On the Satisfiability of Random 3-SAT Formulas with k-Wise Independent Clauses

Ioannis Caragiannis
Department of Computer Science, Aarhus University, Denmark

Nick Gravin
Key Laboratory of Interdisciplinary Research of Computation and Economics,
Shanghai University of Finance and Economics, Ministry of Education, China

Zhile Jiang
Department of Computer Science, Aarhus University, Denmark

Abstract

The problem of identifying the satisfiability threshold of random 3-SAT formulas has received a lot of attention during the last decades and has inspired the study of other threshold phenomena in random combinatorial structures. The classical assumption in this line of research is that, for a given set of n Boolean variables, each clause is drawn uniformly at random among all sets of three literals from these variables, independently from other clauses. Here, we keep the uniform distribution of each clause, but deviate significantly from the independence assumption and consider richer families of probability distributions. For integer parameters n, m, and k, we denote by $\mathcal{F}_k(n,m)$ the family of probability distributions that produce formulas with m clauses, each selected uniformly at random from all sets of three literals from the n variables, so that the clauses are k-wise independent. Our aim is to make general statements about the satisfiability or unsatisfiability of formulas produced by distributions in $\mathcal{F}_k(n,m)$ for different values of the parameters n, m, and k.

Our technical results are as follows: First, all probability distributions in $\mathcal{F}_2(n,m)$ with $m \in \Omega(n^3)$ return unsatisfiable formulas with high probability. This result is tight. We show that there exists a probability distribution $\mathcal{D} \in \mathcal{F}_3(n,m)$ with $m \in O(n^3)$ so that a random formula drawn from $\mathcal{D}$ is almost always satisfiable. In contrast, for $m \in \Omega(n^2)$, any probability distribution $\mathcal{D} \in \mathcal{F}_4(n,m)$ returns an unsatisfiable formula with high probability. This is our most surprising and technically involved result. Finally, for any integer $k \geq 2$, any probability distribution $\mathcal{D} \in \mathcal{F}_k(n,m)$ with $m \in O(n^{1-1/k})$ returns a satisfiable formula with high probability.

2012 ACM Subject Classification Theory of computation → Generating random combinatorial structures; Mathematics of computing → Combinatoric problems; Mathematics of computing → Probability and statistics

Keywords and phrases Random 3-SAT, k-wise independence, Random bipartite graph

Digital Object Identifier 10.4230/LIPIcs.ESA.2025.103

Related Version *Full Version*: `https://arxiv.org/abs/2411.03813`

Funding The research is supported by National Key R&D Program of China (2023YFA1009500) and "the Fundamental Research Funds for the Central Universities".

1 Introduction

Satisfiability of propositional formulas (SAT) is one of the most renowned problems in theoretical computer science. It appeared in the first lists of NP-complete problems independently proposed by Cook and Levin [44], and is pivotal for many developments in modern complexity theory. Today, many lower bounds on the running time of algorithms rely on the Exponential Time Hypothesis for solving SAT [11, 18, 36, 37]. On the practical side, SAT solvers are frequently deployed in hardware circuit design, model checking, program verification, automated planning and scheduling, as well as in solving real-life instantiations

33rd Annual European Symposium on Algorithms (ESA 2025).
Editors: Anne Benoit, Haim Kaplan, Sebastian Wild, and Grzegorz Herman; Article No. 103; pp. 103:1–103:17
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of combinatorial optimization problems such as FCC spectrum auctions. Modern SAT solvers often find solutions to large industrial instances with thousands or even millions of variables despite the NP-hardness of the problem. However, there is still a large discrepancy between the performance of SAT solvers on those instances and theoretical average-case predictions, which have been studied in great depth under the line of research on *random SAT*.

Random SAT

A j-CNF formula ϕ over n variables is composed of m OR-clauses, each containing exactly j literals of j different variables. In the most commonly studied random SAT model, a formula ϕ is generated uniformly at random from all possible j-CNF formulas over n variables and m clauses. The most prominent theoretical question related to random SAT is to identify the satisfiability threshold r_j such that $\lim_{n\to\infty} \mathbf{Pr}[\phi \text{ is satisfiable}]$ is equal to 0 when $m/n > r_j$, and equal to 1 when $m/n < r_j$. It has been established [15] that 2-SAT has $r_2 = 1$, and its phase transition window [10] is $m \in [n - \Theta(n^{1/3}), n + \Theta(n^{1/3})]$. For $j \geq 3$, the asymptotic j-SAT threshold was shown to be $2^j \log 2 - \frac{1}{2}(1 + \log 2) \pm o_j(1)$ as $j \to \infty$ [17] (improving previous results from [3]), while for large enough j the exact value of r_j was determined in [20]. However, the question of identifying r_j for small values of j remains open. In particular, random 3-SAT has attracted a lot of attention. For the lower bound part, it has been shown in a series of papers [15, 12, 32, 1, 35, 39] that $r_3 \geq 3.52$ (the currently best known bound is due to [35, 39]). The upper bound part is studied by [27, 38, 21, 19]; the currently best known bound is $r_3 < 4.49$ due to [19]. The estimate $r_3 \approx 4.26$ was derived from numerical experiments [40] (see also [14, 41]).

A more recent line of work [29, 30, 31, 42] extends the standard model of random j-SAT to non-uniform distributions. Their motivation comes from the empirical observation that, in practice, CNF formulas often have rather different frequencies/probabilities for the n variables to appear in each clause (following a power-law distribution instead of a uniform one). Namely, Friedrich and Rothenberger [31] proposed a non-uniform random model, where the literals $\{x_i, \overline{x_i}\}_{i\in[n]}$ are selected independently at random in each clause c of the random j-CNF with $\mathbf{Pr}[x_i \in c] = \mathbf{Pr}[\overline{x_i} \in c] = p_i$ and where probabilities $\mathbf{p} = (p_i)_{i\in[n]}$ may vary across different variables. They find satisfiability threshold $r_2(\mathbf{p})$ of non-uniform random 2-SAT for certain regimes depending on $\mathbf{p}$. However, the non-uniform model of [31] does not capture the community biases/correlations (i.e., the fact that certain variables are more likely to appear together in a clause), which are often observed in practice [6]. This leads us to the question of whether it is possible to relax the strong independence assumption in the existing random SAT literature.

Relaxation of independence

We first observe that it does not make much sense to study distributions of SAT formulas with arbitrary correlations over the clauses. Indeed, by allowing correlation between several clauses, one may enforce that the random formula ϕ contains large fixed sub-formulas corresponding to NP-hard SAT variants. This would be at odds with our goal of studying average-case complexity. Therefore, we must keep a certain degree of independence in the distribution of instances. We propose to consider the relaxation of mutual independence over m clauses in a random formula ϕ *to k-wise independence* for a small constant k. To keep the new model tractable, we focus on 3-SAT and uniform distribution of literals within each clause. I.e., we assume that (i) every 3-OR-clause c of a random 3-CNF formula ϕ has three literals of three distinct variables drawn uniformly at random among all such triplets of literals and that (ii) given this marginal distribution of each clause $c \sim F_{\text{uni.}}$, the distribution $\mathcal{D}$ over the clause set C in ϕ is only k-wise independent instead of the mutually independent distribution $\mathcal{D}_{\text{Ind.}} = (F_{\text{uni.}})^{\otimes m}$ in the standard model. This is a natural generalization that has been considered in a number of different settings but, to the best of our knowledge, not in the context of random SAT. Note that the smaller k is, the bigger the set of possible distributions $\mathcal{D}$. Furthermore, for small values of k, a k-wise independent distribution $\mathcal{D}$ can still capture a large class of dependencies among clauses but at the same time does not allow correlation between any k-tuples of clauses. In mathematical terms, the family of discrete k-wise independent distributions naturally appears when we map the set of distributions to the set of their low-degree moments. Specifically, if a distribution $\mathcal{D}$ is supported on the n-dimensional binary cube $\mathtt{supp}(D) = \{-1, 1\}^n$, then all its moments of degree up to k can be described as $\mu(D) = (\mathbf{E}[\prod_{i \in S} x_i])_{|S| \le k}$. As low-degree moments (basically, the image of μ) are extremely important in statistical analysis, it is equally important to study the kernel of the aforementioned mapping, which exactly corresponds to the family of k-wise independent distributions. Let us provide additional justifications of our framework by discussing some of the theoretical work on random 3-SAT and on other settings with a similar k-wise independence relaxation.

