisfiable}\} ,$$
$$gc(k) := \max\{d \in \mathbb{N}_0 \mid \text{ every } k\text{-CNF formula } F \text{ with } e(F) \leq d \text{ is satisfiable}\} .$$

The abbreviations lc and gc stand for *local conflicts* and *global conflicts*, respectively. From the above discussion, $\frac{2^k}{e} - 1 \leq lc(k) \leq 2^k - 2$, hence we know $lc(k)$ up to a constant factor. In contrast, it does not seem to be easy to prove nontrivial upper and lower bounds on $gc(k)$. Let us see what we get: Surely, $gc(k) \geq lc(k) \geq \frac{2^k}{e} - 1$. For an upper bound, $gc(k) \leq e(\mathcal{K}_k) - 1 = \binom{2^k}{2} - 1$. Ignoring constant factors, $gc(k)$ lies somewhere between 2^k and 4^k. This leaves much space for improvement. In [4], Zumstein and I proved that $gc(k) \in \Omega(2.27^k)$ and $gc(k) \leq \frac{4^k}{\log^3 k} k$. In this paper, we significantly improve upon these bounds. Somehow surprisingly, $gc(k)$ is exponentially smaller than 4^k.

Theorem 1. *Any unsatisfiable k-CNF formula contains $\Omega\left(2.69^k\right)$ conflicts. On the other hand, there is an unsatisfiable k-CNF formula with $O\left(3.51^k\right)$ conflicts.*

We obtain the lower bound by a more sophisticated application of the idea used in [4]. The upper bound follows from a construction that is partially probabilistic, and inspired in parts by Erdős' construction in [5] of small k-uniform hypergraphs that are not 2-colorable.

1.1 Related Work

Let F be a CNF formula and u be a literal. We write $\mathrm{occ}_F(u) := |\{C \in F \mid u \in C\}|$. For a variable x, we write $d_F(x) = \mathrm{occ}_F(x) + \mathrm{occ}_F(\bar{x})$. So $d_F(x)$, the *degree* of x, counts the number of clauses containing the variable x, irrespective of its polarity. We write $d(F) = \max_x d_F(x)$. It is easy to see that for a k-CNF formula, $\Delta(F) \leq k(d(F) - 1)$. We define

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

The function $f(k)$ has been subject of some research. By an application of Hall's Theorem, Tovey [6] 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 $f(k) \geq \frac{2^k}{ek}$: In our terminology, they showed that $lc(k) \geq \frac{2^k}{e} - 1$ and then used the fact that $\Delta(F) \leq k(d(F) - 1)$. As for an upper bound, in [7] the authors show that $f(k) \leq 2^{k-1} - 2^{k-4} - 1$. This was improved by Savický and Sgall [8] to $f(k) \in O(k^{-0.26} 2^k)$, by Hoory and Szeider [9] to $f(k) \in O\left(\frac{\log(k) 2^k}{k}\right)$, and only recently, by Gebauer [10] to $f(k) \leq \frac{2^{k+2}}{k} - 1$ clauses, closing the gap between lower and upper bound on $f(k)$ up to a constant factor. Finally, Gebauer, Szabó, Tardos [11] proved that $f(k) = (1 \pm o(1)) 2^{k+1}/ek$, which even determines the constant factor.

1.2 Conflicts Generated by a Single Variable

Let F be a CNF formula and x a variable. Every clause containing x conflicts with every clause containing $\bar{x}$, thus $e(F) \geq \mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x})$. In fact,

$$e(F) \geq \frac{1}{k} \sum_x \mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x}) \tag{1}$$

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. By [7], every unsatisfiable k-CNF formula F contains a variable x with $d_F(x) \geq \frac{2^k}{ek}$. If this variable is *balanced*, i.e. $\mathrm{occ}_F(x)$ and $\mathrm{occ}_F(\bar{x})$ are both at least $\frac{2^k}{\mathrm{poly}(k)}$, then $e(F) \geq \frac{4^k}{\mathrm{poly}(k)}$. Indeed, in the formulas constructed in [10], all variables are balanced. The same holds for the complete k-CNF formula $\mathcal{K}_k$. Thus, it might be the case that in every unsatisfiable k-CNF formula, there is a single variable that already generates many conflicts:

Conjecture 1. There exists a number $a > 2$ such that every unsatisfiable k-CNF formula F contains a variable x such that $\mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x}) \geq \Omega\left(a^k\right)$.

