



# Satisfiability Thresholds for Non-Uniform Random k-SAT

**Ralf Rothenberger**

Universitätsdissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium
*(Dr. rer. nat.)*

in der Wissenschaftsdisziplin
Theoretische Informatik

eingereicht an der
Digital-Engineering-Fakultät
der Universität Potsdam

**Datum der Disputation:** 6. April 2022

## Betreuer

**Prof. Dr. Tobias Friedrich**
Hasso Plattner Institute, University of Potsdam

## Gutachter

**Prof. Dr. Amin Coja-Oghlan**
Goethe University Frankfurt

**Prof. Dr. Samuel R. Buss**
University of California, San Diego

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-54970
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-549702

# Abstract

Boolean Satisfiability (SAT) is one of the problems at the core of theoretical computer science. It was the first problem proven to be NP-complete by Cook and, independently, by Levin. Nowadays it is conjectured that SAT cannot be solved in sub-exponential time. Thus, it is generally assumed that SAT and its restricted version $k$-SAT are hard to solve. However, state-of-the-art SAT solvers can solve even huge practical instances of these problems in a reasonable amount of time.

Why is SAT hard in theory, but easy in practice? One approach to answering this question is investigating the average runtime of SAT. In order to analyze this average runtime the random $k$-SAT model was introduced. The model generates all $k$-SAT instances with $n$ variables and $m$ clauses with uniform probability. Researching random $k$-SAT led to a multitude of insights and tools for analyzing random structures in general. One major observation was the emergence of the so-called *satisfiability threshold*: A phase transition point in the number of clauses at which the generated formulas go from asymptotically almost surely (a. a. s. ) satisfiable to a. a. s. unsatisfiable. Additionally, instances around the threshold seem to be particularly hard to solve.

In this thesis we analyze a more general model of random $k$-SAT that we call *non-uniform random $k$-SAT*. In contrast to the classical model each of the $n$ Boolean variables now has a distinct probability of being drawn. For each of the $m$ clauses we draw $k$ variables according to the variable distribution and choose their signs uniformly at random. Non-uniform random $k$-SAT gives us more control over the distribution of Boolean variables in the resulting formulas. This allows us to tailor distributions to the ones observed in practice. Notably, non-uniform random $k$-SAT contains the previously proposed models random $k$-SAT, power-law random $k$-SAT and geometric random $k$-SAT as special cases.

We analyze the satisfiability threshold in non-uniform random $k$-SAT depending on the variable probability distribution. Our goal is to derive conditions on this distribution under which an equivalent of the satisfiability threshold conjecture holds. We start with the arguably simpler case of non-uniform random 2-SAT. For this model we show under which conditions a threshold exists, if it is sharp or coarse, and what the leading constant of the threshold function is. These are exactly the three ingredients one needs in order to prove or disprove the satisfiability threshold conjecture. For non-uniform random $k$-SAT with $k \geq 3$ we only prove sufficient conditions under which a threshold exists. We also show some properties of the variable probabilities under which the threshold is sharp in this case. These are the first results on the threshold behavior of non-uniform random $k$-SAT.

# Zusammenfassung

Das Boolesche Erfüllbarkeitsproblem (SAT) ist eines der zentralsten Probleme der theoretischen Informatik. Es war das erste Problem, dessen NP-Vollständigkeit nachgewiesen wurde, von Cook und Levin unabhängig voneinander. Heutzutage wird vermutet, dass SAT nicht in subexponentialler Zeit gelöst werden kann. Darum wird allgemein angenommen, dass SAT und seine eingeschränkte Version $k$-SAT nicht effizient zu lösen sind. Trotzdem können moderne SAT solver sogar riesige Echtweltinstanzen dieser Probleme in angemessener Zeit lösen.

Warum ist SAT theoretisch schwer, aber einfach in der Praxis? Ein Ansatz um diese Frage zu beantworten ist die Untersuchung der durchschnittlichen Laufzeit von SAT. Um diese durchschnittliche oder typische Laufzeit analysieren zu können, wurde zufälliges $k$-SAT eingeführt. Dieses Modell erzeugt all $k$-SAT-Instanzen mit $n$ Variablen und $m$ Klauseln mit gleicher Wahrscheinlichkeit. Die Untersuchung des Zufallsmodells für $k$-SAT führte zu einer Vielzahl von Erkenntnissen und Techniken zur Untersuchung zufälliger Strukturen im Allgemeinen. Eine der größten Entdeckungen in diesem Zusammenhang war das Auftreten des sogenannten *Erfüllbarkeitsschwellwerts*: Ein Phasenübergang in der Anzahl der Klauseln, an dem die generierten Formeln von asymptotisch sicher erfüllbar zu asymptotisch sicher unerfüllbar wechseln. Zusätzlich scheinen Instanzen, die um diesen Übergang herum erzeugt werden, besonders schwer zu lösen zu sein.

In dieser Arbeit analysieren wir ein allgemeineres Zufallsmodell für $k$-SAT, das wir *nichtuniformes zufälliges k-SAT* nennen. Im Gegensatz zum klassischen Modell, hat jede Boolesche Variable jetzt eine bestimmte Wahrscheinlichkeit gezogen zu werden. Für jede der $m$ Klauseln ziehen wir $k$ Variablen entsprechend ihrer Wahrscheinlichkeitsverteilung und wählen ihre Vorzeichen uniform zufällig. Nichtuniformes zufälliges $k$-SAT gibt uns mehr Kontrolle über die Verteilung Boolescher Variablen in den resultierenden Formeln. Das erlaubt uns diese Verteilungen auf die in der Praxis beobachteten zuzuschneiden. Insbesondere enthält nichtuniformes zufälliges $k$-SAT die zuvor vorgestellten Modelle zufälliges $k$-SAT, skalenfreies zufälliges $k$-SAT und geometrisches zufälliges $k$-SAT als Spezialfälle.

Wir analysieren den Erfüllbarkeitsschwellwert in nichtuniformem zufälligen $k$-SAT abhängig von den Wahrscheinlichkeitsverteilungen für Variablen. Unser Ziel ist es, Bedingungen an diese Verteilungen abzuleiten, unter denen ein Äquivalent der Erfüllbarkeitsschwellwertsvermutung für zufälliges $k$-SAT gilt. Wir fangen mit dem wahrscheinlich einfacheren Modell nichtuniformem zufälligen 2-SAT an. Für dieses Modell zeigen wir, unter welchen Bedingungen ein Schwellwert existiert, ob er steil oder flach ansteigt und was die führende Konstante der Schwellwertfunktion ist. Das sind genau die Zutaten, die man

benötigt um die Erfüllbarkeitsschwellwertsvermutung zu bestätigen oder zu widerlegen. Für nichtuniformes zufälliges $k$-SAT mit $k \geq 3$ zeigen wir nur hinreichende Bedingungen, unter denen ein Schwellwert existiert. Wir zeigen außerdem einige Eigenschaften der Variablenwahrscheinlichkeiten, die dazu führen, dass der Schwellwert steil ansteigt. Dies sind unseres Wissens nach die ersten allgemeinen Resultate zum Schwellwertverhalten von nichtuniformem zufälligen $k$-SAT.

# Acknowledgments

The journey towards my PhD was a long and sometimes strenuous one. I experienced the highs and lows of succeeding and failing at theorem proving, the joy and dread of having publications accepted and declined, and the ups and downs of teaching students and working with others. I learned a lot and matured as a scientist and a person. And the good times far outweighed the bad ones. So much so that I was not in a hurry to finish my PhD. But all good things come to an end and it is time for me to take the next step in academia. And I'd like to thank the work group and the many people around me, who always supported me on the way to that next step.

The person I owe most to on the way to my PhD is my supervisor Tobias Friedrich. He supported me for more than seven years and taught me many things about being a good scientist and succeeding in academia. He always had an open ear for my problems and time to discuss. He gave me time to work at my own pace and to follow my own ideas. And he pushed me to pursue high-ranking conferences and ambitious results. I see him as a role model and thank him a lot.

I'd also like to thank Timo Kötzing for always being a mentor to anyone in our group. He always listened to anyone's problems and offered good advice. Timo also hosted regular game nights at his place which I enjoyed greatly.

Another great thank you goes out to the many current and former members of the Algorithm Engineering group. The group made me feel at home and created a great working environment. There were always people around to have a chat, play board games after lunch, and of course discuss problems and ideas. A special thank you goes to Anton Krohmer and Martin Krejca, for going much of the way to a PhD with me, for always helping me with problems, and for many great nights at the game club. Another special thank you goes to Thomas Bläsius and Ralf Teusner, with whom I spent many hours of free-time, mostly playing board and role-playing games, and who also always supported me. Thank you to Martin Schirneck, Maximilian Katzmann, Vanja Doskoč, and Ziena Zeif for being great office mates! I'd also like to thank Katrin Heinrich for taking care of all the small and big tasks around the office and for guiding me through the bureaucracy of the HPI. Also for all the nice chats and kind words. Apart from the people at the Algorithm Engineering group I would like to thank my reviewers Amin Coja-Oghlan and Sam Buss for taking the time to read this thesis.

Last but not least I would like to thank my family and friends. My parents, Achim Rothenberger and Regina Rothenberger, supported me all my life and could not wait for me to finish my PhD. They were also the ones who pointed me to the position at Tobias Friedrich's group. I am eternally grateful for their

ongoing support. Thank you to my girlfriend Laura Renau for caring for me and motivating me. And for being there for me and bringing joy to my life even when times get tough. One last thank you goes out to Daniel Korinek, Fabian Kittler, and Richard Sieder for being great friends!

# Contents

# 1 Introduction

Boolean Satisfiability (SAT) is one of the problems at the core of theoretical computer science. It was the first problem proven to be NP-complete by Stephen A. Cook in 1971 [Coo71] and, independently, by Leonid Levin in 1973 [Lev73]. The notion of NP-hardness gave rise to a whole class of NP-complete problems via polynomial-time reductions from SAT [GJ79; Kar72]. Answering the question whether SAT or any other NP-complete problem can be solved in polynomial time is still one of the most important open questions in theoretical computer science.

Despite decades of research, no polynomial time algorithm for solving SAT or its restricted version $k$-SAT has been found. In fact, no algorithm has been found that substantially improves upon a running time of $\Theta(2^n)$ for SAT or $\Theta(2^{c \cdot n})$ for $k$-SAT, where $c > 0$ is a small constant. This caused the proposal of another conjecture called the Exponential Time Hypothesis (ETH) [IP99], which claims that 3-SAT can only be solved in time $\Omega(2^{s_3 \cdot n})$ for some constant $s_3 > 0$. A slightly stronger version, the Strong Exponential Time Hypothesis (SETH) [CIP09], claims that SAT cannot be solved in time $O(2^{s \cdot n})$ for a constant $s < 1$. The assumptions of ETH and SETH, just like assuming P$\neq$ NP, can be used to derive lower bounds on the time to solve other problems, even such problems that are solvable in polynomial time [LMS11]. Thus, it is generally assumed that SAT and $k$-SAT are hard to solve. However, state-of-the-art SAT solvers can solve huge instances with millions of variables that stem from practical problems (so called *industrial instances*) in a reasonable amount of time. This begs the question: Why is SAT hard in theory, but easy in practice?

There are several approaches that try to explain the gap between the theoretical and practical hardness of SAT. One possible explanation is that SAT is generally easy except for a small core of hard-to-solve instances. In order to see if this can be the case, one can imagine the following experiment: Draw uniformly at random from all possible $k$-SAT instances with $n$ variables and $m$ clauses and see how hard to solve they are on average. If the number of hard-to-solve instances was sufficiently small, the average hardness of SAT would still be small as well. This approach of drawing formulas in $k$-CNF uniformly is known as *random $k$-SAT* and has been studied extensively. The analysis of random $k$-SAT resulted in a multitude of insights and new techniques for the analysis of Boolean Satisfiability and random structures in general.

It was observed that for random $k$-SAT there is a phase transition in the number of clauses $m$, where instances transition from being satisfiable with probability tending to one to being satisfiable with probability tending to zero. The number of clauses at which this happens is called the *satisfiability threshold*.

Proving the existence, behavior, and exact position of the satisfiability threshold rigorously turned out to be a challenging task for $k \geq 3$. It has been and still is the subject of many theoretical works.

The position and behavior of the satisfiability threshold also seem to have influence on the running time of SAT solvers. Mitchell et al. [MSL92] found that the median running time of DPLL scales exponentially for random $k$-SAT instances generated around or slightly above the threshold. As it turned out, this is due to a deep-seeded connection between the DPLL algorithm and resolution proofs. The resolution proof system is a refutation technique for propositional and first-order logic. The technique only uses a single rule to resolve two clauses into a resolvent clause. If a number of resolution steps on the original or derived clauses yields a contradiction, the formula is unsatisfiable. The sequence of clauses in these steps can then be used as a certificate of unsatisfiability. The minimum number of steps necessary to arrive at a contradiction is called the *resolution size* of the formula. The resolution rule was introduced by Davis and Putnam, who also introduced the DP [DP60] (later DPLL) algorithm for SAT solving as an application of the resolution proof system. The algorithm was later extended to the conflict-driven clause learning algorithm (CDCL) [JS97; SS96], which is the basis for many state-of-the art SAT solvers. Thus, there is a direct connection between resolution, DPLL, and CDCL. More precisely, it was shown that DPLL is polynomially equivalent to tree-like resolution [Bee06] and CDCL with unlimited restarts is polynomially equivalent to resolution [BS14; PD11]. Therefore, lower bounds on the (treelike) resolution size directly translate to lower bounds on the running time of (DPLL) CDCL. Due to Chvátal and Szemerédi [CS88] with probability approaching one the treelike and general resolution size of random $k$-SAT instances generated around the satisfiability threshold is exponential. Thus, both DPLL and CDCL need exponential time on random $k$-SAT instances generated around the threshold. This means, for state-of-the-art solvers the average $k$-SAT instance is still hard. This indicates that our assumption that SAT is easy except for a small core of hard-to-solve instances might be wrong.

Another possible explanation for the observed discrepancy between theory and practice is that SAT is generally hard, but industrial SAT instances form a class, whose properties make them easier to solve. Ansótegui et al. [AGL12] found that industrial SAT instances exhibit unusually high community structure, i. e. there are variables that tend to appear together in clauses. They also found [ABL09a; ABL09b; Ans+15] that in some families of industrial instances the frequencies of variables follow a power law, i. e. the fraction of variables that appear $i$ times is proportional to $i^{-\beta}$ for some constant $\beta$. It is still an open question which properties exactly some classes of practical SAT instances have in common. However, if we assume the ones we know or suspect, we can concentrate on instances with those properties. Again, there are several avenues to pursue this idea. Either we define a class of instances with the suspected proper-

ties or we consider random models generating instances with those properties to analyze their average-case behavior.

## 1.1 Scope of this Thesis

In this thesis we want to study random $k$-SAT models with given expected distributions of variable frequencies. We want to analyze how these distributions influence the satisfiability threshold and the hardness of solving instances. This would help to answer the question if the efficiency of state-of-the-art SAT solvers on industrial instances is due to the variable frequency distribution of those instances only or if it is due to other properties as well. To this end, we introduce a generalization of the random $k$-SAT model with given expected variable frequencies. We call this model *non-uniform random $k$-SAT* as it draws the Boolean variables for each clause according to a non-uniform probability distribution. This probability distribution on the Boolean variables acts as an expected frequency distribution.

We are going to analyze how the input distribution influences the position and behavior of the satisfiability threshold. More precisely, we are interested in the following questions:

1. Is there a threshold? We say that there is a satisfiability threshold, if there is a function $m^\star$ for the number of clauses such that if we draw asymptotically fewer clauses the probability to generate satisfiable instances tends to one and if we draw asymptotically more clauses, the probability to generate satisfiable instances tends to zero. This function $m^\star$ can also depend on the input parameters.

2. Is the threshold sharp? If we already know that there is a threshold at some function $m^\star$, we can investigate how steep the probability to generate satisfiable instances declines in that range. We say that the threshold is sharp if there is a range asymptotically smaller than $m^\star$ in which the probability drops from tending to one to tending to zero. If this is not the case, e. g. if the probability slowly decreases as we increase the leading constant of $m^\star$, we say that the threshold is coarse. More intuitively, a sharp threshold approaches a step function as we increase the number of variables $n$, while a coarse threshold does not.