Pseudo-randomness. Historically, the k-wise relaxation of independence has been actively used in the literature on derandomization and pseudo-randomness, as it allows to significantly reduce the amount of random bits needed to generate random objects. For example, Alon and Nussboim [5] consider random Erdős-Rényi graphs and examine the minimal degree k of independence needed to achieve a variety of graph properties and statistics (such as connectivity, existence of perfect matchings, existence of Hamiltonian cycles, clique and chromatic numbers, etc.) that match those in the mutually independent case. Benjamini et al. [9] consider similar questions for monotone boolean functions. The motivation in [5] comes from the fact that there are efficient constructions of k-wise independent distributions with "low degree of independence" (say $k = O(\log n)$) that utilize only polylog(n) random bits, i.e., much fewer than the polynomial number of random bits required to generate mutually independent distributions. While some of this motivation can be applied to our setting of random 3-SAT, it is a conceptually different story. Indeed, the perspective of pseudo-random generation is through the lenses of "probability theory", where one controls the distributions and can simply choose one that satisfies necessary conditions such as, e.g., $(\log n)$-wise independence. On the other hand, our motivation stems from "statistics", as our ideal model should have a reasonable fit to empirical observations. So, we would like to use as minimal assumptions as possible and study small (constant) degrees of independence.

Refutability of 3-SAT. While the research on lower bounds for random 3-SAT often comes up with certain simple heuristics that efficiently find a satisfying assignment (see the surveys by Achlioptas [2] and Flaxman [26]), it is extremely hard to find an efficient refutation of a unsatisfiable 3-SAT formula. Indeed, the common approach to refute a given SAT formula is proof in resolution. Chvatal and Szemeredi [16] first showed that a random 3-CNF formula with $m = \Theta(n)$ clauses (which is almost surely unsatisfiable) almost surely admits exponential size proof in resolution. Later, Ben-Sasson and Wigderson [8] derived similar result for much larger $m = O(n^{3/2-\varepsilon})$. On the positive side, [28] gave the first polynomial time algorithm via spectral techniques that almost surely[1] refutes a random

[1] Refutation in this case is an algorithm with one-sided error: it always refutes the formula correctly by

3-SAT formula with $m = n^{3/2+\varepsilon}$ clauses. The best known bound on m is due to Feige and Ofek [25] who proved that, for a sufficiently large constant c, random 3-SAT formulas with $m = c \cdot n^{3/2}$ clauses can be almost surely refuted in polynomial time using another spectral graph algorithm. We note that a similar situation (extremely high probability of unsatisfiability for a random formula and inability to efficiently confirm it) is unlikely to happen in our k-clause independent model for constant k. Indeed, the main proof approach for dealing with arbitrary k-wise independent distribution is to define a k-wise statistic, which differentiates any satisfiable formula from a typical unsatisfiable one.

Testing k-wise independence. The property of k-wise independence of a distribution with n components can be tested using $n^{O(k)} = \text{poly}(n)$ many samples in polynomial time, when k is a constant [4, 43]. This is a useful property to have, as it allows one to verify with only polynomially many instances of random 3-SAT, whether these instances conform to k-wise independence or not.

Robust mechanism design. A recent line of work in robust mechanism design also considers families of k-wise independent Bayesian priors in single and multi-unit auctions [13, 22, 33, 34]. Their motivation is similar to ours, as they also rely on the statistical point of view to justify the extension of the results for mutual independent priors typically assumed in Bayesian mechanism design to k-wise independent ones.

1.1 Problem formulation

We consider random 3-CNF formulas with n variables generated from a distribution $\mathcal{D}$ over m clauses, where the mutual independence assumption over clauses is relaxed to k-wise independence. We use the term *k-clause independence* to refer to such distributions. We denote such families of distributions by $\mathcal{F}_k(n, m)$, where each $\mathcal{D} \in \mathcal{F}_k(n, m)$ has identical marginals uniformly distributed over all possible OR-clauses and those marginals are only assumed to be k-wise independent in $\mathcal{D}$. We would like to understand the following question for small values of k:

> How does the satisfiability threshold r_3 of random 3-SAT formulas behave under any k-clause independent distribution $\mathcal{D} \in \mathcal{F}_k(n, m)$?

As the distribution $\mathcal{D}$ is not unique, there might be a large gap between lower and upper estimates of r_3. To this end, we formally define the *lower satisfiability threshold* $\texttt{LST}_k(n)$ as an upper bound on m, such that a random formula ϕ drawn from a distribution in $\mathcal{F}_k(n, m)$ with $m \leq \texttt{LST}_k(n)$ clauses has $\mathbf{Pr}[\phi \text{ is satisfiable}] \geq \frac{2}{3}$. Similarly, the *upper satisfiability threshold* $\texttt{UST}_k(n)$ is a lower bound on m, such that the random formula ϕ with $m \geq \texttt{UST}_k(n)$ clauses has $\mathbf{Pr}[\phi \text{ is satisfiable}] \leq \frac{1}{3}$. What kind of bounds on upper $\texttt{UST}_k(n)$ and lower $\texttt{LST}_k(n)$ thresholds should we expect?

Reasonable expectations

The condition $\mathcal{D} \in \mathcal{F}_k(n, m)$ only says something about configurations of at most k clauses and does not put any other restrictions on the random formula $\phi \sim \mathcal{D}$. As the degree of independence k is a small constant, any argument that gives bounds on $\texttt{LST}_k$ or $\texttt{UST}_k$ can only rely on statistics of at most a constant number of clauses. Hence, it is rather likely that bounds on $\texttt{LST}_k$ and $\texttt{UST}_k$ come together with efficient procedures of, respectively, finding a satisfying assignment for a random formula ϕ, or certifying that ϕ is not satisfiable. Hence, given the prior work on random 3-SAT for $\mathcal{D}_{\text{Ind.}} \in \mathcal{F}_k(n, m)$, we get the following picture:

producing certain certificates, or says that the formula might be correct.

Upper satisfiability threshold. The best known result for refuting 3-CNF formulas efficiently is due to Feige and Ofek [25], who show how to do it only for a large number of clauses $m = c \cdot n^{3/2}$. Furthermore, for any smaller number of clauses $m = O(n^{3/2-\varepsilon})$, a random 3-CNF formula is likely to have only exponential in n proof size for any unsatisfiability proof in resolution [8]. Hence, it is out of reach to aim for a better bound on $\texttt{UST}_k(n)$ than $O(n^{3/2})$ while relying only on k-wise independence for some constant k. In fact, the best known positive result on *efficiently computable proofs* of unsatisfiability in resolution is due to Beame et al. [7], who show that an ordered DLL algorithm executed on a random 3-SAT instance with $m = \Omega(n^2/\log n)$ clauses terminates in polynomial time.

Lower satisfiability threshold. As the proofs for the lower bounds on r_3 often establish simple procedures that find satisfying assignments with high probability, it is still possible that $\texttt{LST}_k(n)$ is of similar order $\Theta(n)$ as the lower bounds on r_3 for $\mathcal{D}_{\text{Ind.}}$. Thus, the most ambitious result would be to show that $\texttt{LST}_k(n) \le c_k \cdot n$ for constant c_k that increases with k. A more modest goal is to aim for $\texttt{LST}_k(n) = o(n)$ for a constant k, where $\texttt{LST}_k(n) \to \Theta(n)$ as $k \to +\infty$.

1.2 Our results

We obtain the following bounds on the upper and lower satisfiability thresholds $\texttt{UST}_k(n)$ and $\texttt{LST}_k(n)$ for various values of k.

Upper satisfiability thresholds

We first consider small degrees of independence, i.e., $k \in \{2, 3\}$. In both cases, we show that $\texttt{UST}_k(n) = \Theta(n^3)$, meaning that one needs almost all possible clauses in a 3-clause (as well as 2-clause) independent formula to ensure that it is unsatisfiable (see **Theorem 4** and **Theorem 7**). The most nontrivial part is to construct the distribution $\mathcal{D} \in \mathcal{F}_3(n, m)$ with $m = \Theta(n^3)$ and $\mathbf{Pr}[\phi \text{ is satisfiable}] \ge \frac{2}{3}$. Our construction is based on "3-XOR formulas" (i.e., OR-clauses that have either one or three literals that are satisfied by a randomly planted truth assignment), which aligns well with the intuition developed in previous work [24, 25]. The main technical difficulty is to ensure k-clause independence by adding a small fraction of unsatisfiable instances and checking all k-wise statistics.

