On the Length of an Unsatisfiable Conjunction

Alexandr V. Seliverstov

Abstract. We consider a lower bound on the length of a conjunction of some propositional formulae such that every unsatisfiable conjunction contains an unsatisfiable subformula. In particular, our method is applicable for 2-CNF, symmetric 3-CNF, and conjunctions of voting functions in three literals. The proof is algebraic. So, a large conjunction can be reduced in a non-deterministic way. This reduction improves some upper bounds on the computational complexity.

Introduction

Let us denote by $\bot$ and $\top$ two Boolean constants. For two integers $\alpha < \beta$, the set α-or-β-in-SAT consists of CNFs such that, for some $(\bot, \top)$-evaluation, every clause contains either exactly α or exactly β true literals.

For $k < \alpha < \beta$, the set α-or-β-in-SAT contains no k-CNF.

For $k < \beta$, a k-CNF φ belongs to 1-or-β-in-SAT iff φ belongs to 1-in-k-SAT. This set consists of k-CNF such that, for some $(\bot, \top)$-evaluation, every clause contains exactly one true literal.

A 2-CNF φ belongs to 1-or-2-in-SAT iff φ is satisfiable.

A 3-CNF φ belongs to 1-or-2-in-SAT iff φ belongs to NAE-3-SAT. This set consists of 3-CNF such that, for some $(\bot, \top)$-evaluation, every clause contains both true and false literals. A 3-CNF $\varphi(p_1, \ldots, p_n)$ belongs to NAE-3-SAT iff $\varphi(p_1, \ldots, p_n) \wedge \varphi(\neg p_1, \ldots, \neg p_n)$ is satisfiable.

It is well known that both problems whether a given 3-CNF belongs to NAE-3-SAT and whether it belongs to 1-in-3-SAT are NP-complete. The formula length serves as a natural parameter for estimating the runtime. Therefore, the possibility of replacing the original 3-CNF with a subformula is interesting. On complexity upper bounds refer to [1]

Let us fix arbitrarily small $\varepsilon > 0$. If the number of clauses is less than the threshold $(1 - \varepsilon)n$, then almost all 2-CNFs in n variables are feasible. On the

contrary, if the number of clauses is greater than the threshold $(1+\varepsilon)n$, then almost all 2-CNFs in n variables are unsatisfiable [2, 3].

Such a threshold is usually called a phase transition and is also shown by random samples of several types of formulae.

A k-CNF is called d-regular when each clause contains exactly k literals and each variable appears in exactly d clauses. For any sufficiently large number k, the membership of a random d-regular k-CNF to the set NAE-k-SAT undergoes a phase transition with increasing d at some critical value d_k, which depends on k. As the number of variables increases, for $d < d_k$, the fraction of d-regular k-CNF belonging to NAE-k-SAT tends to one. For $d > d_k$, this fraction tends to zero [4]. A similar result is known for d-regular k-CNFs having exactly two true literals per clause [5].

Next, let us consider a bound that holds for all, and not just almost all, formulae in consideration. In the proof, we replace a CNF with a system of algebraic equations depending on auxiliary variables, one per clause. So, the original satisfiability problem is reduced to the problem of the incidence of an affine subspace defined by a system of linear equations and a set of points with coordinates from the set $\{0,1\}$. From a geometric point of view, removing a clause corresponds to a projection onto some coordinate subspace [6]. In turn, the projection corresponds to eliminating the auxiliary variable. The solution to a system of equations in which each variable takes values from the set $\{0,1\}$ is called a $(0,1)$-solution. The existence of a $(0,1)$-solution to a system of linear equations over the field of rational numbers is also a well-known computationally difficult problem [7].

1. Results

Theorem 1. *Given a system of m linear equations of the type*

$$y_j = \ell_j(x_1, \ldots, x_n)$$

in $m+n$ variables $y_1, \ldots, y_m, x_1, \ldots, x_n$, where ℓ_j denotes a linear function over a certain field. If this system has no $(0,1)$-solution and the inequality $m > 2n+2$ holds, then there is an equation in the system such that the subsystem obtained by removing this equation also has no $(0,1)$-solution.

Theorem 2. *Given a propositional CNF $\varphi(p_1, \ldots, p_n)$ with m clauses in n variables. If φ does not belong to α-or-β-in-SAT and the inequality $m > 2n+2$ holds, then there exists a CNF that does not belong to α-or-β-in-SAT and is obtained by removing some clause from φ.*

Proof. Let us define by induction a function f that maps a clause to a pseudo-Boolean linear function over the field of rational numbers. $f(\bot) = 0$ and $f(\top) = 1$. For variables $f(p_i) = x_i$. For the negation $f(\neg p_i) = 1 - x_i$. Next, the jth clause $\varphi_j = \ell_1 \vee \cdots \vee \ell_k$ corresponds to the expression $f(\ell_1) + \cdots + f(\ell_k) - \alpha - (\beta - \alpha)y_j$, where the new variable y_j appears only once. Note that $\alpha \neq \beta$.