3. What is the exact threshold position? If we know that there is a sharp threshold, we can try and find the leading constant of $m^\star$ at which the probability to generate satisfiable instances declines from tending to one to tending to zero.

The properties we are interested in were first defined for random $k$-SAT. The *satisfiability threshold conjecture* makes an assumption on how the threshold of random $k$-SAT behaves. Intuitively, the conjecture states that for every $k \geq 2$

there is a sharp threshold at some $m^\star = r^\star \cdot n$ and that the leading constant of that threshold converges to some constant $r_k$ as the number of variables increases. We want to see for which input distributions an equivalent of this conjecture holds for non-uniform random $k$-SAT.

### 1.1.1 State of the Art

There is a large body of work on random $k$-SAT, but different random SAT models have been proposed as well. In regular random $k$-SAT [BC16; Bou+05; CW18; Rat+10] instances are generated so that each variable appears at most one time more often than any other variable. In $(2+p)$-SAT [Ach+01; Mon+96; Mon+99; MZ97] instances are generated such that a $p$ fraction of clauses contain 3 literals (variables or their negation) while all others contain 2 literals. In random geometric $k$-SAT [BP14] literals are distributed uniformly in the euclidean plane and a clause is generated for each set of $k$ literals with a certain distance to each other. However, these models are not motivated by modeling the properties of industrial instances.

So, what are the properties of industrial SAT instances? At least for some families of industrial instances their properties include community structure [AGL12], i. e. certain sets of variables tend to appear together in clauses, and power-law distributed variable frequencies [ABL09a; ABL09b; Ans+15], i. e. there is a $i^{-\beta}$ fraction of variables that appear $i$ times in total. In the following we present some random SAT models that take these properties into account explicitly.

Giráldez-Cru and Levy [GL15] proposed the Community Attachment Model, which creates random formulas with clear community structure. This model has already been studied by Mull et al. [MFS16], who show that unsatisfiable instances generated by it have exponentially long resolution proofs with high probability. Thus, instances generated with the model cannot be solved fast by CDCL- and DPLL-based SAT solvers. Ansótegui et al. [ABL09b] proposed two models, *power-law random k-SAT*, which assumes a power-law distribution, and *geometric random k-SAT*, which assumes a geometric distribution. They show empirically, that instances of their models generated at the satisfiability threshold can be solved faster by state-of-the-art solvers than instances of random $k$-SAT generated at the satisfiability threshold. However, they do not show any rigorous results on the satisfiability thresholds of their models or the proof complexity of unsatisfiable instances generated by them. Recently, Giráldez-Cru and Levy [GL17] also introduced the popularity-similarity model, which incorporates both power-law degree distribution and community structure. Like almost all other models inspired by industrial instances this one lacks theoretical work regarding the satisfiability threshold.

Our thesis aims at proving properties of the satisfiability threshold for non-uniform random $k$-SAT, a generalization of the power-law random $k$-SAT and geometric random $k$-SAT models by Ansótegui et al. [ABL09b]. These properties and the satisfiability threshold conjecture were originally defined for random

$k$-SAT and there is a large body of work on them. Chvátal and Reed [CR92] and, independently, Goerdt [Goe96] proved the conjecture for $k = 2$ and showed that $r_2 = 1$. For larger values of $k$ upper and lower bounds have been established, e. g. , $3.52 \leq r_3 \leq 4.4898$ [Día+09; HS03; KKL06]. Methods from statistical mechanics [MPZ02] were used to derive a numerical estimate of $r_3 \approx 4.26$. Coja-Oghlan and Panagiotou [Coj14; CP16] showed a bound (up to lower order terms) of $r_k = 2^k \log 2 - \frac{1}{2}(1+\log 2) \pm o_k(1)$ for $k \geq 3$. Finally, Ding et al. [DSS15] proved the exact position of the threshold for sufficiently large values of $k$. Their results imply that the satisfiability threshold conjecture holds for these large values. Still, for $k$ between 3 and the values determined by Ding et al. the conjecture remains open. Except for the special case of random $k$-SAT, no rigorous results on the threshold behavior of non-uniform random $k$-SAT were known prior to our work.

## 1.2 Contribution and Outline

In this thesis we contribute to the research on random SAT models. We analyze non-uniform random $k$-SAT, a generalization of the seminal random $k$-SAT model. Non-uniform random $k$-SAT differs from random $k$-SAT by including a probability distribution over Boolean variables according to which the variables for each clause are drawn. This input probability distribution acts as an expected frequency distribution of the Boolean variables that appear in generated instances. For this model we want to answer two questions: First, how does the satisfiability threshold behave? Second, how hard is it to solve instances of the model? Answers to both questions depend on the variable probability distribution the model gets as input. Our goal is to identify how these distributions influence the behavior of the satisfiability threshold and the resolution size of generated instances. This will allow us to judge if non-uniform random $k$-SAT with "realistic" input distributions can be used to explain the behavior of state-of-the-art solvers on industrial SAT instances or if a more involved model, which captures more properties of real-world instances, will be necessary. However, due to space limitations this thesis only aims to answer the first question in detail. Results regarding the hardness and resolution size of non-uniform random $k$-SAT will only be discussed briefly in the last chapter.

Some of the chapters of this thesis are based on joint work with other researchers. In this case, we mention and highlight their contributions at the beginning of a chapter. We now give an overview over the chapters of this thesis and their contents.

In Chapter 2 we introduce the mathematical background and notation necessary for this work. As our results heavily rely on probability theory, we will introduce the stochastic tools and knowledge necessary to derive them.

In Chapter 3 we formally introduce the random $k$-SAT model and satisfiability thresholds. This especially includes the following properties and concepts

related to satisfiability thresholds: 1. asymptotic thresholds, 2. sharp and coarse thresholds, and 3. the satisfiability threshold conjecture. Afterward, we formally introduce non-uniform random $k$-SAT and show how to generalize the satisfiability threshold and related concepts to this new model. We conclude the chapter by presenting some notable special cases of non-uniform random $k$-SAT, which will serve as examples throughout this thesis.

In Chapter 4 we analyze the threshold behavior of *non-uniform random 2-SAT*. Random 2-SAT exhibits a similar threshold behavior as random $k$-SAT for $k \geq 3$, but due to the simpler structure of formulas in 2-CNF this behavior is much easier to analyze. Chvátal and Reed [CR92] derived the exact threshold position and the sharpness of the threshold for random 2-SAT. Their results proved the satisfiability threshold conjecture for $k = 2$. We use techniques similar to those of Chvátal and Reed to analyze non-uniform random 2-SAT. Depending on the input probability distribution, we derive the asymptotic threshold position, if the threshold is coarse or sharp, and, in case of a sharp threshold, the exact threshold position up to leading constant factors. This completely characterizes the behavior of the satisfiability threshold for non-uniform random 2-SAT and generalizes the results of Chvátal and Reed [CR92].

Chapter 5 is dedicated to proving the existence of satisfiability thresholds in non-uniform random $k$-SAT with $k \geq 3$. Due to the more complex nature of formulas in 3-CNF compared to those in 2-CNF this requires more involved tools and techniques. We derive a range of results that allow us to prove the existence and asymptotic position of satisfiability thresholds in non-uniform random $k$-SAT depending on the input probability distributions.

In Chapter 6 we study the sharpness of the satisfiability threshold in non-uniform random $k$-SAT with $k \geq 3$. We derive sufficient conditions for the satisfiability threshold to be sharp depending on both the input probability distribution and the asymptotic threshold position. The main result of this chapter generalizes a result from the seminal work of Friedgut [Fri99], who showed that the satisfiability threshold of random $k$-SAT is sharp, even if its exact position is not known.

Chapter 7 concludes this thesis with a discussion of the results and a compilation of open problems regarding satisfiability thresholds and resolution size of non-uniform random $k$-SAT. It also contains some of our more recent results that did not make it into the thesis.

# 2 Preliminaries

In this chapter we introduce notation, mathematical concepts, and probabilistic methods used throughout this thesis. We assume that the reader knows the basics of mathematics and probability theory. Thus, we will only introduce more advanced concepts.

## 2.1 Notation

We use blackboard bold letters to denote number sets. $\mathbb{N}$ denotes the set of natural numbers including zero and $\mathbb{R}$ denotes the set of real numbers. We let $\mathbb{R}^+$ denote the set of positive real numbers. For any $x, y \in \mathbb{R}$ with $x \leq y$ we let $[x, y] = \{z \in \mathbb{R} \mid x \leq z \leq y\}$ denote the closed interval of real numbers from $x$ to $y$. We denote open intervals with round instead of square brackets. For any $m, n \in \mathbb{N}$ we let $[m \dots n] = [m, n] \cap \mathbb{N}$ and $[n] = [1 \dots n]$. Also, we let $\mathcal{P}(\cdot)$ denote the power set and let $\mathcal{P}_k(\cdot)$ denote the set of cardinality-$k$ elements of the power set.

For a real-valued function $f$ and $c \in \mathbb{R}$ we let $\lim_{x \to c} f(x)$ denote the limit of $f$ as $x$ approaches $c$. For a sequence $a_1, a_2, \dots$ of real numbers we let $\lim_{n \to \infty} a_n$ denote the limit of $a_n$ as $n$ approaches infinity. It holds that $\lim_{n \to \infty} a_n = L$ if and only if for every real number $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ so that for all $n > n_0$ we have $|a_n - L| < \varepsilon$. Furthermore, we will use Landau notation. That means, for two real-valued functions $f$ and $g$ defined on the same unbounded subset of $\mathbb{R}^+$ we use the following notation:

- $f \in O(g) \Leftrightarrow \exists\, \varepsilon > 0\ \exists\, n_0\ \forall\, n > n_0 \colon |f(n)| \leq \varepsilon \cdot g(n)$,
- $f \in \Theta(g) \Leftrightarrow \exists\, \varepsilon_1 > 0\ \exists\, \varepsilon_2 > 0\ \exists\, n_0\ \forall\, n > n_0 \colon \varepsilon_1 \cdot g(n) \leq f(n) \leq \varepsilon_2 \cdot g(n)$,
- $f \in \Omega(g) \Leftrightarrow \exists\, \varepsilon > 0\ \exists\, n_0\ \forall\, n > n_0 \colon f(n) \geq \varepsilon \cdot g(n)$,
- $f \in o(g) \Leftrightarrow \forall\, \varepsilon > 0\ \exists\, n_0\ \forall\, n > n_0 \colon |f(n)| \leq \varepsilon \cdot g(n)$, and
- $f \in \omega(g) \Leftrightarrow \forall\, \varepsilon > 0\ \exists\, n_0\ \forall\, n > n_0 \colon |f(n)| \geq \varepsilon \cdot |g(n)|$.

The definitions of limits and Landau symbols will be used heavily when dealing with satisfiability thresholds in this thesis. Thus, it is important to state those definitions explicitly. Another definition we use to compare functions is the following. For two functions $f, g \colon X \to \mathbb{R}$ which are defined on the same domain $X$ we write $f \leq g$ iff for all $x \in X$ it holds that $f(x) \leq g(x)$.

## 2.2 Probability Theory

In this section we introduce concepts related to probability theory that we will use in this thesis. This includes probability spaces, random variables, conditional probabilities, and expected values. For a more thorough introduction to the topic, we refer to the text book by Mitzenmacher and Upfal [MU05].

### 2.2.1 Probability Spaces and Events

A *probability space* is a triple $(\Omega, \mathcal{F}, \Pr)$, where the *sample space* $\Omega$ is the set of all possible outcomes of the random process, the family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ represents all allowable *events*, and $\Pr\colon \mathcal{F} \to [0,1]$ is a *probability measure*, assigning probabilities to all events from $\mathcal{F}$. An element of $\Omega$ is also called an *elementary* event. Throughout this thesis we will mostly omit the probability space if it is clear from context.

We write $\Pr[A]$ to denote the probability of an event $A$ and we say that an event $A$ occurs *with high probability* (w. h. p.) iff $\Pr[\overline{A}] \in O(1/\mathrm{poly}(n))$ and *asymptotically almost surely* (a. a. s.) iff $\Pr[\overline{A}] \in o(1)$.

For events $A$ and $B$ with $\Pr[B] > 0$ the *conditional probability* that $A$ occurs given that event $B$ occurs is $\Pr[A \mid B] = \Pr[A \cap B]/\Pr[B]$. Essentially, we define $B$ as the new sample space and normalize all probabilities by dividing by $\Pr[B]$. In this thesis we will make use of the following theorem, which derives the probability of an event $A$ if only conditional probabilities of the event on a partitioning of the sample space $\Omega$ are known.

▶ **Theorem 2.1 (Law of Total Probability [MU05, Theorem 1.6]).** Let $B_1$, $B_2, \ldots, B_n$ be mutually disjoint events in the sample space $\Omega$, and let $\bigcup_{i=1}^{n} B_i = \Omega$. Then

$$\Pr[A] = \sum_{i=1}^{n} \Pr[A \cap B_i] = \sum_{i=1}^{n} \Pr[A \mid B_i] \cdot \Pr[B_i].$$

◀

### 2.2.2 Random Variables

In this thesis we will analyze the number of certain sub-structures appearing in randomly generated discrete structures. In order to do so we use the concept of *random variables*. Formally a random variable is any function $X\colon \Omega \to \mathbb{R}$. However, we will only consider *discrete* random variables, i. e. random variables with a countable range.

Formally for a random variable $X$ with range $\mathrm{rng}(X)$ the probability that $X$ takes a value $x \in \mathrm{rng}(X)$ is

$$\Pr\left[X^{-1}(x)\right] = \sum_{s \in \Omega\colon X(s) = x} \Pr[s].$$

We denote this event with $\{X = x\}$ and write $\Pr[X = x]$ to denote its probability. Furthermore, we use $\{X \geq x\}$ to denote the union of all events $\{X = y\}$ with $y \geq x$.

We say that two random variables $X$ and $Y$ are *independent* iff for all values $x \in \operatorname{rng}(X)$ and $y \in \operatorname{rng}(Y)$ $\Pr[(X = x) \cap (Y = y)] = \Pr[X = x] \cdot \Pr[Y = y]$. Random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* iff for any subset $I \subseteq [n]$ and any values $x_i \in \operatorname{rng}(X_i)$, $i \in I$, $\Pr[\bigcap_{i \in I} X_i = x_i] = \prod_{i \in I} \Pr[X_i = x_i]$.

### 2.2.3 Expected Values

One important feature of a random variable that we will use throughout this thesis is its expected value. Intuitively, this is the average value that a random variable will take according to its distribution. The *expected value* of a discrete random variable $X$, denoted by $\mathbb{E}[\,X\,]$, is

$$\mathbb{E}[\,X\,] = \sum_{x \in \operatorname{rng}(X)} x \cdot \Pr[X = x] = \sum_{\omega \in \Omega} X(\omega) \cdot \Pr[\{\omega\}].$$

The expected value is finite if $\sum_{x \in \operatorname{rng}(X)} |x| \cdot \Pr[X = x]$ converges, otherwise it is unbounded. In this thesis we will only consider finite expected values.

The following very useful theorem holds for the expected value of a sum of random variables.

▶ **Theorem 2.2 (Linearity of Expectations [MU05, Theorem 2.1]).** For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[\,X_i\,].$$

◀

Furthermore, the following simple lemma holds.

▶ **Lemma 2.3 ([MU05, Lemma 2.2]).** For any constant $c$ and discrete random variable $X$,

$$\mathbb{E}[\,c \cdot X\,] = c \cdot \mathbb{E}[\,X\,].$$

◀

## 2.3 Probabilistic Inequalities

We will make use of several probabilistic inequalities in this thesis. The most important ones are stated in this section. The following theorem will be used extensively to derive upper bounds on the probability of a union of events. Although we usually use this theorem implicitly, we state it here explicitly.

▶ **Theorem 2.4 (Union Bound [MU05, Lemma 1.2]).** Let $I \subseteq \mathbb{N}$ and let $\{E_i\}_{i \in I}$ be a family of events. Then

$$\Pr\left[\bigcup_{i \in I} E_i\right] \leq \sum_{i \in I} \Pr[E_i].$$

◀

The next theorem is useful to bound the probability that a non-negative random valuable reaches a certain value when only its expected value is known.