Our most exciting and technically involved result (see **Theorem 8**) is our proof that $\texttt{UST}_4(n) = O(n^2)$, i.e., a random formula $\phi \sim \mathcal{D}$ with $m = \Omega(n^2)$ clauses is unsatisfiable with large probability for any 4-clause independent distribution $\mathcal{D} \in \mathcal{F}_4(n, m)$. It is worth noting that such a bound is much harder to get under the 4-wise independence assumption than in the case of a mutually independent distribution $\mathcal{D}_{\text{Ind.}}$. Indeed, Feige and Ofek [25] describe a very simple refutation algorithm for $m = \Theta(n^2)$ that fixes a variable x and considers all clauses containing x or $\overline{x}$ (there will be $\Theta(n)$ such clauses in expectation). Then, after deleting x (or $\overline{x}$), one can reduce the problem to the refutation of the respective random 2-CNF sub-instance, which can be easily verified in polynomial time and has a low satisfiability threshold of $r_2 = 1$. This simple approach obviously fails for 4-clause independent distributions. We instead construct a bipartite multigraph $G(\phi)$ between pairs of distinct literals on one side and all singleton literals on the other, in which every OR-clause in ϕ corresponds to three different edges. We then carefully examine the statistic $\kappa(\phi)$ that counts $K_{2,2}$ subgraphs in $G(\phi)$ for a random $\phi \sim \mathcal{D}$. We find that the expected value of $\kappa(\phi)$ for random ϕ is only slightly larger than its absolute minimal value, while at the same time $\kappa(\phi)$ is significantly larger than its expectation when ϕ is satisfiable. Our argument bears certain similarities with the argument in [25], which also looked at intersections of two literals between pairs of clauses but used the 3-XOR principle and a differently constructed non-bipartite graph.

Lower satisfiability thresholds

We show (in **Theorem 14**) that $\texttt{LST}_k(n) \geq \Omega(n^{1-1/k})$ for any $k \geq 2$. I.e., any k-clause independent random formula is satisfiable with high probability if it contains at most $O(n^{1-1/k})$ clauses. The argument is simple: for any k-clause independent distribution, we look at the 3-uniform hypergraph that corresponds to the variables of a random formula ϕ produced according to this distribution, and argue that this graph does not have Berge-cycles, with high probability. We also provide an informal justification that this bound is asymptotically tight, i.e., that $\texttt{LST}_k(n) = O(n^{1-1/k})$. Specifically, we outline a plausible approach for constructing a k-clause independent distribution with $m = O(n^{1-1/k})$ clauses such that most of its formulas are unsatisfiable. Our approach is built upon existing constructions of dense hyper-graphs with large girth. It is interesting to note that, in both the proof of the $\texttt{LST}_k(n) = \Omega(n^{1-1/k})$ result and the approach for showing that $\texttt{LST}_k(n) = O(n^{1-1/k})$, we only need to consider variables and can completely ignore the distribution over the literals.

2 Preliminaries

Let $x_1, x_2, \cdots, x_n$ be n boolean variables. A literal ℓ is a boolean variable or the negation of it. For convenience, we usually represent a literal as a variable-sign pair, i.e., $\ell = (x_i, s)$, where s is the positive sign $\texttt{+}$ if the literal is a boolean variable and the negative sign $\texttt{-}$ if it is its negation. We define $\Sigma(n) = \{\texttt{+}, \texttt{-}\}^n$. An instance of the *Satisfiability* problem in conjunctive normal form (or *SAT instance*, for short) is a boolean formula over a subset of the n variables which is a conjunction of disjunctive clauses, each clause containing literals with different variables. In a 3-SAT instance, every clause has exactly 3 literals. Each SAT instance C can be described as a multiset of clauses. We denote the size of an instance C as $m = |C|$.

Given the number of variables n, let $X(n)$ be the set of variables. Let $T(n)$ be the set of all unordered triplets of variables $T(n) = \{(x_i, x_j, x_k) \mid x_i, x_j, x_k \in X(n), i < j < k\}$ and $\mathcal{T}(n)$ be the set of all possible clauses, i.e., all triplets with literals of different variables. We will usually omit the dependency of X, T, and $\mathcal{T}$ on n, when the value of n is clear from the context. To refer to the variable (the set of variables) or the sign (the set of signs) of a literal (a clause $c \in C$), we define operators $\mathrm{Var}(\cdot)$ and $\mathrm{Sign}(\cdot)$, so that if, e.g., $c = ((x_1, \texttt{+}), (x_2, \texttt{-}), (x_3, \texttt{+}))$, then $\mathrm{Var}(c) = (x_1, x_2, x_3)$ and $\mathrm{Sign}(c) = (\texttt{+}, \texttt{-}, \texttt{+})$.

A *truth assignment* is a vector $\sigma \in \{0, 1\}^n$, where 1 or 0 at the σ_i coordinate corresponds to the "true" or "false" value of x_i, respectively. A truth assignment σ is *satisfying* for an instance C when each clause $c \in C$ has at least one true literal under σ. An instance C is *satisfiable* if there exists a satisfying assignment; otherwise, C is *unsatisfiable.*

The random 3-SAT model assumes that the instance is drawn from a probability distribution over all possible 3-SAT instances with m clauses and n variables. The main object of study in such a model is the probability of such a random instance being satisfiable as a function of n and m.

2.1 Random 3-SAT without mutual independence

In the standard random 3-SAT model, each clause is drawn uniformly at random among all clauses in $\mathcal{T}(n)$, independently from the other clauses. A constructive definition is given by Model 1. We denote the distribution over instances generated by Model 1 as $\mathcal{D}_{\text{Ind.}}$.

MODEL 1 Selects a 3-SAT instance uniformly at random from all possible instances.

```
Input: Integers $n \geq 3$ and $m \geq 1$
Output: A 3-SAT instance with $n$ variables and $m$ clauses
$C \leftarrow \emptyset$;
for $l \leftarrow 1, \ldots, m$ do
    pick a literal triplet $(\ell_1, \ell_2, \ell_3)$ uniformly at random from $\mathcal{T}(n)$;
    $c \leftarrow (\ell_1, \ell_2, \ell_3)$;
    $C \leftarrow C \cup \{c\}$;
end
return $C$;
```

Our main focus will be on k-wise relaxations of the independent distribution.

▶ **Definition 1** (k-clause independent random SAT). *A distribution $\mathcal{D}$ for selecting a random SAT instance is k-clause independent if the set of any fixed k clauses $S \subseteq C$ is distributed uniformly at random over all k-tuples of possible clauses, i.e.,*

$$\mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}\left[c_i = t_i, \forall i \in [k]\right] = \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}_{Ind.}}\left[c_i = t_i, \forall i \in [k]\right] = \prod_{i\in[k]} \mathbf{Pr}\left[c_i = t_i\right].$$

for every $c_1, \ldots, c_k \in C$ and $t_1, \ldots, t_k \in \mathcal{T}$. We denote the family of all k-clause independent distributions with m clauses over n variables as $\mathcal{F}_k(n, m)$.

We remark that, in contrast to the random 3-SAT model (Model 1), which defines a single distribution for given n and m, the family $\mathcal{F}_k(n, m)$ contains many different distributions.

By definition, the probability of any event A_S that depends only on a subset S of at most k clauses in the k-clause independent distribution $\mathcal{D}$ is the same as for $\mathcal{D}_{\text{Ind.}}$. I.e., the expectations of the indicator function $\mathbb{I}[A_S]$ are the same for $\mathcal{D}$ and $\mathcal{D}_{\text{Ind.}}$. Hence, by linearity of expectation, any statistic that involves only $k' \leq k$ clauses must be the same for $\mathcal{D}$ and $\mathcal{D}_{\text{Ind.}}$, i.e.,

$$\mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\sum_{\substack{S\subseteq C,\\ |S|=k'}} \mathbb{I}\left[A_S\right]\right] = \mathop{\mathbf{E}}_{C\sim\mathcal{D}_{\text{Ind.}}}\left[\sum_{\substack{S\subseteq C,\\ |S|=k'}} \mathbb{I}\left[A_S\right]\right]. \tag{1}$$

We would like to understand what are the largest/smallest possible number of clauses for a random 3-SAT formula to be almost surely satisfiable/unsatisfiable under any k-clause independent distribution. Using $SAT(C)$ to denote the event that the SAT instance C is satisfiable, we define the following satisfiability "thresholds".

▶ **Definition 2** (Upper satisfiability threshold). *The upper satisfiability threshold $\mathtt{UST}_k(n)$ is defined as follows. For integer $n \geq 3$, $\mathtt{UST}_k(n)$ is the minimum integer m such that*

$$\mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}\left[SAT(C)\right] \leq 1/3, \quad \forall \mathcal{D} \in \mathcal{F}_k(n, m).$$

▶ **Definition 3** (Lower satisfiability threshold). *The lower satisfiability threshold* $\mathtt{LST}_k(n)$ *is defined as follows. For integer* $n \geq 3$, $\mathtt{LST}_k(n)$ *is the maximum integer* m *such that*

$$\Pr_{C\sim\mathcal{D}}[\neg SAT(C)] \leq 1/3, \quad \forall \mathcal{D} \in \mathcal{F}_k(n,m).$$

Extending the line of research on the standard random 3-SAT model, we would like to have as tight estimates of $\mathtt{UST}_k(n)$ and $\mathtt{LST}_k(n)$ as possible.