Next, the conjunction of clauses φ_j corresponds to the system of linear equations $f(\varphi_j) = 0$, where $1 \le j \le m$. This system depends on $m + n$ variables. All m equations are linearly independent, since each one depends on its own auxiliary variable. Every $(0,1)$-solution to the system corresponds to a $(\bot, \top)$-evaluation of propositional variables such that in each clause either exactly α or exactly β literals are true. Conversely, for such a $(\bot, \top)$-evaluation of propositional variables, there is a $(0,1)$-solution to the system of linear equations. If $p_i = \bot$, then $x_i = 0$. If $p_i = \top$, then $x_i = 1$. If α literals are satisfied in the jth clause, then $y_j = 0$. If β literals are satisfied, then $y_j = 1$. According to Theorem 1, if the system has no $(0,1)$-solution, then this property is preserved after eliminating some additional variable y_j, i.e., after removing the jth clause from φ.

2. Discussion

The bound on the number of clauses in an unsatisfiable subformula in 2-CNF is close to optimal. There is an unsatisfiable 2-CNF with $m = 2n$ clauses in n variables for which the subformula obtained by removing any clause is satisfiable. An example is 2-CNF

$$(\neg p_1 \vee p_2) \wedge (p_1 \vee \neg p_2) \wedge \cdots \wedge (\neg p_{n-1} \vee p_n) \wedge (p_{n-1} \vee \neg p_n) \wedge (p_n \vee p_1) \wedge (\neg p_n \vee \neg p_1),$$

where each variable enters twice positively and twice negatively. This 2-CNF is equivalent to the conjunction of formulae expressing the equivalence of the variables p_j and p_{j+1} for $j < n$, as well as the equivalence of the variable p_n and the negation of the variable p_1. It is impossible. But removing one clause from this 2-CNF corresponds to replacing some equivalence with an implication. The resulting formula is satisfiable for some $(\bot, \top)$-evaluation for which the antecedent of this implication is false.

Note that the bound on reducing the number of clauses lies in the segment where almost all 2-CNFs are unsatisfiable [2]. So, the possibility of removing some clause from an unsatisfiable 2-CNF does not impose unexpected additional restrictions on the unsatisfiable subformula. Also, the results obtained for other classes of formulae provide an upper estimate for the phase transition boundary, when it exists.

It is possible to consider conjunctions of formulae of another type. Let us denote by $\mathrm{maj}(p_1, p_2, p_3)$ the voting function (majority). Its value is equal to the most frequently occurring among $(\bot, \top)$-values of the propositional variables p_1, p_2, and p_3. For literals ℓ_{ij} the conjunction $\wedge_j \mathrm{maj}(\ell_{1j}, \ell_{2j}, \ell_{3j})$ is satisfiable if and only if 3-CNF $\wedge_j (\ell_{1j} \vee \ell_{2j} \vee \ell_{3j})$ belongs to the set 2-or-3-in-SAT. Therefore, Theorem 2 is applicable to formulae of this type.

Conclusion

Reducing the formula length leads to a decrease in some estimates of computational complexity. However, it does not guarantee a reduction in the running time of some heuristic algorithms. On the other hand, the main result is a pure existence theorem, which does not provide a fast algorithm for finding an unsatisfiable subformula. The result is consistent with the hypothesis that the satisfiability problem is computationally hard in the worst-case.

References

[1] J. Peng, M. Xiao, Further improvements for SAT in terms of formula length. Information and Computation, 2023, vol. 294, no. 105085. https://doi.org/10.1016/j.ic.2023.105085

[2] A. Goerdt, A threshold for unsatisfiability. Journal of Computer and System Sciences, 1996, vol. 53, no. 3, pp. 469–486. https://doi.org/10.1006/jcss.1996.0081

[3] D. Achlioptas, A. Coja-Oghlan, M. Hahn-Klimroth, J. Lee, N. Müller, M. Penschuck, G. Zhou, The number of satisfying assignments of random 2-SAT formulas. Random Structures and Algorithms, 2021, vol. 58, no. 4, pp. 609–647. https://doi.org/10.1002/rsa.20993

[4] J. Ding, A. Sly, N. Sun, Satisfiability threshold for random regular NAE-SAT. Communications in Mathematical Physics, 2016, vol. 341, pp. 435–489. https://doi.org/10.1007/s00220-015-2492-8

[5] *G. Nie, D. Xu, X. Wang, X. Wang*, The phase transition analysis for the random regular exact 2-(d, k)-SAT problem, Symmetry, 2021, vol. 13, no. 7, article 1231. https://doi.org/10.3390/sym13071231

[6] A.A. Boykov, A.V. Seliverstov, On a cube and subspace projections. Vestn. Udmurtsk. Univ. Mat. Mekh. Komp. Nauki, 2023, vol. 33, no. 3, pp. 402–415. (In Russian.) https://doi.org/10.35634/vm230302

[7] O.A. Zverkov, A.V. Seliverstov, Effective lower bounds on the matrix rank and their applications. Programming and Computer Software, 2023, vol. 49, no. 5, pp. 441–447. https://doi.org/10.1134/S0361768823020160

Alexandr V. Seliverstov
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)
Moscow, Russia
e-mail: slvstv@iitp.ru