▶ **Theorem 2.5 (Markov's Inequality [MU05, Theorem 3.1]).** Let $X$ be a non-negative random variable. Then, for all $a > 0$ it holds that

$$\Pr[X \geq a \cdot \mathbb{E}[X]] \leq \frac{1}{a}.$$

◀

We also use the following inequality, which is applicable to sums of independent random variables.

▶ **Theorem 2.6 (Chernoff's Inequality [DP09, Theorem 1.1]).** Let $X_1, X_2, \ldots, X_n$ be independent binary random variables and let $X = \sum_{i=1}^{n} X_i$. Then, for $\varepsilon > 0$

- $\Pr[X > (1+\varepsilon) \cdot \mathbb{E}[X]] \leq \exp\left(-\frac{\varepsilon^2}{3} \cdot \mathbb{E}[X]\right)$,
- $\Pr[X < (1-\varepsilon) \cdot \mathbb{E}[X]] \leq \exp\left(-\frac{\varepsilon^2}{2} \cdot \mathbb{E}[X]\right)$.

◀

The last theorem of this section is used in Chapter 4. It can be used to derive lower bounds on the probability that a random variable is non-zero.

▶ **Theorem 2.7 (Second Moment Method [Jan96]).** If $X$ is a non-negative random variable with finite variance, then

$$\Pr[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

◀

# 3 Random SAT and Satisfiability Thresholds

*Our definition of non-uniform random k-SAT in this chapter is based on joint work with Anton Krohmer, Tobias Friedrich, Thomas Sauerwald, and Andrew M. Sutton [FR18; FR19; Fri+17a; Fri+17b].*

This chapter formally introduces Boolean Satisfiability, the random $k$-SAT model, the satisfiability threshold, and concepts related to them. Afterward, we will formally introduce the non-uniform random $k$-SAT model. As the concepts related to the satisfiability threshold are only defined for random $k$-SAT, we generalize these concepts to our model. Since non-uniform random $k$-SAT generalizes some well-known random SAT models, we will highlight notable special cases apart from random $k$-SAT.

Note that the topics related to SAT are very wide and we only cover a small range of them that are relevant in this work. For further information we refer to the "Handbook of Satisfiability" [Bie+09].

## 3.1 Boolean Satisfiability

This section introduces Boolean Satisfiability (SAT). We assume that the reader is familiar with basic propositional logic as well as basic logic operators.

We let $X_1, X_2, \dots, X_n$ denote Boolean variables that can be either true or false. A *literal* $\ell$ is a Boolean variable $X_i$ or its negation $\overline{X_i}$. For a literal $\ell$ let $|\ell|$ denote the variable of the literal. A *clause* $c = (\ell_1 \vee \ell_2 \vee \dots \vee \ell_l)$ is a disjunction of distinct literals. However, to simplify notation, we will also interpret clauses as sets of literals. We will also call a clause with exactly $l$ literals an *$l$-clause*. A Boolean formula in *conjunctive normal form (CNF)* is a conjunction of clauses $\Phi = c_1 \wedge c_2 \wedge \dots \wedge c_m$. A formula is in *$k$-CNF* if it is in CNF and each clause consists of exactly $k$ literals. We conveniently interpret Boolean formulas $\Phi$ in CNF as sets of clauses. Thus, for a Boolean formula $\Phi$ in CNF we let $|\Phi|$ denote the number of its clauses.

A *truth assignment* is a vector $\alpha \in \{0,1\}^n$, which assigns the values true or false to the Boolean variables $X_1, X_2, \dots, X_n$. It assigns true to $X_i$ iff $\alpha_i = 1$. A clause is satisfied by an assignment $\alpha$ if *at least one* of its literals evaluates to true, i. e. a non-negated variable is set to true or a negated variable is set to false. A Boolean formula in CNF is satisfied by an assignment $\alpha$ if *all* of its clauses are satisfied by $\alpha$. In this case, we call $\alpha$ a *satisfying assignment* of $\Phi$. Boolean Satisfiability (SAT) is the problem of deciding if a given Boolean formula $\Phi$ in CNF has a satisfying assignment. $k$-Satisfiability ($k$-SAT) is the problem of deciding if a given Boolean formula $\Phi$ in $k$-CNF has a satisfying assignment. We

call a formula $\Phi$ *satisfiable* if it has at least one satisfying assignment. Otherwise, we call it *unsatisfiable*.

SAT was the first problem proven to be NP-complete [Coo71; Lev73]. Thus, we cannot expect to find an algorithm that solves the problem in polynomial time. The same holds for $k$-SAT with $k \geq 3$ [Kar72]. However, 1- and 2-SAT can be solved in time $O(n + m)$ [APT79]. Despite being NP-complete, there is a large body of work on exact algorithms for solving SAT and $k$-SAT. At the time of writing this, the best known general algorithm for 3-SAT runs in time $O(1.307^n)$ [Han+19]. Additionally, there are several parameterized algorithms for SAT. These algorithms have a running time of $O(f(k) \cdot \text{poly}(n+m))$ for some computable function $f$ and some parameter $k$ of the input instance. For example, SAT is fixed-parameter tractable for tree-width, branch-width, and clique-width [Sze03]. However, to the best of our knowledge there is no parameter that explains why SAT can be solved fast on industrial instances.

### 3.1.1 Different definitions for SAT

Throughout the paper we will assume the definitions stated above, since they are the most commonly used ones. However, note that there are slightly different definitions, which are widely accepted as well. For example, one could define SAT as deciding if a Boolean formula *in arbitrary form* has a satisfying assignment. In that case, one would refer to our problem definition as the SAT problem for CNF (CNFSAT). However, any Boolean formula can be transformed to a Boolean formula in conjunctive normal form [Tse83]. This can be done by introducing new variables and increasing the total number of literals of the formula only linearly. Thus, we do not consider it a restriction to assume that input formulas are in CNF. Another possible difference is if we allow duplicate literals per clause or duplicate clauses per CNF. We decided to disallow duplicate literals per clause, but to allow duplicate clauses per CNF. This is in line with many seminal works on random $k$-SAT.

## 3.2 Random k-SAT

In this section we define the seminal random $k$-SAT model and the satisfiability threshold as well as concepts related to them. These definitions are due to [CR92; Fri05; Fri99]. Note that the definition of a satisfiability threshold in this section only applies to random $k$-SAT. We will generalize this notion in Section 3.3.1.

Random $k$-SAT is a model that takes a number $n$ of Boolean variables, a clause length $k$, and a number $m$ of clauses as input. It generates a Boolean formula $\Phi$ in $k$-CNF with these parameters uniformly at random. Formally, we define the model as follows.

▶ **Definition 3.1 (Random $k$-SAT (drawing)).** Let $n, k, m$ be given. The *random $k$-SAT (drawing)* model $\mathcal{D}^R(n, k, m)$ constructs a random formula $\Phi$ in

$k$-CNF by sampling $m$ clauses independently at random. Each clause is sampled as follows:

1. Select $k$ variables independently and uniformly at random. Repeat until no variables coincide.
2. Negate each of the $k$ variables independently at random with probability $1/2$.

Thus, the probability to sample a certain clause is $\left(\binom{n}{k} \cdot 2^k\right)^{-1}$. ◀

Note that our definition of random $k$-SAT allows sampling formulas with duplicate clauses, but not with duplicate variables per clause, independent of their signs. However, there is a different definition of random $k$-SAT, which is highly-related to the former one: Imagine, instead of drawing $m$ clauses, for each of the $\binom{n}{k} \cdot 2^k$ possible clauses we flip a coin and add it to the formula with probability $p$ independently at random. Formally, we get the following model.

▶ **Definition 3.2 (Random $k$-SAT (flipping)).** Let $n, k, p$ be given. The *random $k$-SAT (flipping)* model $\mathcal{F}^R(n, k, p)$ constructs a random formula $\Phi$ in $k$-CNF by sampling each of the $\binom{n}{k} \cdot 2^k$ possible clauses $c$ independently at random:

1. With probability $p$ add $c$ to $\Phi$.
2. With probability $1 - p$ do not add it.

◀

In the literature, both models are referred to as random $k$-SAT and they are only distinguished by their set of parameters. We choose to call them $\mathcal{D}^R$ (for drawing) and $\mathcal{F}^R$ (for flipping). The difference between the two models is the same as the difference between a $G(n, m)$ and a $G(n, p)$ in graph theory: In $\mathcal{D}^R$ we draw (with replacement) $m$ of the $\binom{n}{k} \cdot 2^k$ clauses, in $\mathcal{F}^R$ we flip a coin for each of the $\binom{n}{k} \cdot 2^k$ clauses and add it to the formula with a certain probability. However, if we refer to random $k$-SAT, we mean the drawing version $\mathcal{D}^R$. We will state explicitly if we talk about $\mathcal{F}^R$.

### 3.2.1 Satisfiability Threshold

If we fix the number of variables $n$ and increase the number of clauses $m$ the probability that $\mathcal{D}^R$ generates satisfiable instances decreases. This is not surprising, since each clause is a constraint on the satisfying assignments of a Boolean formula in $k$-CNF. Thus, the more clauses a formula has, the more likely it is that the formula is not satisfiable. This property of satisfiability is called *monotonicity*. More formally, if we have a sample space $V = \{0, 1\}^N$, we call a property $P \subseteq V$ *monotone* if

$$\forall x \in P \, \forall y \in V : (\forall i \in [N] : y_i \geq x_i) \Rightarrow y \in P.$$

Intuitively, a property is monotone if adding additional elements to something with the property cannot violate it. This is true for unsatisfiability of Boolean formulas: If we have a set of clauses that is unsatisfiable, we cannot make it satisfiable by adding more clauses to it.

Monotonicity will play a crucial role in the random $k$-SAT models we consider. It implies that the probability for a property to hold increases with the scaling parameter we consider. This holds for unsatisfiability with regard to parameter $m$ of model $\mathcal{D}^R$ as we will show in Lemma 3.8. It also holds for unsatisfiability with regard to parameter $p$ of model $\mathcal{F}^R$ as we will show in Lemma 3.9.

However, there is a point at which the probability that random $k$-SAT generates unsatisfiable instances suddenly increases from close to zero to close to one. This point is called the *satisfiability threshold*. The range of $m$ in which the probability increases from close to zero to close to one is called the *threshold interval*. Formally, we define the satisfiability threshold as follows.

▶ **Definition 3.3 (Satisfiability Threshold).** $m^\star = m^\star(n,k)$ is an *asymptotic threshold function for satisfiability* if for every $m = m(n,k)$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{D}^R(n,k,m)}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } m \in o_n(m^\star) \\ 0, & \text{if } m \in \omega_n(m^\star). \end{cases}$$

We say that a satisfiability threshold exists if there is an asymptotic threshold function for satisfiability. ◀

It is important to realize what this definition actually says. For example, we know that for random $k$-SAT with $k \geq 2$ there is a satisfiability threshold and the asymptotic threshold function is $m^\star(n,k) = n$ [AP04; Kir+98]. This means, if we draw an instance with $m = n^{0.5} \in o(n)$ clauses, then the probability that this instance is satisfiable is close to one. If we draw an instance with $m = n^{1.2} \in \omega(n)$ clauses, then the probability that this instance is satisfiable is close to zero. However, if we draw an instance with $m = \varepsilon \cdot n \pm o(n)$ clauses for any constant $\varepsilon > 0$, we do not know what happens. This is in line with our definition of a satisfiability threshold: We do not care what happens at $m \in \Theta(n)$, as long as the probability to generate satisfiable instances is $1-o(1)$ for $m \in o(n)$ and $o(1)$ for $m \in \omega(n)$. See Figure 3.1 for a visual representation.

But what if we want to know what happens at $m \in \Theta(n)$? There are two ways that the probability function could behave in the range $\Theta(n)$. Either, there is a small interval of size $o(n)$, where it suddenly drops from close to one to close to zero. If this is the case, we call the threshold *sharp*. This is what we observe for random $k$-SAT. Intuitively, a sharp threshold means that the size of the threshold interval grows asymptotically slower than the actual threshold. Thus, the threshold interval seems to vanish in the limit, which makes the probability function look more and more like a step function. If the threshold is not sharp, we call it *coarse*. This could mean that the function decrease more slowly, in an interval of size $\Theta(n)$, in which the probability function is bounded away from

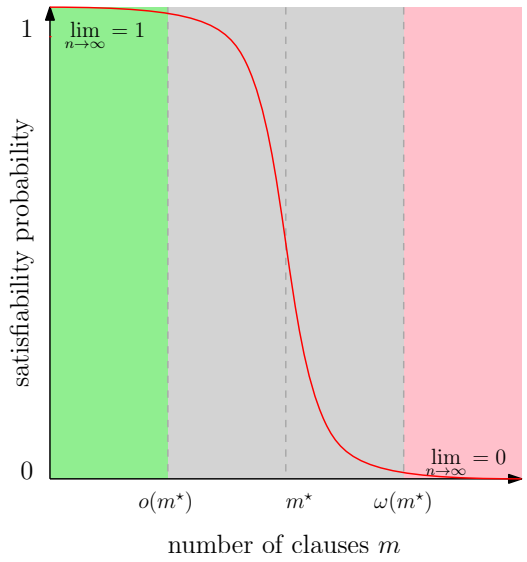



**Figure 3.1:** Visual representation of a satisfiability threshold with asymptotic threshold function $m^\star$. For all functions $m \in o(m^\star)$ the probability tends to one (green region), for all functions $m \in \omega(m^\star)$ the function tends to zero (red region), for all functions $m \in \Theta(m^\star)$ the function is not restricted (gray region).

zero and one by a constant. However, it could also mean that the limit of the probability function is not defined in that region. Formally, we define sharp and coarse thresholds as follows.

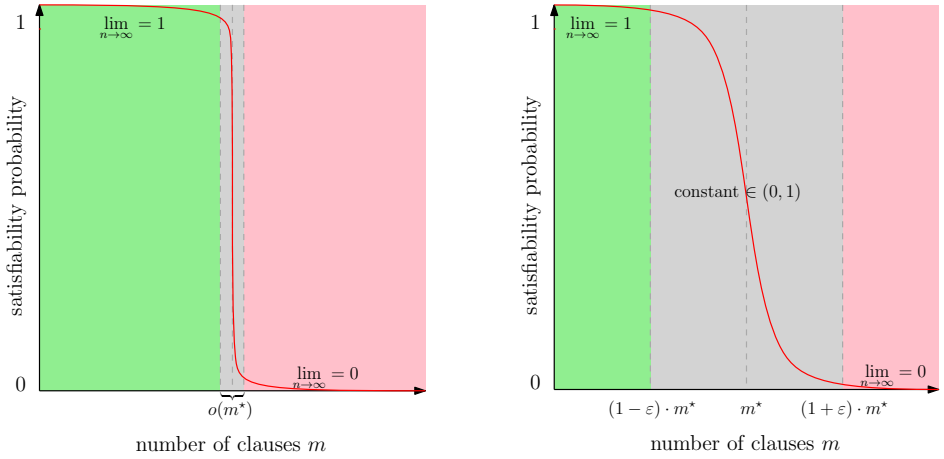
▶ **Definition 3.4 (Sharpness).** Let $m^\star = m^\star(n,k)$ be an asymptotic threshold function for satisfiability. We call the threshold *sharp* if for every constant $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{D}^R(n,k,m)}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } m = (1-\varepsilon)\cdot m^\star \\ 0, & \text{if } m = (1+\varepsilon)\cdot m^\star. \end{cases}$$

Otherwise, we call the threshold *coarse*. ◀

Those two cases are mutually exclusive and one of the two has to hold if an asymptotic threshold function exists. Thus, if there is a satisfiability threshold, we can always classify it as either sharp or coarse. See Figure 3.2 for a visual representation of sharp and coarse thresholds. Friedgut [Fri99] proved that random $k$-SAT with $k \geq 2$ has a sharp threshold, although he did not determine the exact threshold function.