3 Tight bounds for the upper satisfiability threshold $\mathtt{UST}_2(n)$

We begin with a technical warm up and prove asymptotically tight bounds on the upper satisfiability threshold of 2-clause independent random 3-SAT.

▶ **Theorem 4.** $\mathtt{UST}_2(n) = \Theta(n^3)$.

Theorem 4 follows by the next two lemmas. Lemma 5 provides an upper bound on $\mathtt{UST}_2(n)$. In the proof, we introduce the technique that we will use later in Section 5 to get a much more involved upper bound on $\mathtt{UST}_4(n)$.

▶ **Lemma 5.** *For any* $\mathcal{D} \in \mathcal{F}_2(n,m)$ *with* $m \geq 56\binom{n}{3}$, $\mathbf{Pr}_{C\sim\mathcal{D}}[SAT(C)] \leq 56\binom{n}{3}/m$.

Proof. Let $\xi(C) \overset{\text{def}}{=\!=} \sum_{c_i,c_j\in C} \mathbb{I}[c_i = c_j]$ be the number of identical clause pairs in the instance C. On the one hand, Equation (1) implies that the expectation of $\xi(C)$ is the same when C is drawn from a 2-clause independent distribution and $\mathcal{D}_{\text{Ind.}}$. On the other hand, if an instance C has a satisfying assignment, at least 1/8 of possible clauses from $\mathcal{T}$ must not appear in C, which means that the value of $\xi(C)$ is significantly higher than its expectation. These two observations will allow us to get the desired bound on the probability $\mathbf{Pr}_{C\sim\mathcal{D}}[SAT(C)]$.

Let $\mathcal{D} \in \mathcal{F}_2(n,m)$ and define $p \overset{\text{def}}{=\!=} \mathbf{Pr}_{C\sim\mathcal{D}}[SAT(C)]$. We shall derive two lower bounds on the value the random variable $\xi(C)$ can get: an unconditional lower bound $\mathtt{LB}[\xi]$ (not far from $\mathbf{E}_{C\sim\mathcal{D}}[\xi(C)]$) and a lower bound $\mathtt{LB}[\xi \mid SAT]$ on $\xi(C)$ when C is satisfiable (this will be significantly larger than $\mathbf{E}_{C\sim\mathcal{D}}[\xi(C)]$). We note that

$$\begin{aligned}\mathbf{E}_{C\sim\mathcal{D}}[\xi(C)] &= \mathbf{Pr}[SAT(C)] \cdot \mathbf{E}[\xi(C) \mid SAT(C)] \\ &\quad + \mathbf{Pr}[\neg SAT(C)] \cdot \mathbf{E}[\xi(C) \mid \neg SAT(C)] \\ &\geq p \cdot \mathtt{LB}[\xi \mid SAT] + (1-p)\cdot \mathtt{LB}[\xi]. \end{aligned} \tag{2}$$

We next derive $\mathbf{E}_{C\sim\mathcal{D}}[\xi(C)]$, $\mathtt{LB}[\xi]$, and $\mathtt{LB}[\xi \mid SAT]$. We denote as $M \overset{\text{def}}{=\!=} |\mathcal{T}| = 8\binom{n}{3}$ the number of possible clauses and $\lambda \overset{\text{def}}{=\!=} m/M > 1$. First, we observe that

$$\mathbf{E}_{C\sim\mathcal{D}}[\xi(C)] = \sum_{c_i,c_j\in C} \mathbf{E}\left[\mathbb{I}\left[c_i = c_j\right]\right] = \frac{m^2-m}{2}\cdot \mathbf{Pr}[c_i = c_j] = \frac{m^2-m}{2\cdot M} = m\cdot\frac{\lambda\cdot M-1}{2\cdot M}. \tag{3}$$

To derive the lower bounds $\mathtt{LB}[\xi]$ and $\mathtt{LB}[\xi \mid SAT]$, we observe that any given clause type $t \in \mathcal{T}$ contributes $\binom{d_t}{2}$ pairs to $\xi(C)$, where $d_t \overset{\text{def}}{=\!=} \sum_{c\in C}\mathbb{I}[c = t]$ is the number of type t clauses in C. I.e.,

$$\xi(C) = \sum_{t\in\mathcal{T}} \frac{d_t^2 - d_t}{2}, \qquad \text{where } \sum_{t\in\mathcal{T}} d_t = m. \tag{4}$$

The minimum of $\xi(C) = \frac{1}{2}\sum d_t^2 - \frac{m}{2}$ under the constraint $\sum d_t = m$ is achieved when all the d_t variables are equal, i.e., equal to $\frac{m}{|\mathcal{T}|} = \lambda$. Thus, we get the following lower bound $\texttt{LB}[\xi]$ on $\xi(C)$:

$$\xi(C) \geq \texttt{LB}[\xi] \overset{\text{def}}{=\!=} \frac{1}{2}\sum_{t\in\mathcal{T}} \lambda^2 - \frac{m}{2} = m \cdot \frac{\lambda - 1}{2}. \tag{5}$$

To derive the lower bound $\texttt{LB}[\xi \mid SAT]$ on $\xi(C)$ for a satisfiable formula C, we note that at least a $\frac{1}{8}$-fraction of the clause types are not present in C. I.e., at least $\frac{M}{8}$ of the d_t variables have value equal to 0 in (4). Similarly to (5), the minimum value of $\xi(C)$ is achieved when all the remaining $\frac{7M}{8}$ d_t variables are equal to each other, getting the value $\frac{8\cdot m}{7\cdot M}$. That is,

$$\xi(C) \geq \texttt{LB}[\xi \mid SAT] \overset{\text{def}}{=\!=} \frac{7\cdot M}{8} \cdot \left(\frac{8\lambda}{7}\right)^2 - \frac{m}{2} = m \cdot \frac{\frac{8}{7}\lambda - 1}{2}. \tag{6}$$

We plug the bounds (3), (5), and (6) into (2) to get

$$m \cdot \frac{\lambda \cdot M - 1}{2 \cdot M} \geq p \cdot m \cdot \frac{\frac{8}{7}\lambda - 1}{2} + (1-p) \cdot m \cdot \frac{\lambda - 1}{2}.$$

After simple algebraic transformation, this is equivalent to the inequality $1 - \frac{1}{M} \geq \frac{p}{7}\lambda$, which implies that $p \leq \frac{7\cdot M}{m} = 56\binom{n}{3}/m$. ◀

To lower-bound $\texttt{UST}_2(n)$, we use a specific 2-clause independent distribution, which is depicted as Model 2. The next lemma proves the correctness of our construction. The proof is deferred to the full version.

▶ **Lemma 6.** *Model 2 defines a 2-clause independent probability distribution that generates 3-SAT instances of size $\Omega(n^3)$ that are satisfiable with probability at least $1 - O(n^{-3})$.*

MODEL 2 Selects a 3-SAT instance according to a 2-clause independent distribution.

```
Input: Integer $n \geq 3$
Output: A 3-SAT instance $C$ with $n$ variables and $m = \binom{n}{3}$ clauses
- With probability $1 - \binom{n}{3}^{-1}$, construct a satisfiable instance $C$:
    1. Match $m = \binom{n}{3}$ clauses to all different $T(n)$ variable triplets uniformly at random;
    2. Pick a random truth assignment $\sigma \sim \texttt{Uni}[\{0,1\}^n]$;
    3. For each clause $c \in C$ matched to the variable triplet $(x_i, x_j, x_k) \in T(n)$
        - Pick a random sign triplet $(s_1, s_2, s_3)$ of the same parity with $(\sigma_i, \sigma_j, \sigma_k)$
          (i.e., $\mathbb{I}[s_1 = \texttt{-}] + \mathbb{I}[s_2 = \texttt{-}] + \mathbb{I}[s_3 = \texttt{-}] + \sigma_i + \sigma_j + \sigma_k = 0 \mod 2$);
        - Let clause $c \leftarrow ((x_i, s_1), (x_j, s_2), (x_k, s_3))$;
- With probability $\binom{n}{3}^{-1}$, sample $C$ from distribution $\mathcal{D}_{\texttt{uni-var}}$ defined as follows:
    1. Pick a random single variable triplet $(x_i, x_j, x_k) \sim \texttt{Uni}[T(n)]$;
    2. For each clause $c \in C$
        - Pick a random sign triplet $(s_1, s_2, s_3) \sim \texttt{Uni}[\{\texttt{+},\texttt{-}\}^3]$;
        - Let clause $c \leftarrow ((x_i, s_1), (x_j, s_2), (x_k, s_3))$;
```

4 A tight lower bound for $\texttt{UST}_3(n)$

Clearly, Lemma 5 also provides an upper bound on $\texttt{UST}_3(n)$. The proof of the next statement follows by presenting a matching lower bound through Model 3.