We do not know whether this conjecture is true. However, we will give non-trivial *upper* bounds on $\mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x})$:

Theorem 2. *For all sufficiently large k, there is an unsatisfiable k-CNF formula with* $\mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x}) \leq 3.01^k$ *for all variables x.*

2 Notation and Tools

Throughout the paper, we regard formulas as sets of clauses and clauses as sets of literals. This is purely to simplify notation. For a truth assignment α and a clause C, we will write $\alpha \models C$ if α satisfies C. Similarly $\alpha \not\models C$ if it does not. If α satisfies a formula F, we write $\alpha \models F$.

We will state a version of the Lopsided Lovász Local Lemma formulated in terms of satisfiability. For a derivation of this version see [12].

Lemma 1 (SAT Version of the Lopsided Lovász Local Lemma). *Let F be a CNF formula not containing the empty clause. Sample a truth assignment α by independently setting each variable x to* `true` *with $p(x) \in [0,1]$. If for any clause $C \in F$, it holds that*

$$\sum_{D \in F\colon\, C \text{ and } D \text{ conflict}} \Pr[\alpha \not\models D] \leq \frac{1}{4} \tag{2}$$

then F is satisfiable.

In our proofs, it will be difficult to apply Lemma 1 to a formula F which we want to prove satisfiable. Instead, we apply it to a formula F' we obtain from F in the following way:

Definition 1. *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, $(x \vee y) \wedge (\bar{y} \vee z)$ is a truncation of $(x \vee y \vee \bar{z}) \wedge (\bar{x} \vee \bar{y} \vee z)$. A truncation of a k-CNF formula is not a k-CNF formula anymore. It is easy to see that any truth assignment satisfying a truncation F' of F also satisfies F. In our proofs, we will often find it easier to apply Lemma 1 to a special truncation of F than to F itself. We need a technical lemma on the binomial coefficient.

Lemma 2. *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.

$$\binom{a}{b} = \frac{a(a-1)\cdots(a-b+1)}{b!} = \frac{a^b}{b!} \prod_{j=0}^{b-1} \frac{a-j}{a} > \frac{a^b}{b!} e^{-\frac{2}{a}\sum_{j=0}^{b-1} j} > \frac{a^b}{b!} e^{-b^2/a} ,$$

where we used the fact that $1 - x > e^{-2x}$ for $0 \leq x \leq 0.75$. □

3 Upper Bounds – Probabilistic Constructions of Unsatisfiable Formulas

As we have argued in Section 1.2, 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. $\mathrm{occ}_F(x)$ is exponentially larger than $\mathrm{occ}_F(\bar{x})$. The central idea is that we do not construct an unsatisfiable k-CNF formula, but allow certain clauses to be smaller. In a second step, we expand these clauses to size k.

Definition 2. *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 $\mathcal{K}_{k-k'}$ over $k-k'$ new variables $y_1^C, \ldots, y_{k-k'}^C$. We replace C by $C \vee \mathcal{K}_{k-k'}$. Using distributivity, we expand it into a k-CNF formula G called the* k-CNFification of F.

For example, the 3-CNFification of $(x \vee y) \wedge (\bar{x} \vee y \vee z)$ is $(x \vee y \vee y_1) \wedge (x \vee y \vee \bar{y}_1) \wedge (\bar{x} \vee y \vee z)$. It is easy to see that a truth assignment satisfies F if and only if it satisfies its k-CNFification G.

Definition 3. *Let $\ell, k \in \mathbb{N}_0$. An (ℓ, k)-CNF formula is a formula consisting of ℓ-clauses containing only positive literals, and k-clauses containing only negative literals.*

If F is an (ℓ, k)-CNF formula, we write $F = F^+ \wedge F^-$, where F^+ consists of purely positive ℓ-clauses and F^- of purely negative k-clauses.