So far we know about the existence and sharpness of satisfiability thresholds. For random $k$-SAT there is a sharp satisfiability threshold at $m^\star \in \Theta(n)$. This has been observed experimentally and proven rigorously. What remains is to determine *where* the satisfiability threshold is exactly. For a fixed $k$ the position of the threshold always seems to converge to the same clause-variable ratio




**(a)** Sharp threshold: $m^\star$ is an asymptotic threshold function. For any constant $\varepsilon > 0$ the probability tends to one at $m = (1-\varepsilon)\cdot m^\star$ (green region) and to zero at $m = (1+\varepsilon)\cdot m^\star$ (red region). The range where the function is not restricted (gray region) is of size $o(m^\star)$.

**(b)** Coarse threshold: $m^\star$ is an asymptotic threshold function. There is a constant $\varepsilon > 0$ such that the satisfiability probability is bounded away from zero and one by a constant for all $m \in [(1-\varepsilon)\cdot m^\star, (1+\varepsilon)\cdot m^\star]$.

**Figure 3.2:** Visual representation of a sharp and a coarse satisfiability threshold with asymptotic threshold function $m^\star$.

$m/n = r_k$. For random 3-SAT this point is at roughly $r_k \approx 4.26$. This lead to the following conjecture.

▶ **Conjecture 3.5 (Satisfiability Threshold Conjecture).** For each $k \geq 2$ there is a constant $r_k$, which might depend on $k$, such that for every constant $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{D}^R(n,k,m)}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } m = (1-\varepsilon)\cdot r_k \cdot n \\ 0, & \text{if } m = (1+\varepsilon)\cdot r_k \cdot n. \end{cases}$$

◀

The satisfiability threshold conjecture states that there is a sharp threshold with threshold function $m^\star = r_k \cdot n$. Note that this conjecture is not implied by our definitions of a satisfiability threshold and sharpness alone, even if we know that there is a sharp threshold at $m \in \Theta(n)$. As a counter-example imagine a sharp threshold function

$$m^\star(n) = \begin{cases} 2\cdot n & n \text{ odd} \\ 2.1\cdot n & n \text{ even} \end{cases}.$$

It holds that there is a sharp threshold at $m^\star \in \Theta(n)$, but we cannot determine a single leading constant $r_k$ such that $m^\star(n)$ converges to $r_k \cdot n$. At the time of

writing this thesis the satisfiability threshold conjecture has been proven with $r_2 = 1$ by Chvátal and Reed [CR92] and with $r_k = 2^k \log 2 - \frac{1}{2}(1 + \log 2)$ for very large values of $k$ by Ding et al. [DSS15]. For everything from $k = 3$ to these very large values, it remains open.

## 3.3 Non-Uniform Random k-SAT

In this section we introduce a generalization of random $k$-SAT, which we call *non-uniform random k-SAT*. The model is inspired by power-law random $k$-SAT and geometric random $k$-SAT by Ansótegui et al. [ABL09b]. These two models are also notable special cases of non-uniform random $k$-SAT (cf. Section 3.3.2). Afterward, we generalize the concepts related to the satisfiability threshold to our new model.

As in random $k$-SAT we draw $m$ clauses independently at random. However, the clause probabilities are now non-uniform. More precisely, each Boolean variable is assigned a probability. Then, $k$ Boolean variables are drawn without replacement according to that probability and then negated independently with probability $1/2$ each. The variable probabilities act as expected frequencies of those variables in the resulting formula. This allows us to model different frequency distributions. Formally, we define our model as follows.

▶ **Definition 3.6 (Clause-Drawing Non-Uniform Random $k$-SAT).** Let $m$, $n$, $k$ be given, and consider an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}} = ({p_1}^{(n)}, \ldots, {p_n}^{(n)})_{n\in\mathbb{N}}$, where each distribution $\vec{p}^{(n)}$ is defined over $n$ Boolean variables with ${p_1}^{(n)}, \ldots, {p_n}^{(n)} > 0$ and $\sum_{i=1}^{n} {p_i}^{(n)} = 1$. The *clause-drawing non-uniform random k-SAT (non-uniform random k-SAT)* model $\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ constructs a random formula $\Phi$ in k-CNF by sampling $m$ clauses independently at random. Each clause is sampled as follows:

1. Select $k$ variables independently at random according to the distribution $\vec{p}^{(n)}$. Repeat until no variables coincide.
2. Negate each of the $k$ variables independently at random with probability $1/2$.

◀

W. l. o. g. we will assume for all $n \in \mathbb{N}$ ${p_1}^{(n)} \geq {p_2}^{(n)} \geq \ldots {p_n}^{(n)}$. To simplify notation, we denote $p^{(n)}(X_i) := \Pr[X = X_i] = {p_i}^{(n)}$. Throughout this thesis we will consider the limit behavior of probabilities in our ensembles. Thus, based on an ensemble of discrete probability distributions $\left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}}$ we define for all $i \in \mathbb{N}$ the functions $p_i \colon \mathbb{N} \setminus [i-1] \to \mathbb{R}^+$ with $p_i(n) = {p_i}^{(n)}$. However, for the sake of brevity we omit the input parameter $n$ of those functions if it is not necessary, i. e. most of the expressions we derive are actually functions in $n$ if not stated otherwise.

The clause-drawing non-uniform random $k$-SAT model is equivalent to drawing each clause independently at random from the set of all $k$-clauses which contain no variable more than once. The probability to draw a $k$-clause $c = (\ell_1 \vee \ell_2 \vee \ldots \vee \ell_k)$ over $n$ variables in this model is

$$q_c = \frac{\prod_{\ell \in c} p(|\ell|)}{2^k \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j}. \tag{3.1}$$

The factor $2^k$ in the denominator comes from the different possibilities to negate variables. Note that $k! \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j$ is the probability of choosing a $k$-clause that contains no variable more than once. We define

$$C := \left( k! \cdot \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j \right)^{-1} \tag{3.2}$$

and write

$$q_c = C \frac{k!}{2^k} \prod_{\ell \in c} p(|\ell|). \tag{3.3}$$

Remember that both $C$ and $q_c$ are actually functions in $n$. The representation of equation (3.3) makes clause probabilities easier to handle. Since clauses are also drawn independently, the probability to generate a formula $\Phi$ with non-uniform random $k$-SAT essentially comes down to a product of variable probabilities for Boolean variables it contains. This makes the analysis of formulas a lot easier.

We are mainly interested in the clause-drawing version of non-uniform random $k$-SAT. However, as in the case of random $k$-SAT, we can define a clause-flipping equivalent of this model, in which we flip a coin for each possible clause $c$ with probability proportional to $q_c$ as defined in equation (3.3). We will use this model and its relation to the clause-drawing version to derive our results on the sharpness of the satisfiability threshold in Chapter 6. Formally, we define the clause-flipping version of non-uniform random $k$-SAT as follows.

▶ **Definition 3.7 (Clause-Flipping Non-Uniform Random $k$-SAT).** Let $n$, $k$ be given, and consider an ensemble of probability distributions $(\vec{p}^{(n)})_{n \in \mathbb{N}} = ({p_1}^{(n)}, \ldots, {p_n}^{(n)})_{n \in \mathbb{N}}$, where each distribution $\vec{p}^{(n)}$ is defined over $n$ Boolean variables with ${p_1}^{(n)}, \ldots, {p_n}^{(n)} > 0$ and $\sum_{i=1}^{n} {p_i}^{(n)} = 1$. The *clause-flipping non-uniform random $k$-SAT* model $\mathcal{F}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, s)$ constructs a random Boolean formula $\Phi$ over $n$ variables in k-CNF by independently flipping a coin for each of the $\binom{n}{k} 2^k$ possible $k$-clauses. The coin flip for a clause $c$ is a success with probability

$$q_c(s) = \min(s \cdot q_c, 1) = \min\left( s \cdot \frac{\prod_{\ell \in c} p(|\ell|)}{2^k \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j}, 1 \right),$$

where $s \in [0, 1/\min_{c \in C}(q_c)]$ is a scaling factor. If successful, the clause is added to the random formula. ◀

In the *uniform* clause-flipping model $\mathcal{F}^R$ the scaling factor is the uniform probability $p \in [0, 1]$ with which a clause is taken into the formula. There are several ways to generalize this to non-uniform distributions.

If we assume to have a probability vector $\vec{q} = (q_c)_{c \in C}$, where $C$ is the set of all $k$-clauses over $n$ variables, scaling with a factor $s \in [0, 1/\min_{c \in C}(q_c)]$ makes sense. Then, $s$ represents the expected number of clauses in the random formula.

Alternatively, we could assume to be given a vector of clause weights $\vec{w}$ with $\vec{w} = (w_c)_{c \in C}$ and $\min_{c \in C}(w_c) = 1$ for all $n \in \mathbb{N}$. Then, the clause probabilities would be $\min(s \cdot w_c, 1)$ with a scaling factor of $s \in [0, 1]$. This gives a nicer scaling and resembles the uniform case, but the scaling factor would not actually represent clause probabilities properly as it does in the uniform case.

In both cases we get a skewed distribution as soon as the probabilities get capped at one. We decided for $s$ representing the expected number of clauses for two reasons. First, it makes the two models easily comparable, because the asymptotic thresholds of $\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ and $\mathcal{F}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, s)$ will be of the same size (c.f. Section 6.1). Second, it makes the notation in some of the proofs easier.

### 3.3.1 Satisfiability Threshold

As in the case of random $k$-SAT, we want to study the satisfiability threshold as the number of clauses increases. Thus, we are only interested in satisfiability thresholds in parameter $m$ of $\mathcal{D}^N$ and in parameter $s$ of $\mathcal{F}^N$. Both parameters capture the (expected) number of clauses of instances generated by those models. As such they are a natural choice for scaling parameters in accordance with random $k$-SAT.

As in the case of random $k$-SAT, we first want to see if the probability to generate unsatisfiable instances is non-decreasing in $m$. The following lemma establishes that this holds for the probability that *any* monotone property $P$ is fulfilled in $\mathcal{D}^N$ with respect to $m$.

▶ **Lemma 3.8.** Fix $n \in \mathbb{N}$, $k \in \mathbb{N}$, and a probability ensemble $(\vec{p}^{(n)})_{n \in \mathbb{N}}$. Let $P$ be any monotone property. Then, the probability to generate an instance with property $P$ in $\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ is non-decreasing in $m$. ◀

*Proof.* Let $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $(\vec{p}^{(n)})_{n \in \mathbb{N}}$ be arbitrary, but fixed. Since $m$ is the only free parameter, we let $\mathcal{D}^N(m)$ denote $\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ for the sake of simplicity. Now choose some $m \in \mathbb{N}$ arbitrarily. We are going to show that

$$\Pr_{\Phi \sim \mathcal{D}^N(m+1)}[\, P(\Phi) = 1 \,] \geq \Pr_{\Phi \sim \mathcal{D}^N(m)}[\, P(\Phi) = 1 \,].$$

We interpret each formula $\Phi$ as a sequence of (not necessarily distinct) clauses $(c_1, c_2, \ldots, c_m)$. Since clauses are drawn independently with replacement in $\mathcal{D}^N$,

for two formulas $x = (c_1, c_2, \ldots, c_m)$ and $y = (c_1, c_2, \ldots, c_m, c_{m+1})$ it holds that

$$\Pr_{\Phi \sim \mathcal{D}^N(m+1)}[\,\Phi = y\,] = \Pr_{\Phi \sim \mathcal{D}^N(m)}[\,\Phi = x\,] \cdot \Pr_{\Phi \sim \mathcal{D}^N(1)}[\,\Phi = (c_{m+1})\,].$$

Now let $P_{k,m}$ denote the set of all formulas with property $P$ in $k$-CNF with at most $m$ clauses and let $C$ denote the set of all possible $k$-clauses over $n$ variables. Then,

$$\begin{aligned}
&\Pr_{\Phi \sim \mathcal{D}^N(m+1)}[\,P(\Phi) = 1\,] \\
&= \sum_{y=(c_1,\ldots,c_m,c) \in P_{k,m+1}} \Pr_{\Phi \sim \mathcal{D}^N(m+1)}[\,\Phi = y\,] \\
&= \sum_{x=(c_1,\ldots,c_m) \in C^m} \left( \Pr_{\Phi \sim \mathcal{D}^N(m)}[\,\Phi = x\,] \cdot \sum_{\substack{c \in C: \\ (c_1,\ldots,c_m,c) \in P_{k,m+1}}} \left( \Pr_{\Phi \sim \mathcal{D}^N(1)}[\,\Phi = (c)\,] \right) \right) \\
&\geq \sum_{x=(c_1,\ldots,c_m) \in P_{k,m}} \left( \Pr_{\Phi \sim \mathcal{D}^N(m)}[\,\Phi = x\,] \cdot \sum_{\substack{c \in C: \\ (c_1,\ldots,c_m,c) \in P_{k,m+1}}} \left( \Pr_{\Phi \sim \mathcal{D}^N(1)}[\,\Phi = (c)\,] \right) \right).
\end{aligned}$$

Due to the monotonicity of $P$, if $x \in P_{k,m}$, then any $y$ which extends $x$ by one clause is in $P_{k,m+1}$. Thus,

$$\begin{aligned}
&= \sum_{y=(c_1,\ldots,c_m,c_{m+1}) \in P_{k,m}} \left( \Pr_{\Phi \sim \mathcal{D}^N(m)}[\,\Phi = x\,] \cdot \sum_{c \in C} \left( \Pr_{\Phi \sim \mathcal{D}^N(1)}[\,\Phi = (c)\,] \right) \right) \\
&= \Pr_{\Phi \sim \mathcal{D}^N(m)}[\,P(\Phi) = 1\,],
\end{aligned}$$

since

$$\sum_{c \in C} \Pr_{\Phi \sim \mathcal{D}^N(1)}[\,\Phi = (c)\,] = 1.$$

This proves that the probability for $P$ is non-decreasing in $\mathcal{D}^N$ as $m$ increases. ■

Monotonicity will be crucial for many proofs in this thesis. We will now see that the probability for a monotone property to hold is also non-decreasing in $\mathcal{F}^N$ as $s$ increases.