▶ **Theorem 7.** *$\texttt{UST}_3(n) = \Theta(n^3)$.*

MODEL 3 Selects a 3-SAT instance according to a 3-clause independent distribution.

```
Input: Integer $n \geq 3$ and even integer $m \leq \frac{1}{3} \cdot \binom{n}{3}$.
Output: A 3-SAT instance $C$ with $n$ variables and $m$ clauses.
Set $p \overset{\text{def}}{=\!=} (m-1) \cdot \left( \binom{n}{3}^{-1} - \frac{1}{3}\binom{n}{3}^{-2} \right)$ and $q \overset{\text{def}}{=\!=} \binom{n}{3}^{-2}$;
- With probability $1 - p - q$, construct a satisfiable instance $C$:
    1. Match $C$ clauses to $m$ different variable triplets in $T(n)$ uniformly at random;
    2. Pick a random truth assignment $\sigma \sim \mathtt{Uni}[\{0,1\}^n]$;
    3. For each clause $c \in C$ matched to the variable triplet $(x_i, x_j, x_k) \in T(n)$
        - Pick a random sign triplet $(s_1, s_2, s_3)$ of the same parity with $(\sigma_i, \sigma_j, \sigma_k)$
          (i.e., $\mathbb{I}\left[s_1 = \texttt{-}\right] + \mathbb{I}\left[s_2 = \texttt{-}\right] + \mathbb{I}\left[s_3 = \texttt{-}\right] + \sigma_i + \sigma_j + \sigma_k = 0 \mod 2$);
        - Let clause $c \leftarrow ((x_i, s_1), (x_j, s_2), (x_k, s_3))$;
- With probability $p$, construct an instance $C$ with $m/2$ different variable triplets:
    1. Uniformly at random match $C$ to $\frac{m}{2}$ different variable triplets in $T(n)$
       (exactly 2 clauses in $C$ per one variable triplet);
    2. For each clause $c \in C$ assigned to the variable triplet $(x_i, x_j, x_k) \in T(n)$
        - Pick a random sign triplet $(s_1, s_2, s_3) \sim \mathtt{Uni}[\{\texttt{+},\texttt{-}\}^3]$;
        - Let clause $c \leftarrow ((x_i, s_1), (x_j, s_2), (x_k, s_3))$;
- With probability $q$, sample $C$ from distribution $C \sim \mathcal{D}_{\texttt{uni-var}}$ as follows:
    1. Pick a random single variable triplet $(x_i, x_j, x_k) \sim \mathtt{Uni}[T(n)]$;
    2. For each clause $c \in C$
        - Pick a random sign triplet $(s_1, s_2, s_3) \sim \mathtt{Uni}[\{\texttt{+},\texttt{-}\}^3]$;
        - Let clause $c \leftarrow ((x_i, s_1), (x_j, s_2), (x_k, s_3))$;
```

We give an explicit construction (see Model 3) of a 3-clause independent distribution $\mathcal{D} \in \mathcal{F}_3(n, m)$ with n variables and (an even number of) $m \leq \frac{1}{3} \cdot \binom{n}{3}$ clauses, such that $\mathbf{Pr}_{C \sim \mathcal{D}}[SAT(C)] \geq 2/3 - O(n^{-6})$. Our construction in Model 3 follows the same pattern as in Model 2, but uses an additional step and minor modifications to ensure 3-clause independence.[2] The proof is deferred to the full version.

5 An upper bound for $\mathtt{UST}_4(n)$

Our next result is rather surprising as it indicates that 4-wise independence allows for a steep decrease in the upper satisfiability threshold compared to 2- and 3-wise independence.

▶ **Theorem 8.** *$\mathit{UST}_4(n) = O(n^2)$.*

Proof. We will prove the theorem by showing that for any positive integers n and m and any 4-clause independent probability distribution $\mathcal{D} \in \mathcal{F}_4(n, m)$, it holds $\mathbf{Pr}_{C \sim \mathcal{D}}[SAT(C)] \leq O\left(\max\left\{\frac{n^2}{m}, \frac{1}{\sqrt{n}}\right\}\right)$. The claim is obvious for $n < 10$. We will assume that $n \geq 10$ and $m \leq \sqrt{10} \cdot n^{5/2}$, and will show that $\mathbf{Pr}_{C \sim \mathcal{D}}[SAT(C)] \leq \frac{4288 \cdot n^2}{m}$ for every 4-clause independent probability distribution $\mathcal{D} \in \mathcal{F}_4(n, m)$. Note that for $m > \sqrt{10} \cdot n^{5/2}$, the probability bound of $O\left(\frac{1}{\sqrt{n}}\right)$ follows by selecting uniformly at random a subset of $\sqrt{10} \cdot n^{5/2}$ clauses.

We will use a graph representation of 3-SAT instances defined as follows. Given a 3-SAT instance C consisting of m clauses over n variables, the bipartite multi-graph $G(C) = (L \cup R, E)$ has a node corresponding to each (unordered) pair of literals $\{\ell_1, \ell_2\}$ from different variables at the left node side L and a node corresponding to each literal ℓ at the right node

[2] The first and third block of Model 3 correspond to the two blocks of Model 2. The only difference in the first block is that the matching of C is not to all the variable triplets in $T(n)$ but only to m of them.

side R. Hence, $|L| = 4\binom{n}{2}$ and $|R| = 2n$. For every clause $c = (\ell_1, \ell_2, \ell_3)$ of C, $G(C)$ has the three edges between the node corresponding to the pair of literals (ℓ_i, ℓ_j) and the node corresponding to literal ℓ_{6-i-j} for $(i,j) \in \{(1,2),(1,3),(2,3)\}$.

The main proof idea of Theorem 8 is to analyze the statistic $\kappa(C)$ defined as the number of distinct $K_{2,2}$ subgraphs in graph $G(C)$. Namely, we first derive an upper bound on the expectation $\mathbf{E}_{C\sim\mathcal{D}}[\kappa(C)]$ (see Lemma 9) when C is drawn from the 4-clause independent probability distribution $\mathcal{D} \in \mathcal{F}_4(n,m)$. We then give two lower bounds on the value of the random variable $\kappa(C)$ by considering an underlying simple subgraph of $G(C)$. Note that the underlying simple subgraph corresponds to a smaller instance $\widetilde{C}$ with $\widetilde{m}$ distinct clauses, in which we remove all duplicated clauses in C. As we show in Lemma 10, this does not significantly reduce the size of the instance. Our first lower bound (Lemma 11) on $\kappa(\widetilde{C})$ holds for any instance $\widetilde{C}$ and is very close to the upper bound on the expectation of $\kappa(C)$. On the other hand, when $\widetilde{C}$ is satisfiable, we manage to give a significantly stronger lower bound on $\kappa(\widetilde{C})$ in Lemma 12. We then conclude the proof of Theorem 8 by relating all these upper and lower bounds with $\mathbf{E}_{C\sim\mathcal{D}}[\kappa(C)]$ in Lemma 13. We defer the proofs of Lemmas 9, 10, 11, and 12 to the full version.

▶ **Lemma 9.** *For any $\mathcal{D} \in \mathcal{F}_4(n,m)$, it holds that $\mathbf{E}_{C\sim\mathcal{D}}[\kappa(C)] \le \frac{81m^4}{64n^6} + \frac{729m^3}{32n^4}$.*

▶ **Lemma 10.** $\mathbf{E}_{C\sim\mathcal{D}}[\widetilde{m}^4] \ge m^4 - \frac{125}{6} \cdot m^3 \cdot n^2$.