Proposition 1. *Let $\ell \leq k$, and let $F = F^+ \wedge F^-$ be an (ℓ, k)-CNF formula. Let G be the k-CNFification of F. Then*

(i) $e(G) \leq 4^{k-\ell}|F^+| + 2^{k-\ell}|F^+| \cdot |F^-|$,
(ii) $\mathrm{occ}_G(x) \cdot \mathrm{occ}_G(\bar{x}) \leq \max\{4^{k-\ell}, 2^{k-\ell}|F^+| \cdot |F^-|\}$ *for every variable x.*

Proof. Every edge in $CG(F)$ runs between a positive ℓ-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. Replacing C by $2^{k-\ell}$ copies introduces less than $4^{k-\ell}$ edges. This proves (i). To prove (ii), there are two cases. First, if x appears in F, then $\mathrm{occ}_G(\bar{x}) = \mathrm{occ}_F(\bar{x})$ and $\mathrm{occ}_G(x) = \mathrm{occ}_F(x)2^{k-\ell}$, thus $\mathrm{occ}_G(x)\mathrm{occ}_G(\bar{x}) \leq 2^{k-\ell}|F^+| \cdot |F^-|$. Second, if x does not appear in F, it has been introduced in the k-CNFification. Then $\mathrm{occ}_G(x) = \mathrm{occ}_G(\bar{x}) = 2^{k-\ell-1}$, and $\mathrm{occ}_G(x) \cdot \mathrm{occ}_G(\bar{x}) \leq 4^{k-\ell}$. □

We will explore for which values of $|F^+|$ and $|F^-|$ there are unsatisfiable (ℓ, k)-CNF formulas. Then we use Proposition 1 to derive the upper bounds of Theorem 1 and Theorem 2.

Lemma 3. *(i) For any $\rho \in (0, 1)$, there is a constant c such that for all k and $\ell \leq k$, there exists an unsatisfiable (ℓ, k)-CNF formula $F = F^+ \wedge F^-$ with $|F^-| \leq ck^2\rho^{-k}$ and $|F^+| \leq ck^2(1-\rho)^{-\ell}$.*

(ii) Let $F = F^+ \wedge F^-$ be an (ℓ, 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. We begin with (ii), which is easier. Sample a truth assignment α by setting each variable independently to `true` with probability ρ. For a negative k-clause C, it holds that $\Pr[\alpha \not\models C] = \rho^k$. Similarly, for a positive ℓ-clause D, $\Pr[\alpha \not\models D] = (1-\rho)^\ell$. Hence the expected number of clauses in F that are unsatisfied by α is $\rho^k|F^-| + (1-\rho)^\ell|F^+| < \frac{1}{2} + \frac{1}{2} = 1$. Therefore, with positive probability α satisfies F.

For (i), we choose a set $V = \{x_1, \ldots, x_n\}$ of $n = k^2$ variables. Let c be a constant, to be determined later. We form F^- by sampling, with replacement, $ck^2\rho^{-k}$ negative k-clauses from all $\binom{n}{k}$ possible. Similarly, we form F^+ by sampling $ck^2(1-\rho)^{-\ell}$ positive ℓ-clauses. We claim that for a suitable choice of c this formula is unsatisfiable with high probability. Let α be any truth assignment. There are two cases. First, suppose α sets at least ρn variables to `true`. For a random negative clause C,

$$\Pr[\alpha \not\models C] \geq \frac{\binom{\rho n}{k}}{\binom{n}{k}} \geq \frac{\frac{(\rho n)^k}{k!} \cdot e^{-k^2/(\rho n)}}{\frac{n^k}{k!}} = \rho^k e^{-1/\rho} = c' \rho^k$$

By independence, $\Pr[\alpha \models F^-] \leq (1 - c'\rho^k)^{ck^2\rho^{-k}} < e^{-cc'k^2}$. Second, suppose α sets at most ρn variables to `true`. By a similar argument, $\Pr[\alpha \models F^+] \leq (1 - c''(1-\rho)^\ell)^{ck^2(1-\rho)^{-\ell}} < e^{-cc''k^2}$. For suitable c, we obtain $\Pr[\alpha \models F] < e^{-k^2} = e^{-n}$ for any α. The expected number of satisfying assignments of F is thus less than $2^n e^{-n} < 1$. With high probability F is unsatisfiable. □