▶ **Lemma 3.9.** Fix $n \in \mathbb{N}$, $k \in \mathbb{N}$, and a probability ensemble $\left(\vec{p}^{(n)}\right)_{n \in \mathbb{N}}$. Let $P$ be any monotone property. Then, the probability to generate an instance with property $P$ in $\mathcal{F}^N\left(n, k, \left(\vec{p}^{(n)}\right)_{n \in \mathbb{N}}, s\right)$ is non-decreasing in $s$. ◀

*Proof.* Again, we fix $n \in \mathbb{N}$, $k \in \mathbb{N}$, and a probability ensemble $\left(\vec{p}^{(n)}\right)_{n \in \mathbb{N}}$. Since the only free parameter is $s$ we let $\mathcal{F}^N(s)$ denote $\mathcal{F}^N\left(n, k, \left(\vec{p}^{(n)}\right)_{n \in \mathbb{N}}, s\right)$ for the sake of simplicity. Now we choose $s, s' \in R^+$ so that $s \leq s'$. We want to show

that

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] \leq \Pr_{\Phi\sim\mathcal{F}^N(s')}[\,P(\Phi)=1\,].$$

We are going to show something more general, namely that increasing *any* clause probability also increases the probability for the monotone property to hold. Let us assume that the clauses are identified by indices $i \in [N]$, where $N = \binom{n}{k}\cdot 2^k$ is the total number of different $k$-clauses. In this context, we interpret a formula $\Phi$ as a subset of those indices. W. l. o. g. we increase the probability of the first clause $c_1$ from $s\cdot q_1$ to $s'\cdot q_1$. Then,

$$\begin{aligned}&\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,]\\ &\qquad=\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\in\Phi\wedge P(\Phi)=1\,]+\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\notin\Phi\wedge P(\Phi)=1\,].\end{aligned}$$

Since clauses are incorporated into formulas independently, we can consider the restriction of $\mathcal{F}^N(s)$ to $[2..N]$, i. e. we ignore $c_1$ and only consider and sample the remaining clauses. We denote this model as $\mathcal{F}^N_{\bar{1}}(s)$. It holds that

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\in\Phi\wedge P(\Phi)=1\,] = s\cdot q_1\cdot\sum_{\Phi'\subseteq[2..N]\colon P(\{1\}\cup\Phi')=1}\Pr_{\Phi\sim\mathcal{F}^N_{\bar{1}}(s)}[\,\Phi=\Phi'\,]$$

and

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\notin\Phi\wedge P(\Phi)=1\,] = (1-s\cdot q_1)\cdot\sum_{\Phi'\subseteq[2..N]\colon P(\Phi')=1}\Pr_{\Phi\sim\mathcal{F}^N_{\bar{1}}(s)}[\,\Phi=\Phi'\,].$$

However, since $P$ is monotone, any $\Phi'$ with $P(\Phi')=1$ also satisfies $P(\{1\}\cup\Phi')=1$. Thus, the factors $s\cdot q_1$ and $1-s\cdot q_1$ for the probabilities of those formulas add up to one. What remains are formulas $\Phi'$ with $P(\Phi')=0$ and $P(\{1\}\cup\Phi')=1$. This yields

$$\begin{aligned}&\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\in\Phi\wedge P(\Phi)=1\,]+\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,c_1\notin\Phi\wedge P(\Phi)=1\,]\\ &=\sum_{\substack{\Phi'\subseteq[2..N]\colon\\ P(\Phi')=1}}\left(\Pr_{\Phi\sim\mathcal{F}^N_{\bar{1}}(s)}[\,\Phi=\Phi'\,]\right)+s\cdot q_1\cdot\sum_{\substack{\Phi'\subseteq[2..N]\colon\\ P(\Phi')=0\wedge P(\Phi'\cup\{1\})=1}}\left(\Pr_{\Phi\sim\mathcal{F}^N_{\bar{1}}(s)}[\,\Phi=\Phi'\,]\right)\end{aligned}$$

Now it is obvious that increasing the probability of the first clause can only increase the total probability for $P$ to hold. If we assume to be given the new clause probabilities with probability $s'\cdot q_1$ for the first clause, we can now repeat the same argument when increasing the second clause probability from $s\cdot q_2$ to $s'\cdot q_2$. Repeating this step implies the desired result

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] \leq \Pr_{\Phi\sim\mathcal{F}^N(s')}[\,P(\Phi)=1\,].$$

■

If we now want to study the threshold behavior of non-uniform random $k$-SAT, we first have to generalize the concept of satisfiability thresholds to the non-uniform case. In the uniform case the probability distribution is $\vec{p}^{(n)} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ for every $n \in N$. This ensemble of distributions is stated implicitly in the model. Since we want to consider different probability distributions and study the behavior of the probability to sample satisfiable instances as $n$ increases, we have to make the ensemble of probability distributions explicit in our model. Now imagine $\mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ with the following ensemble of probability distributions:

$$\vec{p}^{(n)} = \begin{cases} \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) & , n \text{ even} \\ \left(\underbrace{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}}_{\lfloor\sqrt{(}n)/2\rfloor \text{times}}, \frac{1-\left(\lfloor\sqrt{(}n)/2\rfloor\right)/\sqrt{n}}{n-\lfloor\sqrt{n}\rfloor}, \ldots, \frac{1-\left(\lfloor\sqrt{(}n)/2\rfloor\right)/\sqrt{n}}{n-\lfloor\sqrt{n}\rfloor}\right) & , n \text{ odd.} \end{cases} \quad (3.4)$$

How to define the satisfiability threshold for such an ensemble? Can we define it at all? In this more general setting, consider the satisfiability threshold to be defined as follows.

▶ **Definition 3.10 (Satisfiability Threshold).** Let $\mathcal{M}$ be a random SAT model with parameters $n$ and $p$. Fix all parameters of the model except for $n$ and $p$. Let $p^\star$ and $p'$ be functions, which may depend on the other parameters of $\mathcal{M}$. $p^\star$ is an *asymptotic threshold function for satisfiability* of model $\mathcal{M}$ with respect to parameter $p$ if for every $p'$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{M}(p=p')}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } p' \in o_n(p^\star) \\ 0, & \text{if } p' \in \omega_n(p^\star). \end{cases}$$

We say that a satisfiability threshold with respect to $p$ exists if there is an asymptotic threshold function for satisfiability. ◀

Let us return to our example above (equation (3.4)). We can now see that definition 3.10 can be applied as follows: We define a threshold function $m^\star$ that is tailored to each probability distribution from the ensemble separately, in our case to odd $n$ and even $n$, respectively. For any other function $m'$ with $m' \in o(m^\star)$ it has to hold that

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{D}^N(n,2,(\vec{p}^{(n)})_{n\in\mathbb{N}},m')}[\,\Phi \text{ satisfiable}\,] = 1.$$

First, let us make clear what $m' \in o(m^\star)$ means in this context. Keeping in mind our definitions of limits and $O$-notation, $m' \in o(m^\star)$ means that for any

constant $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$m'(n) < \varepsilon \cdot m^\star(n).$$

This definition also holds if the threshold function is defined for odd and even $n$ separately. Thus, we can say with certainty for which functions $m'$ the statements in the threshold definition have to hold. The definition now states something for the probability to generate satisfiable instances. However, if we plug function $m'$ into our model, this probability only depends on $n$, since all other parameters are fixed according to definition 3.10. Therefore, the probability to generate satisfiable instances is a function only depending on $n$. For this function we have to check if its limit is one, i. e. if for every constant $\varepsilon \in (0, 1)$ there is an $n_0 \in \mathbb{N}$ such that the value of the function is at least $1 - \varepsilon$ for all $n \geq n_0$. For $m' \in \omega(m^\star)$ the argumentation is equivalent.

Our definition of satisfiability thresholds is very general. One can also imagine random $k$-SAT, where we fix the number of clauses $m$ (or a function $m(n)$) and increase the clause length $k$. This does not seem intuitive, but it actually makes sense if we remember that a satisfiability threshold is only a phase transition. For example, water undergoes a phase transition from liquid to gaseous state. If we fix all environmental parameters and only increase temperature, this transition happens at some point. Yet, instead of increasing temperature, one can also decrease atmospheric pressure. This will also lead to the water transitioning from one phase to the other. However, as we wrote before, in this work we are only interested in satisfiability thresholds in parameter $m$ of $\mathcal{D}^N$ and in parameter $s$ of $\mathcal{F}^N$.

We can now go on and generalize sharpness and coarseness of satisfiability thresholds similar to definition 3.10.

▶ **Definition 3.11 (Sharpness).** Let $\mathcal{M}$ be a random SAT model with parameters $n$ and $p$. Fix all parameters of the model except for $n$ and $p$. Let $p^\star$ be an asymptotic threshold function of $\mathcal{M}$ with respect to parameter $p$. We call $p^\star$ *sharp* if for every function $p'$ and every constant $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr_{\Phi \sim \mathcal{M}(p=p')}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } p' = (1-\varepsilon)\cdot p^\star \\ 0, & \text{if } p' = (1+\varepsilon)\cdot p^\star. \end{cases}$$

Otherwise we call $p^\star$ coarse. ◀

### 3.3.2 Notable Special Cases

In this section we introduce notable special cases of non-uniform random $k$-SAT from the literature. These will serve as examples for applying our results throughout this thesis.

### Random k-SAT

The most prominent special case of non-uniform random $k$-SAT is (uniform) random $k$-SAT. In this case the probability ensemble is

$$\forall n \in \mathbb{N} \colon \vec{p}^{(n)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

We already presented many results on random $k$-SAT, including the behavior of the satisfiability threshold. Those results can also be derived with the more general theorems we provide in this thesis. They will serve as some kind of sanity check and as a baseline for the other distributions.

### Power-law Random k-SAT

Power-law random $k$-SAT was introduced by Ansótegui et al. [ABL09b] as a more realistic model for industrial SAT instances. The variable probabilities in this model follow a discrete power law with power-law exponent $\beta > 2$. More precisely for some fixed $\beta > 2$ and some $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{(n/i)^{\frac{1}{\beta-1}}}{\sum_{j=1}^{n} (n/j)^{\frac{1}{\beta-1}}}.$$

Ansótegui et al. [ABL09b] claim that instances generated with their model exhibit a satisfiability threshold. They experimentally determine the threshold position and examine the running time of state-of-the-art SAT solvers on instances generated at the threshold. They observe that the running time of solvers can be controlled with the power-law exponent $\beta$. With increasing exponent instances get more similar to those generated by random $k$-SAT. Thus, solvers specialized in random instances perform better. With small exponents, the performance of solvers specialized in industrial instances is better. According to the authors this phenomenon can be used to generate instances on which the performance of state-of-the-art SAT solvers is comparable to the performance of those solvers on industrial instances.

However, Ansótegui et al. do not determine the threshold position rigorously. Furthermore they do not show theoretical bounds on the running times of the solvers they consider. The results of this thesis complement their work with regard to the threshold behavior of power-law random $k$-SAT. In our continuing work [Blä+21] we also show some first lower bounds on the resolution size of this model, which might explain some of the observations of Ansótegui et al. We mention those results briefly in Chapter 7.

In order to derive the results for power-law random $k$-SAT, we need the following bounds.

▶ **Lemma 3.12.** For the discrete power-law distribution $\vec{p}$ with exponent $\beta > 2$ it holds that

$$p_i = (1 + o(1)) \cdot \frac{\beta - 2}{\beta - 1} \cdot n^{-1} \cdot \left(\frac{n}{i}\right)^{1/(\beta-1)},$$

$$\sum_{i=1}^{n} {p_i}^2 = \begin{cases} \Theta\left(n^{-2\frac{\beta-2}{\beta-1}}\right) & \text{for } \beta < 3 \\ (1 \pm o(1)) \cdot \frac{1}{4} \cdot \frac{\ln n}{n} & \text{for } \beta = 3 \\ (1 \pm o(1)) \cdot \frac{(\beta-2)^2}{(\beta-3)\cdot(\beta-1)} \cdot n^{-1} & \text{for } \beta > 3, \text{ and} \end{cases}$$

$$\sum_{j=1}^{i} p_j \leq (1 + o(1)) \cdot \left(\frac{i}{n}\right)^{\frac{\beta-2}{\beta-1}}.$$

◀

*Proof.* It holds that

$$1 + \int_{i=1}^{n} \left(\frac{n}{i}\right)^{1/(\beta-1)} di \leq \sum_{i=1}^{n} \left(\frac{n}{i}\right)^{1/(\beta-1)} \leq n^{1/(\beta-1)} + \int_{i=1}^{n} \left(\frac{n}{i}\right)^{1/(\beta-1)} di.$$

Since

$$\int_{i=1}^{n} \left(\frac{n}{i}\right)^{1/(\beta-1)} di = \frac{\beta - 1}{\beta - 2} \cdot \left(n - n^{1/(\beta-1)}\right),$$

for $\beta > 2$, we have

$$\sum_{i=1}^{n} \left(\frac{n}{i}\right)^{1/(\beta-1)} = (1 - o(1)) \cdot \frac{\beta - 1}{\beta - 2} \cdot n$$

and thus

$$p_i = (1 + o(1)) \cdot \frac{\beta - 2}{\beta - 1} \cdot n^{-1} \cdot \left(\frac{n}{i}\right)^{1/(\beta-1)}.$$

For $\beta = 3$ it holds that

$$\sum_{i=1}^{n} p_i^2 = (1 + o(1)) \cdot \left(\frac{\beta - 2}{\beta - 1}\right)^2 \cdot n^{-1} \cdot \sum_{i=1}^{n} \frac{1}{i} = (1 + o(1)) \cdot \frac{1}{4} \cdot \frac{\ln}{n}.$$

Otherwise, we consider the function

$$\left(\frac{\beta - 2}{\beta - 1}\right)^2 \cdot n^{-2\frac{\beta-2}{\beta-1}} \cdot i^{-\frac{2}{\beta-1}}$$

and the integral

$$\int_{i=1}^{n} \left(\frac{\beta - 2}{\beta - 1}\right)^2 \cdot n^{-2\frac{\beta-2}{\beta-1}} \cdot i^{-\frac{2}{\beta-1}} di = \frac{(\beta - 2)^2}{(\beta - 3) \cdot (\beta - 1)} \cdot \left(n^{-1} - n^{-2\frac{\beta-2}{\beta-1}}\right).$$

Again, we can use the relation of sum and integral to derive

$$\sum_{i=1}^{n} p_i^2 = (1 + o(1)) \cdot \frac{(\beta - 2)^2}{(\beta - 3) \cdot (\beta - 1)} \cdot n^{-1}$$

for $\beta > 3$ and

$$\sum_{i=1}^{n} p_i^2 \in \Theta\left(n^{-2\frac{\beta-2}{\beta-1}}\right)$$

for $\beta < 3$ as desired. Although we do not derive the exact leading factor in the last case, the asymptotic expression is sufficient for our results. The second statement of the theorem holds since

$$\begin{aligned}\sum_{j=1}^{i} p_j &\leq p_1 + \int_{j=1}^{i} p_j \, dj \\ &= (1 + o(1)) \cdot \left(\left(\frac{\beta - 2}{\beta - 1}\right) \cdot n^{-\frac{\beta-2}{\beta-1}} + \left(\frac{i}{n}\right)^{\frac{\beta-2}{\beta-1}} - n^{-\frac{\beta-2}{\beta-1}}\right) \\ &\leq (1 + o(1)) \cdot \left(\frac{i}{n}\right)^{\frac{\beta-2}{\beta-1}}.\end{aligned}$$

■

### Geometric Random k-SAT

Geometric Random $k$-SAT was introduced by Ansótegui et al. [ABL09b] as an alternative to Power-law Random $k$-SAT. In this model the variable probabilities are normalized terms of a geometric series with base $1/b$ for some constant $b > 1$. More precisely for some fixed $b > 1$ and some $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{b \cdot (1 - b^{-1/n})}{b - 1} \cdot b^{-(i-1)/n}.$$

Ansótegui et al. [ABL09b] determine the position of the satisfiability threshold for geometric random $k$-SAT experimentally. For base parameter $b = 1$ the model is equivalent to random $k$-SAT. However, the authors observe that instances get easier as the base parameter $b$ increases. As for power-law random $k$-SAT Ansótegui et al. do not provide any rigorous results regarding the satisfiability threshold. The results of this thesis complement their work in that regard.

We will make use of the following lemma in order to derive our results.