▶ **Lemma 11.** *Let C be an instance with n variables and $\widetilde{m}$ distinct clauses. Then,*

$$\kappa(\widetilde{C}) \ge \frac{81\widetilde{m}^4}{64n^6} - \frac{27\widetilde{m}^3}{8n^4}.$$

▶ **Lemma 12.** *Let C be a statisfiable instance with n variables and $\widetilde{m}$ distinct clauses. Then*

$$\kappa(\widetilde{C}) \ge \frac{82\widetilde{m}^4}{64n^6} - \frac{123\widetilde{m}^3}{16n^4}.$$

▶ **Lemma 13.** *For any $\mathcal{D} \in \mathcal{F}_4(n,m)$, we have $\mathbf{E}_{C\sim\mathcal{D}}[\kappa(C)] \ge \frac{81m^4}{64n^6} - \frac{1215m^3}{32n^4} + \frac{m^4}{64n^6} \cdot \mathbf{Pr}_{C\sim\mathcal{D}}[SAT(C)]$.*

Proof. We prove the lemma with the following derivation:

$$\begin{aligned}
\mathop{\mathbf{E}}_{C\sim\mathcal{D}}[\kappa(C)] &\ge \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\kappa(\widetilde{C})\right] \\
&= \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\kappa(\widetilde{C})|SAT(C)\right] \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[SAT(C)] \\
&\quad + \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\kappa(\widetilde{C})|\neg SAT(C))\right] \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)] \\
&\ge \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\frac{82\widetilde{m}^4}{64n^6} - \frac{123\widetilde{m}^3}{16n^4}\middle|SAT(C)\right] \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[SAT(C)] \\
&\quad + \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\frac{81\widetilde{m}^4}{64n^6} - \frac{27\widetilde{m}^3}{8n^4}\middle|\neg SAT(C)\right] \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)] \\
&\ge \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\frac{82\widetilde{m}^4}{64n^6} - \frac{123\widetilde{m}^3}{16n^4}\right] - \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\frac{\widetilde{m}^4}{64n^6}\middle|\neg SAT(C)\right] \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)] \\
&\ge \frac{82}{64n^6}\mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\widetilde{m}^4\right] - \frac{123m^3}{16n^4} - \frac{m^4}{64n^6} \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)] \\
&\ge \frac{82m^4}{64n^6} - \frac{861m^3}{32n^4} - \frac{123m^3}{16n^4} - \frac{m^4}{64n^6} \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)]
\end{aligned}$$

$$= \frac{81m^4}{64n^6} - \frac{1107m^3}{32n^4} + \frac{m^4}{64n^6} \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[SAT(C)],$$

as desired. The second inequality follows by Lemmas 12 and 11, the fourth one by the fact $\widetilde{m} \le m$, and the fifth one by Lemma 10. ◀

Now, Lemmas 9 and 13 yield

$$\frac{81m^4}{64n^6} - \frac{1215m^3}{32n^4} + \frac{m^4}{64n^6} \cdot \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[SAT(C)] \le \mathop{\mathbf{E}}_{C\sim\mathcal{D}}[\kappa(C)] \le \frac{81m^4}{64n^6} + \frac{729m^3}{32n^4},$$

which implies the desired bound $\mathbf{Pr}_{C\sim\mathcal{D}}[SAT(C)] \le \frac{4288\cdot n^2}{m}$. Theorem 8 follows. ◀

6 Bounds on the lower satisfiability threshold

In this section, we present our upper bound on the lower satisfiability threshold and explain why we believe that it is the tight bound for every degree of independence.

▶ **Theorem 14.** *For every integer $k \ge 2$, $\mathsf{LST}_k(n) = \Omega(n^{1-1/k})$.*

Proof. We prove the theorem by showing that for every integers $n \ge 3$ and $m \le \frac{1}{12} \cdot n^{1-1/k}$, any k-clause independent distribution $\mathcal{D} \in \mathcal{F}_k(n,m)$ satisfies $\mathbf{Pr}_{C\sim\mathcal{D}}[\neg SAT(C)] \le 1/3$.

Let $G(C) = (V, E)$ be a 3-uniform hypergraph defined by an instance C drawn from a k-clause independent distribution $\mathcal{D}$, $V = X(n)$ and $E = \{\mathrm{Var}(c)\}_{c\in C}$, i.e., every node in G represents a variable and every hyperegde in G represents the variable triplet of a clause. We consider simple *Berge-cycles* (or, simply, cycles) of length $\ell \ge 2$. A cycle of length $\ell > 2$ is defined as a set of ℓ distinct hyperedges $e_1, e_2, \ldots, e_\ell \in E$ such that pairs of consecutive edges share exactly one vertex ($|e_i \cap e_{i+1}| = 1$ and $|e_\ell \cap e_1| = 1$) and all other pairs of hyperedges are disjoint. In a cycle of length $\ell = 2$, the two hyperedges e_1 and e_2 have at least two vertices in common. We similarly define a path of length $\ell \ge 2$, where pairs of consecutive edges share exactly one vertex ($|e_i \cap e_{i+1}| = 1$ for $i \in [\ell-1]$) and all other pairs of hyperedges are disjoint. We observe that any unsatisfiable instance must contain a subgraph H in G such that every variable in H appears in at least two hyperedges of H, which in turn implies that G has a cycle. That translates into the following upper bound on the probability that instance $C \sim \mathcal{D}$ is not satisfiable:

$$\begin{aligned}\mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\neg SAT(C)] \le \mathop{\mathbf{Pr}}_{C\sim\mathcal{D}}[\exists \text{ cycle in } G(C)] &\le \sum_{\ell\ge 2} \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\mathsf{Cycle}_\ell(G(C))\right] \\ &\le \mathop{\mathbf{E}}_{C\sim\mathcal{D}}[\mathsf{Path}_k(G(C))] + \sum_{\ell=2}^{k-1} \mathop{\mathbf{E}}_{C\sim\mathcal{D}}\left[\mathsf{Cycle}_\ell(G(C))\right] \\ &= \mathop{\mathbf{E}}_{C\sim\mathcal{D}_{\mathrm{Ind.}}}[\mathsf{Path}_k(G(C))] + \sum_{\ell=2}^{k-1} \mathop{\mathbf{E}}_{C\sim\mathcal{D}_{\mathrm{Ind.}}}\left[\mathsf{Cycle}_\ell(G(C))\right]. \end{aligned} \tag{7}$$

In the above derivation, $\mathsf{Cycle}_\ell(G(C))$ counts the number of distinct cycles of length ℓ in G, and $\mathsf{Path}_k(G(C))$ counts the number of distinct paths of lengths k in G. For each ordered tuple of k hyperedges in G (there are $k! \cdot \binom{m}{k}$ such orderings), we can easily calculate the probability that they form a path. Thus,

$$\mathop{\mathbf{E}}_{C\sim\mathcal{D}_{\mathrm{Ind.}}}[\mathsf{Path}_k(G(C))] = k! \cdot \binom{m}{k} \cdot \frac{3\cdot\binom{n-3}{2}}{\binom{n}{3}} \cdot \prod_{t=3}^{k} \frac{2\cdot\binom{n-2t+1}{2}}{\binom{n}{3}} \le m^k \cdot \frac{9}{n} \cdot \left(\frac{6}{n}\right)^{k-2}. \tag{8}$$

For each ordering of k edges, the first edge e_1 can be chosen arbitrarily; the second edge has exactly one vertex in common with e_1 and the remaining two vertices are chosen from $[n]\setminus e_1$ vertices; each of the remaining edges e_t for $t\geq 3$ has exactly one of the two vertices of e_{t-1} different from $e_{t-1}\cap e_{t-2}$ and has two other vertices chosen from $[n]\setminus(e_1\cup\ldots\cup e_{t-1})$. The upper bound follows after simplifying each of the terms. We similarly derive an upper bound on $\mathbf{E}[\mathsf{Cycle}_\ell(G(C))]$ for $\ell\geq 3$ as follows:

$$\underset{C\sim\mathcal{D}_{\text{Ind.}}}{\mathbf{E}}\left[\mathsf{Cycle}_\ell(G(C))\right]\leq\frac{1}{6}\cdot\ell!\cdot\binom{m}{\ell}\cdot\left[\frac{3\cdot\binom{n-3}{2}}{\binom{n}{3}}\cdot\prod_{t=3}^{\ell-1}\frac{2\cdot\binom{n-2t+1}{2}}{\binom{n}{3}}\right]\cdot\frac{4\cdot(n-2\ell+1)}{\binom{n}{3}},$$

where for each ordering of ℓ edges (now $2\ell\geq 6$ orderings correspond to the same cycle), the probabilities to choose the first $\ell-1$ edges can be computed in the same way as for Path_k; for the last hyper-edge e_ℓ, there are $4=2\cdot 2$ ways to select a vertex of $e_{\ell-1}$ together with a vertex of e_1, and $n-2\ell+1$ choices for the new vertex in $[n]\setminus(e_1\cup\ldots\cup e_{\ell-1})$. Hence,

$$\underset{C\sim\mathcal{D}_{\text{Ind.}}}{\mathbf{E}}\left[\mathsf{Cycle}_\ell(G(C))\right]\leq m^\ell\cdot\left[\frac{9}{n}\cdot\left(\frac{6}{n}\right)^{\ell-3}\right]\cdot\frac{4}{n^2}=\frac{1}{6}\left(\frac{6m}{n}\right)^\ell. \tag{9}$$