It should be pointed out that for $k = \ell$, an (ℓ, 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 [13] raised the question what this number is, and proved a 2^{k-1} lower bound (this is easy, simply choose a random 2-coloring). One year later, he [5] gave a probabilistic construction of a non-2-colorable k-uniform hypergraph using $ck^2 2^k$ hyperedges. For $\ell = k$ and $\rho = \frac{1}{2}$, the above proof is basically the same as Erdős' proof.

Proof (Proof of Theorem 2). Combining Lemma 3 and Proposition 1, we conclude that for any $\rho \in (0,1)$ and $0 \leq \ell \leq k$, there is an unsatisfiable k-CNF formula F with

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

for every variable x. The constant c depends on ρ, but not on k or ℓ. The term $\rho^{-k}(1-\rho)^{-\ell}$ is minimized for $\rho = \frac{k}{k+\ell}$. Choosing $\ell = \lceil 0.2055k \rceil$, we get $\rho \approx 0.83$ and $\mathrm{occ}_F(x) \cdot \mathrm{occ}_F(\bar{x}) \in O(3.01^k)$. □

Proof (Proof of the upper bound of Theorem 1). As in the previous proof, Proposition 1 together with Lemma 3 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 $e(F) \in O(3.51^k)$. □

4 A Lower Bound on the Number of Global Conflicts

Proof (of the lower bound in Theorem 1). Let F be an unsatisfiable k-CNF formula and let $e(F)$ be the number of conflicts in F. We will show that $e(F) \in \Omega\left(2.69^k\right)$. In the proof, x denotes a variable and u a positive or negative literal. We assume $\mathrm{occ}_F(\bar{x}) \leq \mathrm{occ}_F(x)$ 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 $\mathrm{occ}_F(x)$ and $\mathrm{occ}_F(\bar{x})$ are both at least 1, if x occurs in F at all. For x, we define

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

We define a random truth assignment α by setting x to `true` with probability $p(x)$, independently for each variable. Since $\mathrm{occ}_F(u) \leq e(F)$, 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 $p(x) = \sqrt[k]{\frac{\mathrm{occ}_F(x)}{16e(F)}} > \frac{1}{2}$. Second, if $p(u) = \frac{1}{2}$, then both $\sqrt[k]{\frac{\mathrm{occ}_F(x)}{16e(F)}} \leq \frac{1}{2}$ and $\sqrt[k]{\frac{\mathrm{occ}_F(\bar{x})}{16e(F)}} \leq \frac{1}{2}$ 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 $p(u) \geq \frac{1}{2}$.

Lemma 4. *Let $\mathcal{B} \subseteq F$ denote the set of bad clauses. Then $\sum_{C \in \mathcal{B}} \Pr\left[\alpha \not\models C\right] \leq \frac{1}{8}$.*

Proof. For each clause $C \in \mathcal{B}$, let u_C be the literal in C minimizing $p(u)$, breaking ties arbitrarily. This means $\Pr[\alpha \not\models C] \leq p(\bar{u}_C)^k$. Since C is a bad clause, $p(u_C) < \frac{1}{2}$, u_C is a negative literal $\bar{x}_C$, and $p(x_C) = \sqrt[k]{\frac{\mathrm{occ}_F(x_C)}{16e(F)}}$. We can calculate

$$\sum_{C \in \mathcal{B}} \Pr[\alpha \not\models C] \leq \sum_{C \in \mathcal{B}} p(x_C)^k = \sum_{C \in \mathcal{B}} \frac{\mathrm{occ}_F(x_C)}{16e(F)} . \tag{3}$$

Since clause C contains $\bar{x}_C$, it conflicts with all $\mathrm{occ}_F(x_C)$ clauses containing x_C, thus $\sum_{C \in \mathcal{B}} \mathrm{occ}_F(x_C) \leq 2e(F)$. 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 1 to F. Therefore we apply the following sparsification process to F:

Algorithm: Sparsification Process

Let $\mathcal{G} = \{D \in F \mid p(u) \geq \frac{1}{2}, \forall u \in D\}$ be the set of good clauses in F.
$\mathcal{G}' := \mathcal{G}$
while $\exists$ a literal $u :\ \sum_{D:u\in D\in\mathcal{G}'} \Pr[\alpha \not\models D] > \frac{1}{8k}$ **do**
  Let C be some clause maximizing $\Pr[\alpha \not\models C]$ among all clauses in $\mathcal{G}'$ containing u.
  $C' := C \setminus \{u\}$
  $\mathcal{G}' := (\mathcal{G}' \setminus \{C\}) \cup \{C'\}$
end
return $F' := \mathcal{G}' \cup \mathcal{B}$

Lemma 5. *If F' does not contain the empty clause, then F is satisfiable.*

Proof. We will prove this using Lemma 1, the SAT version of the Lopsided Lovász Local Lemma. Fix a clause $C \in F'$. After the sparsification process, every literal u fulfills $\sum_{D:u\in D\in\mathcal{G}'} \Pr[\alpha \not\models D] \leq \frac{1}{8k}$. We combine this with Lemma 4 to show that the condition (2) of the Local Lemma holds:

$$\begin{aligned}
\sum_{D\in F':\ C \text{ and } D \text{ conflict}} \Pr[\alpha \not\models D] &= \sum_{D\in\mathcal{B}} \Pr[\alpha \not\models D] + \sum_{D\in\mathcal{G}':\ C \text{ and } D \text{ conflict}} \Pr[\alpha \not\models D] \\
&\leq \frac{1}{8} + \sum_{u\in C} \sum_{D\in\mathcal{G}':\bar{u}\in D} \\
&\leq \frac{1}{8} + k \cdot \frac{1}{8k} = \frac{1}{4}\ .
\end{aligned}$$

Hence (2) holds and by Lemma 1, F' is satisfiable, and clearly F as well. □

If F is unsatisfiable, the sparsification process produces the empty clause. We will show that in this case, $e(F)$ is large (at least $\Omega\left(2.69^k\right)$). If the sparsification process produces the empty clause, then there is some $C \in \mathcal{G}$ 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 $\mathrm{occ}_F(u_1) \leq \mathrm{occ}_F(u_2) \leq \cdots \leq \mathrm{occ}_F(u_k)$. Since C is a good clause, the definition of $p(x)$ 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 $\mathcal{G}'$ at this point of time. Then $\{u_1, \ldots, u_\ell\} \subseteq C' \in \mathcal{G}'$. By the definition of the process,

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

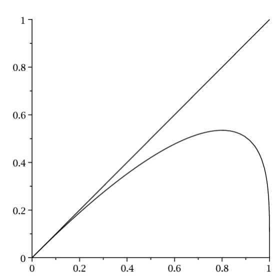
Since $p(u) \geq \sqrt[k]{\frac{\mathrm{occ}_F(u)}{16e(F)}}$ for all literals u in a good clause, it follows that $\frac{1}{128ke(F)} \leq p(u_\ell)^k \prod_{i=1}^{\ell}(1 - p(u_i))$, for every $1 \leq \ell \leq k$.

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

If all q_i are $\frac{1}{2}$, then the k^{th} inequality yields $128ke(F) \geq 4^k$, and we are done. Otherwise, there is some $\ell^* = \min\{i \mid q_i > \frac{1}{2}\}$. For $\ell^* \leq j < k$ both q_j and q_{j+1} are greater than $\frac{1}{2}$, thus $q_{j+1}^k \prod_{i=1}^{j+1}(1 - q_i) = q_j^k \prod_{i=1}^{j}(1 - q_i)$, and $q_j = q_{j+1} \sqrt[k]{1 - q_{j+1}}$. We define

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

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 $f_k^{(0)}(t) = t$. In this notation, $q_j = f_k^{(k-j)}(q_k) > \frac{1}{2}$ for $\ell^* \leq j \leq k$. The figure below shows the graph of $f_4(t)$.