▶ **Lemma 3.13.** For the discrete geometric distribution $\vec{p}$ with base $b > 1$ it

holds that

$$\sum_{i=1}^{n} p_i^2 = \frac{b+1}{b-1} \cdot \frac{1-b^{-1/n}}{1+b^{-1/n}} = (1 \pm o(1)) \cdot \frac{b+1}{b-1} \cdot \frac{\ln b}{2} \cdot n^{-1}.$$

It also holds that

$$p_1 = \frac{b \cdot (1-b^{-1/n})}{b-1} = (1-o(1)) \cdot \frac{b \cdot \ln b}{b-1} \cdot n^{-1}.$$

◀

*Proof.* It holds that

$$\sum_{i=1}^{n} p_i^2 = \frac{b^2 \cdot (1-b^{-1/n})^2}{(b-1)^2} \cdot \sum_{i=0}^{n-1} \left(\frac{1}{b^{2/n}}\right)^i = \frac{b+1}{b-1} \cdot \frac{1-b^{-1/n}}{1+b^{-1/n}},$$

since this is a simple geometric series. We get

$$\begin{aligned}\sum_{i=1}^{n} p_i^2 = \frac{b+1}{b-1} \cdot \frac{1-b^{-1/n}}{1+b^{-1/n}} &= \frac{b+1}{b-1} \cdot \frac{1-e^{-\ln(b)/n}}{1+e^{-\ln(b)/n}} \\ &\leq \frac{b+1}{b-1} \cdot \frac{1-(1-\ln(b)/n)}{1+(1-\ln(b)/n)} \\ &= \left(1+\frac{\ln b}{2n-\ln b}\right) \cdot \frac{b+1}{b-1} \cdot \frac{\ln(b)}{2n}\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n} p_i^2 = \frac{b+1}{b-1} \cdot \frac{1-b^{-1/n}}{1+b^{-1/n}} &= \frac{b+1}{b-1} \cdot \frac{b^{1/n}-1}{b^{1/n}+1} \\ &= \frac{b+1}{b-1} \cdot \frac{e^{\ln(b)/n}-1}{(1+(b-1))^{1/n}+1} \\ &\geq \frac{b+1}{b-1} \cdot \frac{(1+\ln(b)/n)-1}{1+(b-1)/n+1} \\ &= \left(1-\frac{b-1}{2n+b-1}\right) \cdot \frac{b+1}{b-1} \cdot \frac{\ln(b)}{2n},\end{aligned}$$

where we used Bernoulli's inequality in the denominator of the third line. This establishes the first statement. For the second statement observe

$$p_1 = \frac{b \cdot (1-b^{-1/n})}{b-1} = \frac{b \cdot (1-e^{-\ln(b)/n})}{b-1} \leq \frac{b \ln(b)}{b-1} \cdot n^{-1}$$

and

$$p_1 = \frac{b \cdot (1-b^{-1/n})}{b-1} = \frac{b \cdot (b^{1/n}-1)}{(b-1) \cdot b^{1/n}}$$

$$
\begin{aligned}
&= \frac{b \cdot (e^{\ln(b)/n} - 1)}{(b-1) \cdot (1 + (b-1))^{1/n}} \\
&\geq \frac{b \cdot \ln(b)/n}{(b-1) \cdot (1 + (b-1)/n)} \\
&= \left(1 - \frac{b-1}{n+b-1}\right) \cdot \frac{b \ln(b)}{b-1} \cdot n^{-1}.
\end{aligned}
$$

■

# 4 Satisfiability Threshold in Non-Uniform Random 2-SAT

*The content of this chapter is based on the publication [FR19], which is joint work with Tobias Friedrich, and the publication [Fri+17b], which is joint work with Tobias Friedrich, Anton Krohmer, and Andrew M. Sutton. [Fri+17b] contains an early version of the results in Section 4.2 and the experimental results. [FR19] contains a complete collection of the results this chapter is based on. Here, we adjusted the proofs a bit to explicitly capture the fourth case, which was not considered in the conference version. More precisely, we substituted the asymptotic expressions in our statements by exact lower and upper bounds which imply the same results and also yield the statement of the last case.*

In this chapter we analyze the behavior of the satisfiability threshold in non-uniform random 2-SAT. Although 2-SAT can be solved in polynomial time and is therefore not NP-complete, random 2-SAT exhibits a threshold behavior similar to random $k$-SAT for bigger values of $k$. Due to the simpler structure of Boolean formulas in 2-CNF this threshold behavior is much easier to analyze. The insights from analyzing non-uniform random 2-SAT will help us to understand and analyze the satisfiability threshold of non-uniform random $k$-SAT for bigger values of $k$.

Chvátal and Reed [CR92] showed that random 2-SAT has a sharp satisfiability threshold at $m^\star = n$, thus confirming the satisfiability threshold conjecture for $k = 2$. We extend and generalize their proof ideas to non-uniform random 2-SAT. In order to show a lower bound on the threshold, we investigate the existence of bicycles. Bicycles were introduced by Chvatal and Reed. They are sub-formulas which appear in every unsatisfiable formula. We can show with a first moment argument, that these do not appear below a certain number of clauses, thus making formulas satisfiable.

In order to show an upper bound on the threshold, we investigate the existence of snakes. Snakes are unsatisfiable sub-formulas and have also been introduced by Chvatal and Reed. We can show with a second-moment argument that snakes of certain sizes do appear above a certain number of clauses, thus making formulas unsatisfiable. Unfortunately, this method does not work if the two largest variable probabilities are too large asymptotically. In that case we lower-bound the probability that an unsatisfiable sub-formula containing only those two variables exists. This can be done with a simple inclusion-exclusion argument and the resulting lemma also works for $k \geq 3$.

We will see that the threshold position and its sharpness depend on how the functions of the two highest variable probabilities $p_1$ and $p_2$ behave compared to the other variable probabilities. Moreover, it depends on the asymptotic behavior

of those values. The squares of $p_1$ and $p_2$ will be compared to the sum of squares of the other variable probabilities, $\sum_{i=1}^{n} p_i^2$ and $\sum_{i=2}^{n} p_i^2$. Note that these sums of squares are functions in $n$ as well. The conditions on those functions can be checked if we know the ensemble of probability distributions that our model uses. We are going to show that there are four cases depending on $p_1$, $p_2$, $\sum_{i=1}^{n} p_i^2$, and $\sum_{i=2}^{n} p_i^2$:

1. If $p_1^2 \in o\left(\sum_{i=1}^{n} p_i^2\right)$, then there is a sharp threshold at exactly
$$m^\star = \frac{1}{\sum_{i=1}^{n} p_i^2}.$$

2. If $p_1^2 \in \Theta\left(\sum_{i=1}^{n} p_i^2\right)$ and $p_2^2 \in o\left(\sum_{i=2}^{n} p_i^2\right)$, then the asymptotic threshold is at
$$m^\star \in \Theta\left(\frac{1 - \sum_{i=1}^{n} p_i^2}{p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}\right)$$
and it is coarse. The coarseness stems from the emergence of an unsatisfiable sub-formula with 3 variables and 4 clauses. Furthermore, we can show that there is a range of size $\Theta(m^\star)$ around the threshold in which the probability to generate satisfiable instances is a constant bounded away from zero and one in the limit.

3. If $p_1^2 \in \Theta\left(\sum_{i=1}^{n} p_i^2\right)$ and $p_2^2 \in \Theta\left(\sum_{i=2}^{n} p_i^2\right)$, then we can show that the asymptotic satisfiability threshold is at
$$m^\star \in \Theta\left(\frac{1 - \sum_{i=1}^{n} p_i^2}{p_1 \cdot p_2}\right),$$
which is proportional to $1/q_{\max}$, where $q_{\max}$ is the maximum clause probability. We can also show that this threshold is coarse. This time the coarseness stems from the emergence of an unsatisfiable sub-formula of size 4, which contains only the two most probable variables. Again, we can show that there is a range of size $\Theta(m^\star)$ around the threshold in which the probability to generate satisfiable instances is a constant bounded away from zero and one in the limit.

4. If none of the above cases apply, there is a threshold at
$$m^\star \in \Theta\left(\frac{1 - \sum_{i=1}^{n} p_i^2}{\sum_{i=2}^{n} p_i^2 + p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}\right).$$
The threshold is again coarse, but this time the probability in the threshold interval cannot be bounded.

It is important to understand why we want to show that in the second and third case there is a range of size $\Theta(m^\star)$ around the threshold in which the probability

to generate satisfiable instances can be bounded away from zero and one by a constant. This implies that the probability cannot approach zero or one for some functions $m$ that are only a constant factor away from the threshold. According to our definition of sharp thresholds (Definition 3.11), that means in those cases the threshold cannot be sharp, but must be coarse. We will later see that the statements for those two cases also imply coarseness of the threshold in the last case. Moreover, in all four cases the asymptotic threshold is at

$$m^\star \in \Theta\left(\frac{1 - \sum_{i=1}^{n} p_i^2}{\sum_{i=2}^{n} p_i^2 + p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}\right).$$

Together with the conditions on $p_1^2$ and $p_2^2$ this threshold function simplifies to the ones we stated in the first three cases, respectively. The four cases give us a complete dichotomy of coarseness and sharpness for the satisfiability threshold of non-uniform random 2-SAT. This result generalizes the seminal works by Chvátal and Reed [CR92] and by Goerdt [Goe96] to arbitrary ensembles of variable probability distributions and includes their findings as a special case (c. f. Section 4.6).

## 4.1 What we are going to show

First, we are going to discuss which kinds of results we are going to show and why we do not show something more intuitive. Our results will assume certain relations between the functions $m$ and $m^\star$, $p_1$ and $\sum_{i=1}^{n} p_i^2$, and $p_2$ and $\sum_{i=2}^{n} p_i^2$. Intuitively, those relations would be in terms of Landau notation as is suggested by the results we want to show for non-uniform random 2-SAT. However, we are only going to assume that functions are smaller or bigger than other functions by some constant factor that is either given or that we can choose. Instead of assuming $p_1^2 \in o(\sum_{i=1}^{n} p_i^2)$, we only assume that we can choose a constant $\varepsilon_1 \in (0, 1)$ so that $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$. This condition is implied by $p_1^2 \in o(\sum_{i=1}^{n} p_i^2)$ for sufficiently large $n$. Instead of assuming $p_1^2 \in \Theta(\sum_{i=1}^{n} p_i^2)$, we assume that there is some constant $\varepsilon_1 \in (0, 1)$ so that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$. Again, this condition is implied by $p_1^2 \in \Theta(\sum_{i=1}^{n} p_i^2)$ for sufficiently large $n$. Equivalently, we consistently use the factors $\varepsilon_2$ and $\varepsilon_m$ to define relationships between $p_2$ and $\sum_{i=2}^{n} p_i^2$, and $m$ and $m^\star$, respectively. Additionally, we will use the placeholder $\varepsilon$ without any index if we refer to relations from previous results that will only be used in a very local scope. This is mainly to avoid using the same notation twice.

We choose to use these requirements for our results, because it will allow us to prove something in absence of asymptotic behavior as well, i. e. in the fourth case above. Remember what we want to show: If we assume the existence of a

satisfiability threshold at some function $m^\star$, then

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{D}^N\left(n,2,\left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}},m\right)}[\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } m \in o_n(m^\star) \\ 0, & \text{if } m \in \omega_n(m^\star). \end{cases}$$

Let us concentrate on the case that $m \in o(m^\star)$. Remembering the definition of limits, we want for any constant $\varepsilon_P \in (0,1)$ that there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ we have

$$\Pr_{\Phi\sim\mathcal{D}^N\left(n,2,\left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}},m\right)}[\,\Phi \text{ satisfiable}\,] \geq \varepsilon_P.$$

We will now show the following:

1. Assume we can choose $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. Then for any given $\varepsilon_m \in (0,1)$ with $m \leq \varepsilon_m \cdot m^\star$ and for any given $\varepsilon_P \in (0,1)$, we can reach a probability of at least $\varepsilon_P$ by choosing $\varepsilon_1$ small enough.

2. Assume we are given $\varepsilon_1 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and we can choose $\varepsilon_m \in (0,1)$ with $m \leq \varepsilon_m \cdot m^\star$. Then, for any given $\varepsilon_P \in (0,1)$, we can choose $\varepsilon_m$ small enough to reach a probability of at least $\varepsilon_P$.

Let us now consider what these two results imply. Assume we are given some $\varepsilon_P \in (0,1)$ and some $m \in o(m^\star)$. We have to show that the probability to generate satisfiable instances at $m$ is at least $\varepsilon_P$ for all sufficiently large $n$. First we note that, if $m \in o(m^\star)$, then for all $\varepsilon_m \in (0,1)$ there is some $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ we have $m \leq \varepsilon_m \cdot m^\star$.

If $p_1^2 \in o(\sum_{i=1}^n p_i^2)$, then for every $\varepsilon_1 \in (0,1)$ there is some $n_0 \in \mathbb{N}$ so that $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ holds for all $n \geq n_0$. Due to the first result we can now simply choose some $\varepsilon_m \in (0,1)$ (for example $\varepsilon_m = \frac{1}{2}$) and choose $\varepsilon_1$ small enough so that the resulting probability is at least $\varepsilon_P$. The requirements $p_1^2 \in o(\sum_{i=1}^n p_i^2)$ and $m \in o(m^\star)$ guarantee that there is some $n_0 \in \mathbb{N}$ so that both requirements hold for all $n \geq n_0$.

If $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$, then there are $\varepsilon_1 \in (0,1)$ and $n_0 \in \mathbb{N}$ so that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ holds for all $n \geq n_0$. For this value of $\varepsilon_1$, we can now simply choose an $\varepsilon_m$ small enough so that the resulting probability is at least $\varepsilon_P$. Again, this requirement is fulfilled for all sufficiently large $n$, since $m \in o(m^\star)$.

The last case is that neither $p_1^2 \in o(\sum_{i=1}^n p_i^2)$ nor $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$. First, we assume to be able to choose $\varepsilon_1$. Like in the first case, we choose some $\varepsilon_m \in (0,1)$ (e. g. $\varepsilon_m = \frac{1}{2}$) and evaluate how small $\varepsilon_1$ has to be in order to have a probability of at least $\varepsilon_P$ due to our first result. For sufficiently large $n$ the requirement $m \leq \varepsilon_m \cdot m^\star$ will be fulfilled. However, $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ might not. For all values of $n$ that fulfill the requirement, we are done already. Thus, we only have to consider what happens if $p_1^2 > \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. For those values of $n$ we can simply use the second result and evaluate an $\varepsilon_m \in (0,1)$ small enough so that we reach

the given probability $\varepsilon_P$. Again, this requirement is fulfilled for sufficiently large $n$ due to $m \in o(m^\star)$.

This was a simplified example of what our results will look like and how we are going to use them to show the statements in the introduction of this chapter. We will show our results for slightly different functions $m^\star$. However, these functions will asymptotically coincide as we will see later.

### 4.1.1 How we are going to show it

Another note on how we will derive our results might be necessary at this point. As stated in Section 3.3, the probability to draw a certain clause is proportional to the product of variable probabilities for Boolean variables it contains. For example, the probability to draw a clause $c = (X_i \vee \overline{X_j})$ is

$$\Pr\left[(X_i \vee \overline{X_j})\right] = q_c = \frac{C}{2} \cdot p_i \cdot p_j, \tag{4.1}$$

where $C = 1/\left(k! \cdot \sum_{J \in \mathcal{P}_k(\{1,2,\dots,n\})} \prod_{j \in J} p_j\right)$ is the same for all clauses. However, for $k = 2$ the factor $C$ simplifies to $C = 1/\left(1 - \sum_{i=1}^n p_i^2\right)$.

Since clauses are also drawn independently, it holds that the probability of drawing a certain formula in non-uniform random $k$-SAT is proportional to the product of variable probabilities for each appearance of a Boolean variable in it. For example, the probability of drawing $\Phi = (X_i \vee \overline{X_j}) \wedge (\overline{X_h} \vee \overline{X_i})$ is

$$\Pr_{\Phi \sim \mathcal{D}^N\left(n,2,\left(\vec{p}^{\,(n)}\right)_{n\in\mathbb{N}},2\right)}\left[\Phi = (X_i \vee \overline{X_j}) \wedge (\overline{X_h} \vee \overline{X_i})\right] = \left(\frac{C}{2}\right)^2 \cdot p_h \cdot p_i^2 \cdot p_j.$$

We will use this fact heavily in our analysis. If we want to know the probability of drawing a formula, we only have to know which Boolean variables it contains how often. Furthermore, we can also use this fact if we do not know the exact variables, but only how often they appear. For example, if we are looking for a formula $\Phi = (\ell_i \vee \overline{\ell_j}) \wedge (\overline{\ell_h} \vee \overline{\ell_i})$, where $\ell_h$, $\ell_i$, and $\ell_j$ are literals of distinct Boolean variables, the probability is proportional to

$$\left(\sum_{h=1}^n p_h\right) \cdot \left(\sum_{\substack{i=1 \\ i\neq h}}^n p_i^2\right) \cdot \left(\sum_{\substack{j=1 \\ j\neq h,\ j\neq i}}^n p_j\right) \leq \left(\sum_{h=1}^n p_h\right) \cdot \left(\sum_{i=1}^n p_i^2\right) \cdot \left(\sum_{j=1}^n p_j\right).$$

The following lemma shows how we can bound expressions of that kind. It applies to situations where a set of variables all appear the same number of times and we already accounted for the possible ways to arrange them in clauses. For example, if we want the probability for a formula $\Phi = (\ell_h \vee \ell_i) \wedge (\ell_i \vee \ell_j) \wedge (\ell_j \vee \ell_h)$, where $\ell_h$, $\ell_i$, and $\ell_j$ are literals of distinct Boolean variables, the probability is

proportional to

$$3! \cdot \sum_{A \in \mathcal{P}_3([n])} \prod_{a \in A} p_a^2,$$

where 3! accounts for the possibilities to interchange the chosen variables.