For $\ell=2$ we have

$$\underset{C\sim\mathcal{D}_{\text{Ind.}}}{\mathbf{E}}\left[\mathsf{Cycle}_2(G(C))\right]=\binom{m}{2}\cdot\frac{3\cdot(n-3)+1}{\binom{n}{3}}=\frac{3m\cdot(m-1)\cdot(3n-8)}{n\cdot(n-1)\cdot(n-2)}\leq\left(\frac{3m}{n}\right)^2. \tag{10}$$

We conclude the proof by plugging estimates (8),(9), and (10) into (7).

$$\begin{aligned}\mathbf{Pr}\left[\neg SAT(C)\right]&\leq\frac{2}{3}\frac{(6m)^k}{n^{k-1}}+\left(\frac{3m}{n}\right)^2+\frac{1}{6}\sum_{\ell=3}^{k-1}\left(\frac{6m}{n}\right)^\ell\\&\leq\frac{2}{3\cdot 2^k}+\frac{1}{16}+\frac{1}{6}\cdot\left(\frac{1}{2^3}+\frac{1}{2^4}+\ldots+\frac{1}{2^{k-1}}\right)<\frac{1}{3}.\end{aligned}$$

The second inequality follows since $m\leq\frac{1}{12}\cdot n^{1-1/k}$. ◀

6.1 On the tightness of the $O(n^{1-1/k})$ bound for $\mathtt{LST}_k(n)$

Theorem 14 says that k-wise independence of $\mathcal{D}\in\mathcal{F}_k(n,m)$ is enough to guarantee satisfiability of a random formula $C\sim\mathcal{D}$ for the number of clauses m of order $n^{1-1/k}$. On the other hand, for the mutually independent distribution $\mathcal{D}_{\text{Ind.}}$ a random formula is unsatisfiable with high probability for $m=O(n)$. Furthermore, our analysis in Theorem 14 is essentially tight and it seems unlikely that there is a better bound than $m=O(n^{1-1/k})$. We discuss below why this is the case, by outlining a plausible way for constructing k-clause independent distribution $\mathcal{D}\in\mathcal{F}_k(n,m)$ with $\mathbf{Pr}_{C\sim\mathcal{D}}[\neg SAT(C)]\geq\frac{2}{3}$ and $m=\Theta(n^{1-1/k})$.

Informal outline of the construction

In our construction of the k-clause independent distribution $\mathcal{D}$ we are only concerned with the distribution of clauses over variables, as all literals will be assigned uniformly at random and independently. Then, a sufficient condition for a random formula $C\sim\mathcal{D}$ to be unsatisfiable with large probability is that the 3-uniform hypergraph $G(C)$ constructed in Theorem 14 has a dense subgraph, i.e., a subgraph on $|V(H)|$ vertices (corresponding to variables in C) with at least $|E(H)|\geq 100\cdot|V(H)|$ hyperedges. Indeed, one can simply count the expected number of satisfying assignments of a random formula ϕ_H on $V(H)$ variables and

$|E(H)|$ fixed clauses with randomly assigned literals: initially, all $2^{|V(H)|}$ assignment are satisfying, but then, as we add $|E(H)|$ clauses one-by-one, the expected number of satisfying assignments reduces each time exactly by a factor 7/8 (each satisfying assignment disappears with probability 1/8). At the end, we get that the expected number of satisfying assignments is $\left(\frac{7}{8}\right)^{100|V(H)|} \cdot 2^{|V(H)|} < 0.01$, which means that ϕ is unsatisfiable most of the times.

Now, we want to make sure that our hypergraph $G(C)$ often has such a "dense" subgraph H, to get an unsatisfiable random formula. Notice that H needs only have a constant number of vertices. Also, note that according to the analysis in Theorem 14, we need to avoid any cycles of length smaller than or equal to k, as the probability of having such a cycle is of order $o(1)$ in $G(C)$ for $m = c \cdot n^{1-1/k}$ and any k-clause independent distribution $C \sim \mathcal{F}_k(n, m)$. Luckily, there are many construction of such 3-uniform hypergraphs H on $|V(H)| = O(1)$ vertices with large number of hyperedges $|E(H)| \geq 100|V(H)|$, and also of large girth $g(H) \geq k + 1$ (e.g., see [23]). We shall "plant" H (essentially insert H as a connected component into G) with large probability in our construction. However, we need to make sure that by inserting H into our graph G, we still have enough room to match $\mathbf{Pr}[c_i = t_i, \forall i \in S] = \prod_{i \in S} \mathbf{Pr}[c_i = t_i]$ for all $S : |S| \leq k$. Next, we give a high level idea how one could achieve this.

As usual for the construction of distributions with identical marginals, we symmetrize $\mathcal{D}$ over all possible permutations of variables in $C \sim \mathcal{D}$. Then, our goal is to match the expected numbers for each isomorphism class of configurations of k hyperedges in $G(C)$ for $C \sim \mathcal{D}$ with $C \sim \mathcal{D}_{\text{Ind.}}$. It is useful to take note of the structure of hypergraph $G(C)$ for $C \sim \mathcal{D}_{\text{Ind.}}$ when $m = c \cdot n^{1-1/k}$: it consists of connected components, each of which is a tree of size at most k; the number of connected components of size k is a constant that grows slightly faster than linearly in c and, more generally, the number of connected components of size $k - j$ is $\Theta(n^{j/k})$. There is also a $o(1)$ probability event of having a Berge-cycle or a connected component of size at least $k + 1$ in $G(C)$ for $C \sim \mathcal{D}_{\text{Ind.}}$. We can ignore in our construction of $\mathcal{D}$ those events, by adding small probability mass to $\mathcal{D}$ that consists of $\mathcal{D}_{\text{Ind.}}$ conditional on any of these rare events (can be easily achieved via rejection sampling from $\mathcal{D}_{\text{Ind.}}$). In this way, we only need to worry about matching expected numbers of forest configurations of size k in $G(C)$ between $C \sim \mathcal{D}$ and $C \sim \mathcal{D}_{\text{Ind.}}$. We can pick the constant in $m = c \cdot n^{1-1/k}$ sufficiently large, so that the constant size subgraph H contributes fewer trees of each type than their respective expected numbers in $G(C)$ for $C \sim \mathcal{D}_{\text{Ind.}}$. Then, we can add a few more connected components that are trees of size k to $C \sim \mathcal{D}$, so that we match k-wise statistics on all tree configurations with $C \sim \mathcal{D}_{\text{Ind.}}$. By having H and a few trees of size k in the graph $G(C)$ for $C \sim \mathcal{D}$, we only utilize a constant number of variables. Then, we would like to keep adding smaller connected components that consist of trees with strictly less than k hyperedges and eventually match k-wise statistics on all forest-like configurations of size k.

It seems plausible that the approach outlined above should work and yield a k-clause independent distribution $\mathcal{D}$. Apart from heavy notation that would be needed to formalize all steps in the above outline, the main technical hurdle is to ensure that statistics for all "forest-like" configurations of k hyperedges with more than one connected component are perfectly matched with $G(C)$ for $C \sim \mathcal{D}_{\text{Ind.}}$. Note, however, that even if our approach fails and there is a stronger version of Theorem 14 with $m = \omega(n^{1-1/k})$ and $\mathbf{Pr}_{C \sim \mathcal{D}}[SAT(C)] \geq \frac{2}{3}$ for any k-clause independent distribution $\mathcal{D}$, such a theorem would require a rather nontrivial argument that relies on subtle dependencies between multiple k-wise statistics.

References

1 Dimitris Achlioptas. Setting 2 variables at a time yields a new lower bound for random 3-SAT. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 28–37, 2000. doi:10.1145/335305.335309.

2 Dimitris Achlioptas. Random satisfiabiliy. In *Handbook of Satisfiability*, volume 336 of *Frontiers in Artificial Intelligence and Applications*, pages 437–462. IOS Press, 2nd edition, 2021. doi:10.3233/FAIA200993.

3 Dimitris Achlioptas and Yuval Peres. The threshold for random k-sat is $2klog2 - o(k)$. *Journal of the American Mathematical Society*, 17:947–973, 2004. URL: http://www.jstor.org/stable/20161221.

4 Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 496–505, 2007. doi:10.1145/1250790.1250863.

5 Noga Alon and Asaf Nussboim. k-wise independent random graphs. In *Proceedings of 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 813–822, 2008. doi:10.1109/FOCS.2008.61.

6 Carlos Ansótegui, Jesús Giráldez-Cru, and Jordi Levy. The community structure of SAT formulas. In *Proceedings of 15th International Conference on Theory and Applications of Satisfiability Testing (SAT)*, pages 410–423, 2012. doi:10.1007/978-3-642-31612-8_31.