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

We will prove this in the appendix. By Proposition 2, $f_k^{(k-1)}(q_k) \leq \frac{1}{2}$, thus $\ell^* \geq 2$. Therefore $q_1 = \cdots = q_{\ell^*-1} = \frac{1}{2}$, and the $(l^*-1)^{\text{st}}$ inequality reads as

$$\frac{1}{128ke(F)} \leq q_{\ell^*-1}^k \prod_{i=1}^{\ell^*-1} (1-q_i) = 2^{-k-\ell^*+1} \ .$$

We obtain $e(F) \geq \frac{2^{k+\ell^*-1}}{128k}$. How large is ℓ^*? 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 **??**, 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 6. *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 $e(F) \geq \frac{2^{(2-0.572)k}}{128k} \in \Omega\left(2.69^k\right)$. $\square$

5 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 $\text{occ}_F(x) \gg \text{occ}_F(\bar{x})$. The set of clauses conflicting with a specific clause C may contain many clauses containing some x with $\bar{x} \in C$. If x is the only literal in these clauses with $p(x) > \frac{1}{2}$, then each such clause is unsatisfied with probability not much smaller than 2^{-k}, and the sum (2) is greater than $\frac{1}{4}$ 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 want to state some open problems and questions.

Question: Does $\lim_{k\to\infty} \sqrt[k]{gc(k)}$ 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 ℓ-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.

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 4. On the other hand, Lemma 4 proves more than we need for Theorem 1: It proves that $\Pr[\alpha \models D]$, summed up over *all* bad clauses gives at most $\frac{1}{8}$. We only need that the bad clauses conflicting with a specific clause sum up to at most $\frac{1}{8}$. Still, we do not see how to apply or extend our methods to prove that such an $a > 2$ exists.

Acknowledgments

I am very thankful to Philipp Zumstein. We discussed the results of this paper extensively, and in particular the idea behind the proof of Lemma 6 is due to him.

References

1. Erdős, P., Spencer, J.: Lopsided Lovász Local Lemma and Latin transversals. Discrete Appl. Math. **30**(2-3) (1991) 151–154 ARIDAM III (New Brunswick, NJ, 1988).
2. Alon, N., Spencer, J.H.: The probabilistic method. Second edn. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York (2000) With an appendix on the life and work of Paul Erdős.
3. Lu, L., Székely, L.: Using Lovász Local Lemma in the space of random injections. Electron. J. Combin. **14**(1) (2007) Research Paper 63, 13 pp. (electronic)
4. Scheder, D., Zumstein, P.: How many conflicts does it need to be unsatisfiable? In: SAT. (2008) 246–256
5. Erdős, P.: On a combinatorial problem. II. Acta Math. Acad. Sci. Hungar **15** (1964) 445–447
6. Tovey, C.A.: A simplified NP-complete satisfiability problem. Discrete Appl. Math. **8**(1) (1984) 85–89
7. Kratochvíl, J., Savický, P., Tuza, Z.: One more occurrence of variables makes satisfiability jump from trivial to NP-complete. SIAM Journal of Computing **22**(1) (1993) 203–210
8. Savický, P., Sgall, J.: DNF tautologies with a limited number of occurrences of every variable. Theoret. Comput. Sci. **238**(1–2) (2000) 495–498
9. Hoory, S., Szeider, S.: A note on unsatisfiable k-CNF formulas with few occurences per variable. SIAM Journal on Discrete Mathematics **20**(2) (2006) 523–528
10. Gebauer, H.: Disproof of the neighborhood conjecture and its implications to sat (2008) submitted.
11. Gebauer, H., Szabó, T., Tardos, G.: The local lemma is tight for SAT. In Randall, D., ed.: SODA, SIAM (2011) 664–674
12. Scheder, D., Zumstein, P.: How many conflicts does it need to be unsatisfiable? In: Eleventh International Conference on Theory and Applications of Satisfiability Testing (SAT), Lecture Notes in Computer Science, Vol. 4996. (2008) 246–256
13. Erdős, P.: On a combinatorial problem. Nordisk Mat. Tidskr. **11** (1963) 5–10, 40