▶ **Lemma 4.1.** For every set $S \subseteq \{1, \dots, n\}$, every integer $i \leq |S|$, and every integer $l \geq 1$ it holds that

$$\sum_{A \in \mathcal{P}_i(S)} \prod_{a \in A} p_a^l \leq \frac{1}{i!} \left( \sum_{s \in S} p_s^l \right)^i.$$

If $\left(\sum_{s \in S} p_s^l\right) \geq (i-1) \cdot \max\left\{p_s^l \mid s \in S\right\}$, then it also holds that

$$\sum_{A \in \mathcal{P}_i(S)} \prod_{a \in A} p_a^l \geq \frac{1}{i!} \left( \left( \sum_{s \in S} p_s^l \right) - (i-1) \cdot \max\left\{p_s^l \mid s \in S\right\} \right)^i.$$

◀

*Proof.* For the first part, notice that each product $\prod_{a \in A} p_a^l$ for some $A \in \mathcal{P}_i(S)$ appears $i!$-times in $\left(\sum_{s \in S} p_s^l\right)^i$. For the second part, $\sum_{A \in \mathcal{P}_i(S)} \prod_{a \in A} p_a^l$ can be expressed as the following nested sum

$$\sum_{A \in \mathcal{P}_i(S)} \prod_{a \in A} p_a^l = \frac{1}{i!} \cdot \sum_{a_1 \in A} \left( p_{a_1}^l \cdot \sum_{a_2 \in A \setminus \{a_1\}} \left( p_{a_2}^l \cdot \ldots \cdot \sum_{a_i \in A \setminus \{a_0, \ldots, a_{i-1}\}} p_{a_i}^l \right) \right).$$

This sum essentially captures the choices of elements we have for each term, where $a_j$ is the $j$-th chosen element for $j = 1, \dots, i$. Since we only forbid repetitions of elements, the $j$-th element can be anything from $S \setminus \{a_1, a_2, \dots, a_{j-1}\}$. Again, we generate each product $i!$ times on the right-hand side. If we pessimistically assume that forbidden elements have the maximum value $\max\left\{p_s^l \mid s \in S\right\}$, we get

$$\begin{aligned}
&\sum_{A \in \mathcal{P}_i(S)} \prod_{a \in A} p_a^l \\
&\geq \frac{1}{i!} \sum_{a_1 \in A} \left( p_{a_1}^l \cdot \sum_{a_2 \in A \setminus \{a_1\}} \left( p_{a_2}^l \cdot \ldots \cdot \left( \left( \sum_{s \in S} p_s^l \right) - (i-1) \cdot \max\left\{p_s^l \mid s \in S\right\} \right) \right) \right) \\
&\geq \frac{1}{i!} \left( \left( \sum_{s \in S} p_s^l \right) - (i-1) \cdot \max\left\{p_s^l \mid s \in S\right\} \right)^i.
\end{aligned}$$

Now we also see why the requirement for this second statement is necessary. ■

We will use the bounds of the former lemma heavily in the remainder of this thesis.

## 4.2 Bicycles and the First Moment Method

In this section we introduce the concept of bicycles and derive a lower bound on the position of the satisfiability threshold.

Chvátal and Reed [CR92] define the following sub-structure of 2-SAT formulas and show that every unsatisfiable formula in 2-CNF contains this substructure.

▶ **Definition 4.2 (bicycle).** Let $X_1, X_2, \ldots, X_t$ be $t$ distinct Boolean variables and let $w_1, w_2, \ldots, w_t$ be literals such that each $w_l$ is either $X_l$ or $\overline{X_l}$. We define a bicycle of length $t$ to be a sequence of $t+1$ clauses of the form

$$(u, w_1), (\overline{w_1}, w_2), \ldots, (\overline{w_{t-1}}, w_t), (\overline{w_t}, v),$$

where $u, v \in \{w_1, \ldots, w_t, \overline{w}_1, \ldots, \overline{w}_t\}$. ◀

Although a bicycle itself might not be unsatisfiable, Chvátal and Reed [CR92] prove that *every unsatisfiable Boolean formula in 2-CNF must contain a bicycle*. We can use this knowledge in the following way: We show that up to a certain number of clauses $m^\star$ the random formulas our model generates a. a. s. do not contain *any* bicycles. Thus, they must be satisfiable. In order to bound the probability for bicycles to appear, we use the first moment method. This means, we bound the expected number of bicycles that appear. If this number is $o(1)$, we can use Markov's inequality to bound the probability of them appearing as desired. The same approach was used in the proof of Theorem 3 from [CR92].

First, we consider the case $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. We define our threshold function to be $m^\star = (\sum_{i=1}^n p_i^2)^{-1}$. We want to show that this function defines a sharp satisfiability threshold for non-uniform random 2-SAT. Remember our definition of a sharp satisfiability threshold. We need to show that for any constant $\varepsilon_m \in (0,1)$ and all functions $m \leq \varepsilon_m \cdot m^\star$ the probability to generate a satisfiable instance is a function tending to one as $n$ increases. However, as we wrote in the last section, we are going to show something a bit more general. We will show that, given $\varepsilon_m \in (0,1)$ and $\varepsilon_P \in (0,1)$, we can choose a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2)$ small enough so that the probability to generate a satisfiable instance is at least $\varepsilon_P$. If $p_1^2 \in o\big(\sum_{i=1}^n p_i^2\big)$, then there is an $n_0 \in \mathbb{N}$ such that this condition holds for all $n \geq n_0$.

▶ **Lemma 4.3.** Given an ensemble of probability distributions $\big(\vec{p}^{\,(n)}\big)_{n\in\mathbb{N}}$. Let $m^\star = 1/\sum_{i=1}^n p_i^2$. Then, for any constant $\varepsilon_m \in (0,1)$ with $m \leq \varepsilon_m \cdot m^\star$ and any constant $\varepsilon_P \in (0,1)$ we can choose $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2)$ such that the function describing the probability to generate a satisfiable formula $\Phi \sim \mathcal{D}^N\big(n, 2, \big(\vec{p}^{\,(n)}\big)_{n\in\mathbb{N}}, m\big)$ is at least $\varepsilon_P$. ◀

*Proof.* To show this result, we show that the expected number of bicycles is at most $1 - \varepsilon_P$ for the setting we consider. The result then follows by Markov's inequality.

First, choose $n$ arbitrary, but fixed. We want to evaluate the value of the probability function for this value of $n$ and the number of clauses prescribed by

the clause function $m(n)$. We fix a set $S \subseteq [n]$ of variables to appear in a bicycle with $|S| = t \geq 2$ . The probability that a *specific* bicycle $B$ with these variables appears in $\Phi$ is

$$\Pr[B \text{ in } \Phi] \leq \underbrace{\binom{m}{t+1} \cdot (t+1)!}_{\text{positions of } B \text{ in } \Phi} \cdot \Pr[(u \vee w_1)] \cdot \Pr[(\overline{w_t} \vee v)] \cdot \prod_{h=1}^{t-1} \Pr[(\overline{w_h} \vee w_{h+1})].$$

$\Pr\big[(w_j \vee w_i)\big]$ denotes the probability to draw clause $(w_j \vee w_i)$ in non-uniform random 2-SAT. There are at most $t!$ possibilities to arrange the $t$ variables in a bicycle and $2^t$ possibilities to choose literals from the $t$ variables. For the probability that *any* bicycle with the variables from $S$ appears in $\Phi$ it now holds that

$$\Pr[S\text{-bicycle in } \Phi] \leq m^{t+1} \cdot t! \cdot 2^t \cdot \left(\frac{C}{2}\right)^{t+1} \cdot \prod_{i \in S} p_i^2 \cdot \left(2 \cdot \sum_{i \in S} p_i\right)^2,$$

where the last factor accounts for the possibilities to choose $u$ and $v$. It now holds that

$$\Pr[\text{bicycle in } \Phi] \leq \sum_{t=2}^{n} \left( \sum_{S \in \mathcal{P}_t([n])} \left( m^{t+1} \cdot t! \cdot 2^t \cdot \left(\frac{C}{2}\right)^{t+1} 2^2 \cdot \prod_{i \in S} p_i^2 \cdot \left(\sum_{i \in S} p_i\right)^2 \right) \right).$$

If we estimate $\sum_{i \in S} p_i \leq t \cdot p_1$, we get

$$\leq 2 \cdot \sum_{t=2}^{n} \left( (C \cdot m)^{t+1} \cdot t! \cdot t^2 \cdot p_1^2 \cdot \sum_{S \in \mathcal{P}_t([n])} \left( \prod_{i \in S} p_i^2 \right) \right)$$

and with $\sum_{S \in \mathcal{P}_t([n])} \left( \prod_{i \in S} p_i^2 \right) \leq \frac{1}{t!} \cdot \left( \sum_{i=1}^{n} p_i^2 \right)^t$ due to Lemma 4.1 this yields

$$\leq 2 \cdot \sum_{t=2}^{n} \left( (C \cdot m)^{t+1} \cdot t^2 \cdot p_1^2 \cdot \left( \sum_{i=1}^{n} p_i^2 \right)^t \right).$$

Since $m \leq \varepsilon_m \cdot m^\star = \frac{\varepsilon_m}{\sum_{i=1}^{n} p_i^2}$, this is

$$\leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^{n} p_i^2} \cdot \sum_{t=2}^{n} (C \cdot \varepsilon_m)^{t+1} \cdot t^2.$$

Now, it holds that $C = \frac{1}{1 - \sum_{i=1}^{n} p_i^2} \leq 1 + \frac{p_1}{1 - p_1}$, since $\sum_{i=1}^{n} p_i^2 \leq p_1$. Thus,

$$\Pr[\text{bicycle in } \Phi] \leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^n p_i^2} \cdot \sum_{t=2}^{n} \left( \left( \left(1 + \frac{p_1}{1-p_1}\right) \cdot \varepsilon_m \right)^{t+1} \cdot t^2 \right)$$
$$\leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^n p_i^2} \cdot \sum_{t=2}^{\infty} \left( \left( \left(1 + \frac{p_1}{1-p_1}\right) \cdot \varepsilon_m \right)^{t+1} \cdot t^2 \right).$$

We know that $p_1 \leq \sqrt{\varepsilon_1 \cdot \sum_{i=1}^n p_i^2} \leq \sqrt{\varepsilon_1}$. Thus, if we choose $\varepsilon_1$ small enough such that

$$\left(1 + \frac{p_1}{1-p_1}\right) \cdot \varepsilon_m \leq \left(1 + \frac{\sqrt{\varepsilon_1}}{1-\sqrt{\varepsilon_1}}\right) \cdot \varepsilon_m < 1,$$

then

$$\left( \sqrt{\left(1 + \frac{\sqrt{\varepsilon_1}}{1-\sqrt{\varepsilon_1}}\right) \cdot \varepsilon_m} \right)^{t+1} \cdot t^2 \in o(1).$$

Thus, there is some $t_0$ such that for all $t \geq t_0$ this function is at most 1. Therefore,

$$\Pr[\Phi \text{ contains a bicycle}] \leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^n p_i^2} \cdot \sum_{t=2}^{\infty} \left( \left( \left(1 + \frac{p_1}{1-p_1}\right) \cdot \varepsilon_m \right)^{t+1} \cdot t^2 \right)$$
$$\leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^n p_i^2} \cdot \left( t_0^3 + \sum_{t=t_0}^{\infty} \left( \sqrt{\left(1 + \frac{\sqrt{\varepsilon_1}}{1-\sqrt{\varepsilon_1}}\right) \cdot \varepsilon_m} \right)^{t+1} \right)$$
$$\leq 2 \cdot \varepsilon_1 \cdot \left( t_0^3 + \frac{\sqrt{(1 + \frac{\sqrt{\varepsilon_1}}{1-\sqrt{\varepsilon_1}}) \cdot \varepsilon_m}}{1 - \sqrt{(1 + \frac{\sqrt{\varepsilon_1}}{1-\sqrt{\varepsilon_1}}) \cdot \varepsilon_m}} \right),$$

where the second term was bounded by a geometric series. If we choose $\varepsilon_1$ sufficiently small, this expression is at most $1 - \varepsilon_P$. ■

We now turn to the case that $p_1^2 \notin o(\sum_{i=1}^n p_i^2)$. We are going to show that there is an asymptotic threshold at

$$m^\star = \left( C \cdot \left( \sum_{i=2}^{n} p_i^2 \right) + C \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{-1}.$$

However, we are going to show something a bit more general. We only assume that $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2)$ for some constant $\varepsilon_1 > 0$. If $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$, then there is some $n_0 \in \mathbb{N}$ such that this holds for all $n \geq n_0$. Under this condition, we will show that for any $\varepsilon_P \in (0,1)$ we can choose an $\varepsilon_m \in (0,1)$ with $m \leq \varepsilon_m \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ so that the probability to generate a satisfiable instance is at least $\varepsilon_P$. If $m \in o\big((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}\big)$, this condition is met

for all sufficiently large $n$. However, it is also met if $m \in o(m^\star)$. We will show this in more detail in Section 4.5.

▶ **Lemma 4.4.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ with $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2)$ for some constant $\varepsilon_1 \in (0,1)$. Let $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. Then, for any $\varepsilon_P \in (0,1)$ we can choose an $\varepsilon_m \in (0,1)$ such that the probability to generate a satisfiable formula $\Phi \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ is at least $\varepsilon_P$ if $m \leq \varepsilon_m \cdot m^\star$. ◀

*Proof.* As in the proof of Lemma 4.3 it holds that

$$\begin{aligned}\Pr[\Phi \text{ unsat}] &\leq \Pr[\text{bicycle in } \Phi] \\ &\leq \sum_{t=2}^{n}\left(\sum_{S\in\mathcal{P}_t([n])}\left(m^{t+1}\cdot t!\cdot 2^t\cdot\left(\frac{C}{2}\right)^{t+1} 2^2\cdot\prod_{i\in S}p_i^2\cdot\left(\sum_{i\in S}p_i\right)^2\right)\right) \\ &\leq 2\cdot\sum_{t=2}^{n}\left((C\cdot m)^{t+1}\cdot t!\cdot\sum_{S\in\mathcal{P}_t([n])}\left(\prod_{i\in S}p_i^2\right)\cdot\left(\sum_{i\in S}p_i\right)^2\right). \qquad (4.2)\end{aligned}$$

We can analyze the term $\sum_{S\in\mathcal{P}_t([n])}\left(\left(\prod_{i\in S}p_i^2\right)\cdot\left(\sum_{i\in S}p_i\right)^2\right)$ in more detail. By doing a case distinction between the terms with $p_1 \in S$ and $p_1 \notin S$ we get

$$\begin{aligned}&\sum_{S\in\mathcal{P}_t([n])}\left(\left(\prod_{i\in S}p_i^2\right)\cdot\left(\sum_{i\in S}p_i\right)^2\right) \\ &\leq p_1^2\cdot t^2\cdot p_1^2\cdot\frac{1}{(t-1)!}\cdot\left(\sum_{i=2}^{n}p_i^2\right)^{t-1} + t^2\cdot p_2^2\cdot\frac{1}{t!}\cdot\left(\sum_{i=2}^{n}p_i^2\right)^{t}\end{aligned}$$

and since $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2) \geq \varepsilon_1 \cdot (\sum_{i=2}^n p_i^2)$ and $p_2 \leq p_1$ this yields

$$\leq (1+1/\varepsilon_1)\cdot t^3\cdot p_1^4\cdot\frac{1}{t!}\cdot\left(\sum_{i=2}^{n}p_i^2\right)^{t-1}.$$

It holds that $p_1^4 \cdot \left(\sum_{i=2}^n p_i^2\right)^{t-1} \leq \left(\frac{1}{\sqrt{\varepsilon_1}}\cdot p_1\cdot\left(\sum_{i=2}^n p_i^2\right)^{1/2}\right)^{t+1}$ for $t \geq 3$. This yields

$$\sum_{S\in\mathcal{P}_t([n])}\left(\left(\prod_{i\in S}p_i^2\right)\cdot\left(\sum_{i\in S}p_i\right)^2\right) \leq (1+1/\varepsilon_1)\cdot\frac{t^3}{t!}\cdot\left(\frac{1}{\sqrt{\varepsilon_1}}\cdot p_1\cdot\left(\sum_{i=2}^{n}p_i^2\right)^{1/2}\right)^{t+1} \qquad (4.3)$$

for $t \geq 3$.