7 Paul Beame, Richard Karp, Toniann Pitassi, and Michael Saks. The efficiency of resolution and davis–putnam procedures. *SIAM Journal on Computing*, 31(4):1048–1075, April 2002. doi:10.1137/S0097539700369156.

8 Eli Ben-Sasson and Avi Wigderson. Short proofs are narrow—resolution made simple. *Journal of the ACM*, 48(2):149–169, 2001. doi:10.1145/375827.375835.

9 Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled. On k-wise independent distributions and boolean functions. *arXiv:math*, abs/1201.3261, 2012. doi:10.48550/arXiv.1201.3261.

10 Béla Bollobás, Christian Borgs, Jennifer T. Chayes, Jeong Han Kim, and David Bruce Wilson. The scaling window of the 2-SAT transition. *Random Structures and Algorithms*, 18(3):201–256, 2001. doi:10.1002/RSA.1006.

11 Karl Bringmann. Why walking the dog takes time: Frechet distance has no strongly subquadratic algorithms unless SETH fails. In *Proceedings of 55th IEEE Annual Symposium on Foundations of Computer Science, (FOCS)*, pages 661–670, 2014. doi:10.1109/FOCS.2014.76.

12 Andrei Z. Broder, Alan M. Frieze, and Eli Upfal. On the satisfiability and maximum satisfiability of random 3-CNF formulas. In *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 322–330, 1993. URL: http://dl.acm.org/citation.cfm?id=313559.313794.

13 Ioannis Caragiannis, Nick Gravin, Pinyan Lu, and Zihe Wang. Relaxing the independence assumption in sequential posted pricing, prophet inequality, and random bipartite matching. In *Proceedings of 17th International Conference on Web and Internet Economics (WINE)*, pages 131–148, 2021. doi:10.1007/978-3-030-94676-0_8.

14 Peter C. Cheeseman, Bob Kanefsky, and William M. Taylor. Where the really hard problems are. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 331–340, 1991. URL: http://ijcai.org/Proceedings/91-1/Papers/052.pdf.

15 Vasek Chvátal and Bruce A. Reed. Mick gets some (the odds are on his side). In *Proceedings of 33rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 620–627, 1992. doi:10.1109/SFCS.1992.267789.

16 Vašek Chvátal and Endre Szemerédi. Many hard examples for resolution. *Journal of the ACM*, 35(4):759–768, 1988. doi:10.1145/48014.48016.

17 Amin Coja-Oghlan. The asymptotic k-sat threshold. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 804–813, 2014. doi:10.1145/2591796.2591822.

18 Marek Cygan, Jesper Nederlof, Marcin Pilipczuk, Michal Pilipczuk, Johan M. M. van Rooij, and Jakub Onufry Wojtaszczyk. Solving connectivity problems parameterized by treewidth in single exponential time. *ACM Transactions on Algorithms*, 18(2):17:1–17:31, 2022. doi:10.1145/3506707.

19 Josep Díaz, Lefteris M. Kirousis, Dieter Mitsche, and Xavier Pérez-Giménez. On the satisfiability threshold of formulas with three literals per clause. *Theoretical Computer Science*, 410(30-32):2920–2934, 2009. doi:10.1016/J.TCS.2009.02.020.

20 Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. *Annals of Mathematics*, 196(1):1–388, 2022. doi:10.4007/annals.2022.196.1.1.

21 Olivier Dubois, Yacine Boufkhad, and Jacques Mandler. Typical random 3-SAT formulae and the satisfiability threshold. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 126–127, 2000. URL: http://dl.acm.org/citation.cfm?id=338219.338243.

22 Shaddin Dughmi, Yusuf Hakan Kalayci, and Neel Patel. Limitations of stochastic selection problems with pairwise independent priors. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 479–490, 2024. doi:10.1145/3618260.3649718.

23 David Ellis and Nathan Linial. On regular hypergraphs of high girth. *The Electronic Journal of Combinatorics*, 21(1):1, 2014. doi:10.37236/3851.

24 Uriel Feige. Relations between average case complexity and approximation complexity. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 534–543, 2002. doi:10.1145/509907.509985.

25 Uriel Feige and Eran Ofek. Easily refutable subformulas of large random 3cnf formulas. *Theory of Computing*, 3(1):25–43, 2007. doi:10.4086/TOC.2007.V003A002.

26 Abraham Flaxman. Random planted 3-sat. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*, pages 1728–1732. Springer New York, New York, NY, 2016. doi:10.1007/978-1-4939-2864-4_330.

27 John Franco and Marvin C. Paull. Probabilistic analysis of the davis putnam procedure for solving the satisfiability problem. *Discrete Applied Mathematics*, 5(1):77–87, 1983. doi:10.1016/0166-218X(87)90032-1.

28 Joel Friedman, Andreas Goerdt, and Michael Krivelevich. Recognizing more unsatisfiable random k-SAT instances efficiently. *SIAM Journal on Computing*, 35(2):408–430, 2005. doi:10.1137/S009753970444096X.

29 Tobias Friedrich, Anton Krohmer, Ralf Rothenberger, Thomas Sauerwald, and Andrew M. Sutton. Bounds on the satisfiability threshold for power law distributed random SAT. In *Proceedings of 25th Annual European Symposium on Algorithms (ESA)*, pages 37:1–37:15, 2017. doi:10.4230/LIPICS.ESA.2017.37.

30 Tobias Friedrich and Ralf Rothenberger. Sharpness of the satisfiability threshold for non-uniform random k-SAT. In *Proceedings of 21st International Conference on Theory and Applications of Satisfiability Testing (SAT)*, pages 273–291, 2018. doi:10.1007/978-3-319-94144-8_17.

31 Tobias Friedrich and Ralf Rothenberger. The satisfiability threshold for non-uniform random 2-SAT. In *Proceedings of 46th International Colloquium on Automata, Languages, and Programming, (ICALP)*, pages 61:1–61:14, 2019. doi:10.4230/LIPICS.ICALP.2019.61.

32 Alan M. Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of k-SAT. *Journal of Algorithms*, 20(2):312–355, 1996. doi:10.1006/JAGM.1996.0016.

33 Nick Gravin and Zhiqi Wang. On robustness to k-wise independence of optimal bayesian mechanisms. In *Proceedings of the 65th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1275–1293, 2024. doi:10.1109/FOCS61266.2024.00084.

34 Anupam Gupta, Jinqiao Hu, Gregory Kehne, and Roie Levin. Pairwise-independent contention resolution. In *Proceddings of 25th International Conference on Integer Programming and Combinatorial Optimization (IPCO)*, pages 196–209, 2024. doi:10.1007/978-3-031-59835-7_15.

35 Mohammad Taghi Hajiaghayi and Gregory B Sorkin. The satisfiability threshold of random 3-SAT is at least 3.52. *arXiv:math*, abs/0310193, 2003. doi:10.48550/arXiv.math/0310193.

36 Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *Journal of Computer and System Sciences*, 62(2):367–375, 2001. doi:10.1006/JCSS.2000.1727.

37 Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *Journal of Computer and System Sciences*, 63(4):512–530, 2001. doi:10.1006/JCSS.2001.1774.

38 Anil Kamath, Rajeev Motwani, Krishna V. Palem, and Paul G. Spirakis. Tail bounds for occupancy and the satisfiability threshold conjecture. *Random Structures and Algorithms*, 7(1):59–80, 1995. doi:10.1002/RSA.3240070105.

39 Alexis C. Kaporis, Lefteris M. Kirousis, and Efthimios G. Lalas. The probabilistic analysis of a greedy satisfiability algorithm. *Random Structures and Algorithms*, 28(4):444–480, 2006. doi:10.1002/RSA.20104.

40 Marc Mezard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002. doi:10.1126/science.1073287.

41 David G. Mitchell, Bart Selman, and Hector J. Levesque. Hard and easy distributions of SAT problems. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI)*, pages 459–465, 1992. URL: http://www.aaai.org/Library/AAAI/1992/aaai92-071.php.

42 Oleksii Omelchenko and Andrei A. Bulatov. Satisfiability threshold for power law random 2-sat in configuration model. *Theoretical Computer Science*, 888:70–94, 2021. doi:10.1016/J.TCS.2021.07.028.

43 Ronitt Rubinfeld and Ning Xie. Testing non-uniform k-wise independent distributions over product spaces. In *Proceedings of 37th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 565–581, 2010. doi:10.1007/978-3-642-14165-2_48.

44 Michael Sipser. *Introduction to the Theory of Computation.* Cengage Learning, 3rd edition, 2013.