For $t = 2$ we know that each of the three 2-clauses in the bicycle must contain both variables. Thus,

$$\sum_{S\in\mathcal{P}_2([n])}\Pr[S\text{-bicycle in } F]$$

$$\leq m^3 \cdot (C/2)^3 \cdot 2! \cdot 2^2 \cdot 2^2 \sum_{\substack{i,j \in V: \\ i \neq j}} p_i^3 \cdot p_j^3$$

$$\leq (C \cdot m)^3 \cdot t^2 \cdot p_1^3 \cdot \left( \sum_{i=2}^{n} p_i^3 \right) + (C \cdot m)^3 \cdot t^2 \left( \sum_{i=2}^{n} p_i^3 \right)^2$$

and since $\sum_{i=2}^{n} p_i^3 \leq \left( \sum_{i=2}^{n} p_i^2 \right)^{3/2}$ due to the monotonicity of vector norms, this is at most

$$\leq (C \cdot m)^3 \cdot t^2 \cdot p_1^3 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{3/2} + (C \cdot m)^3 \cdot t^2 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^3$$

and due to our condition $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^{n} p_i^2) \geq \varepsilon_1 \cdot (\sum_{i=2}^{n} p_i^2)$, we get

$$\leq \left( 1 + \frac{1}{\varepsilon_1^{3/2}} \right) \cdot (C \cdot m)^3 \cdot t^2 \cdot p_1^3 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{3/2}$$

$$\leq (C \cdot m)^{t+1} \cdot (1 + 1/\varepsilon_1) \cdot t^3 \cdot \left( \frac{1}{\sqrt{\varepsilon_1}} \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{t+1}. \tag{4.4}$$

We can now plug equation (4.3) and equation (4.4) into equation (4.2) to get

$$\Pr[\Phi \text{ unsat}] \leq 2 \cdot (1 + 1/\varepsilon_1) \sum_{t=2}^{n} \left( \left( C \cdot m \cdot \frac{1}{\sqrt{\varepsilon_1}} \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{t+1} \cdot t^3 \right)$$

$$\leq 2 \cdot (1 + 1/\varepsilon_1) \sum_{t=2}^{\infty} \left( \left( C \cdot m \cdot \frac{1}{\sqrt{\varepsilon_1}} \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{t+1} \cdot t^3 \right).$$

We can now choose $m \leq \varepsilon_m \cdot m^\star$ for a constant $\varepsilon_m \in (0, 1)$ to be determined later. Then,

$$\Pr[\Phi \text{ unsat}] \leq 2 \cdot (1 + 1/\varepsilon_1) \sum_{t=2}^{\infty} \left( \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \right)^{t+1} \cdot t^3$$

$$\leq 2 \cdot (1 + 1/\varepsilon_1) \cdot \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \sum_{t=2}^{\infty} \left( \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \right)^{t} \cdot t^3.$$

If we choose $\varepsilon_m$ small enough so that $\frac{\varepsilon_m}{\sqrt{\varepsilon_1}} < 1$, it holds that $\left( \sqrt{\frac{\varepsilon_m}{\sqrt{\varepsilon_1}}} \right)^t \cdot t^3 = o(1)$. Thus, there is a $t_0 \in \mathbb{N}$ so that this function is at most 1 for all $t \geq t_0$. As in the

proof of Lemma 4.3, we have

$$\begin{aligned}\Pr[\Phi \text{ unsat}] &\leq 2 \cdot (1 + 1/\varepsilon_1) \cdot \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \cdot \sum_{t=2}^{\infty} \left(\frac{\varepsilon_m}{\sqrt{\varepsilon_1}}\right)^t \cdot t^3 \\ &\leq 2 \cdot (1 + 1/\varepsilon_1) \cdot \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \cdot \left(t_0^4 + \sum_{t=t_0}^{\infty} \left(\sqrt{\frac{\varepsilon_m}{\sqrt{\varepsilon_1}}}\right)^t\right) \\ &\leq 2 \cdot (1 + 1/\varepsilon_1) \cdot \frac{\varepsilon_m}{\sqrt{\varepsilon_1}} \cdot \left(t_0^4 + \frac{\sqrt{\frac{\varepsilon_m}{\sqrt{\varepsilon_1}}}}{1 - \sqrt{\frac{\varepsilon_m}{\sqrt{\varepsilon_1}}}}\right).\end{aligned}$$

We can now choose $\varepsilon_m$ small enough so that this probability is at most $1 - \varepsilon_P$. ■

Lemma 4.3 and Lemma 4.4 imply the statements we want for $m \leq \varepsilon_m \cdot m^\star$ and $m \in o(m^\star)$ respectively. We will show this formally in Section 4.5.

## 4.3 Snakes and the Second Moment Method

The two lemmas from the previous section provide a lower bound on the satisfiability threshold for non-uniform random 2-SAT. By using the second moment method, we can also derive an upper bound. This proof is inspired by Chvatal and Reed [CR92, Theorem 4], who provide us with the following definition.

▶ **Definition 4.5 (snake).** A *snake* of size $t \geq 2$ is a sequence of literals $(w_1, w_2, \ldots, w_{2t-1})$ over distinct variables. Each snake $A$ is associated with a set $F_A$ of $2t$ clauses $(\overline{w_i}, w_{i+1})$, $0 \leq i \leq 2t-1$, such that $w_0 = w_{2t} = \overline{w_t}$. ◀

We will also call the variable $|w_t|$ of a snake its *central* variable. Note that the set of clauses $F_A$ defined by a snake $A$ is unsatisfiable. Also, the snakes

$$\begin{aligned}&(w_1, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_s),\\ &(\overline{w_{t-1}}, \overline{w_{t-2}}, \ldots, \overline{w_1}, w_t, w_{t+1}, \ldots, w_s),\\ &(w_1, \ldots, w_{t-1}, w_t, \overline{w_s}, \overline{w_{s-1}}, \ldots, \overline{w_{t+1}}), \text{ and}\\ &(\overline{w_{t-1}}, \overline{w_{t-2}}, \ldots, \overline{w_1}, w_t, \overline{w_s}, \overline{w_{s-1}}, \ldots, \overline{w_{t+1}})\end{aligned}$$

create the same set of clauses.

The *variable-variable incidence graph (VIG)* for a formula $\Phi$ is a simple graph $G_\Phi = (V_\Phi, E_\Phi)$ with $V_\Phi$ consisting of all variables appearing in $\Phi$ and two variables being connected by an edge if they appear together in at least one clause of $\Phi$. An example for a snake's VIG can be seen in Figure 4.1. We will use this representation later in the proof of Lemma 4.12.

In order to show our upper bounds, we will prove that snakes of a certain length $t$ appear with sufficiently high probability in a random formula $\Phi \sim$

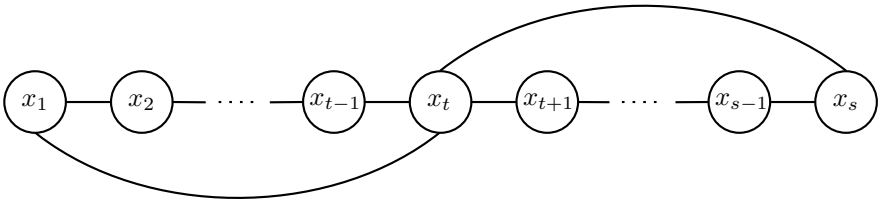



**Figure 4.1:** Variable-variable-incidence graph of a snake $w_1, w_2, \ldots, w_s$ where $|w_i| = x_i$ (the variable of the literal $w_i$) for $1 \leq i \leq s = 2t - 1$.

$\mathcal{D}(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$. To this end we utilize the second moment method: If $X \geq 0$ is a random variable with finite variance, then

$$\Pr[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

We define the following indicator variables for each snake $A$ of size $t$

$$X_A = \begin{cases} 1 & \text{if } F_A \text{ appears exactly once in } \Phi \\ 0 & \text{otherwise} \end{cases}$$

and their sum $X_t = \sum_A X_A$. Throughout the rest of this chapter we let $X_t$ denote the number of snakes of size $t \geq 2$ whose associated clauses appear exactly once in a non-uniform random 2-SAT formula $\Phi \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$.

As before, if we define $t\colon \mathbb{N} \to \mathbb{R}^+$ as a function in $n$, it holds that $\mathbb{E}\left[X_t^2\right]$ and $\mathbb{E}[X_t]$ are functions in $n$ as well. For carefully chosen functions $t$ we will show that $\mathbb{E}\left[X_t^2\right] \leq (1 + \varepsilon_E) \cdot \mathbb{E}[X_t]^2$ for a sufficiently small constant $\varepsilon_E > 0$. Note that for the probability to generate an unsatisfiable instance, it is sufficient to show a large enough lower bound for *any* value of $t$. Thus, we will consider several values of $t$, one of which is guaranteed to give us a bound as desired for sufficiently large values of $n$. More precisely, there are two values of $t$ that are relevant for us, $t = 2$ and $t = f^{1/78}$, where we define

$$f = \frac{\sum_{i=1}^n p_i^2}{p_1^2}.$$

Note that $t$ and $f$ are both functions in $n$ as are $\left(\sum_{i=1}^n p_i^2\right)$ and $p_1$.

$t = 2$ will provide the desired result if there is a constant $\varepsilon_1 \in (0, 1)$ such that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and if we can choose a sufficiently small $\varepsilon_2 \in (0, 1)$ such that $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$. This especially includes the case $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in o(\sum_{i=2}^n p_i^2)$.

$t = f^{1/78}$ will provide the desired result if we can choose a sufficiently small $\varepsilon_1 \in (0, 1)$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. This includes the case $p_1^2 \in o(\sum_{i=1}^n p_i^2)$.

However, if there are constants $\varepsilon_1, \varepsilon_2 \in (0, 1)$ so that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and

$p_2^2 \geq \varepsilon_1 \cdot \sum_{i=2}^{n} p_i^2$, we can show a lower bound directly without having to use the second moment method. We will handle this case in Section 4.4.

Now, if we want to use the second moment method, we first have to ensure that the expected number of snakes of a certain size is large enough. The following lemma provides a lower bound on this expected number.

▶ **Lemma 4.6.** Let $2 \leq t \leq (2 \cdot q_{\max})^{-1}$. Then it holds that

$$\mathbb{E}[X_t] \geq \frac{1}{2} \cdot (m-2t)^{2t} \cdot C^{2t} \cdot (1 - 2t \cdot q_{\max} \cdot m) \cdot \left(\sum_{i=1}^{n} p_i^4\right) \cdot \left(\sum_{i=2}^{n} \left(p_i^2 - (2t-3) \cdot p_2^2\right)\right)^{2t-2}.$$

◀

*Proof.* It holds that

$$\mathbb{E}[X_t] = \sum_{\substack{\text{snake} \\ A=(w_1,\dots,w_{2t-1})}} \left(\binom{m}{2t} \cdot (2t)! \cdot \prod_{i=0}^{2t-1} \Pr[\overline{w_i}, w_{i+1}] \cdot \left(1 - \sum_{c \in F_A} \Pr[c]\right)^{m-2t}\right)$$

and according to equation (4.1) it holds that $\Pr[(\overline{w_i}, w_{i+1})] = \frac{C}{2} \cdot p(|w_i|) \cdot p(|w_{i+1}|)$. Together with the fact that $\sum_{c \in F_A} \Pr[c] \leq 2t \cdot q_{\max}$, we get

$$\mathbb{E}[X_t] \geq (m-2t)^{2t} (1 - 2t \cdot q_{\max})^{m-2t} \left(\frac{C}{2}\right)^{2t} \cdot \sum_{\substack{\text{snake} \\ A=(w_1,\dots,w_{2t-1})}} \left(p(|w_t|)^4 \cdot \prod_{\substack{i=1 \\ i \neq t}}^{2t-1} p(|w_i|)^2\right). \tag{4.5}$$

Now we count how many snakes of size $t$ there are. First, we choose a central variable $X_j$. Then, we choose a set $S$ of $2t-2$ *different variables*. From those variables we can create $2^{2t-1} \cdot (2t-2)!$ different snakes by choosing signs for the $2t-1$ variables and by permuting the order of the $2t-2$ non-central variables.

$$\sum_{\substack{\text{snake} \\ A=(w_1,\dots,w_{2t-1})}} \left(p(|w_t|)^4 \cdot \prod_{\substack{i=1 \\ i \neq t}}^{2t-1} p(|w_i|)^2\right) \geq 2^{2t-1}(2t-2)! \cdot \sum_{j=1}^{n} \left(p_j^4 \cdot \sum_{\substack{S \subseteq [n] \setminus \{j\}: \\ |S|=2t-2}} \prod_{s \in S} p_s^2\right).$$

Due to Lemma 4.1 we have

$$\sum_{\substack{S \subseteq [n] \setminus \{j\}: \\ |S|=2t-2}} \prod_{s \in S} p_s^2 \geq \frac{1}{(2t-2)!} \left(\sum_{i=2}^{n} \left(p_i^2 - (2t-3) \cdot p_2^2\right)\right)^{2t-2}$$

and thus

$$\sum_{\substack{\text{snake} \\ A=(w_1,\ldots,w_{2t-1})}} \left( p(|w_t|)^4 \cdot \prod_{\substack{i=1 \\ i\neq t}}^{2t-1} p(|w_i|)^2 \right) \geq 2^{2t-1} \cdot \left( \sum_{j=1}^{n} p_j^4 \right) \cdot \left( \sum_{i=2}^{n} \left( p_i^2 - (2t-3) \cdot p_2^2 \right) \right)^{2t-2}. \quad (4.6)$$

Due to Bernoulli's inequality, it also holds that

$$(1 - 2t \cdot q_{\max})^{m-2t} \geq (1 - 2t \cdot q_{\max} \cdot (m - 2t)), \quad (4.7)$$

if $2t \cdot q_{\max} \leq 1$. Plugging equation (4.6) and equation (4.7) into equation (4.5) we get the result as desired. ■

### 4.3.1 The coarse threshold case

We want to prove an upper bound on the non-uniform random 2-SAT threshold. To get to know the proof technique, we start with the much simpler case that there is a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ and that we can choose a sufficiently small $\varepsilon_2 \in (0,1)$ such that $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$. For this case, we set

$$m^\star = \left( C \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{-1}.$$

In order to show the desired result, we need the following lower bound on $m^\star$.

▶ **Lemma 4.7.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ so that there is a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ and a constant $\varepsilon_2 \in (0,1)$ so that $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$. Let $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$. Then,

$$m^\star \geq \frac{1 - \varepsilon_2^{1/2}}{\varepsilon_2^{1/4}}.$$

◀

*Proof.* First, we notice

$$\sum_{i=2}^{n} p_i^2 \leq p_2 \cdot \sum_{i=2}^{n} p_i.$$

Due to the requirement $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$, we get

$$\leq \varepsilon_2^{1/2} \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \cdot \sum_{i=2}^{n} p_i.$$

The monotonicity of vector norms yields $\left(\sum_{i=2}^{n} p_i^2\right)^{1/2} \leq \sum_{i=2}^{n} p_i$ and thus

$$\begin{aligned} &\leq \varepsilon_2^{1/2} \cdot \left( \sum_{i=2}^{n} p_i \right)^2 \\ &= \varepsilon_2^{1/2} \cdot (1 - p_1)^2. \end{aligned}$$

It holds that

$$m^\star = \frac{1 - \sum_{i=1}^{n} p_i^2}{p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}.$$

We can now use the inequality $\sum_{i=2}^{n} p_i^2 \leq \varepsilon_2^{1/2} \cdot (1 - p_1)^2$ to get

$$\begin{aligned} &\geq \frac{1 - p_1^2 - \sum_{i=2}^{n} p_i^2}{p_1 \cdot \varepsilon_2^{1/4} \cdot (1 - p_1)} \\ &\geq \frac{1 - p_1^2 - \varepsilon_2^{1/2} \cdot (1 - p_1)^2}{p_1 \cdot \varepsilon_2^{1/4} \cdot (1 - p_1)} \\ &= \frac{(1 - p_1) \cdot (1 + p_1 - \varepsilon_2^{1/2} \cdot (1 - p_1))}{p_1 \cdot \varepsilon_2^{1/4} \cdot (1 - p_1)} \\ &= \frac{1 + p_1 - \varepsilon_2^{1/2} \cdot (1 - p_1)}{p_1 \cdot \varepsilon_2^{1/4}} > \frac{1 - \varepsilon_2^{1/2}}{\varepsilon_2^{1/4}}. \end{aligned}$$

■

The former lemma especially says that $m^\star$ can be arbitrarily large if $\varepsilon_2 \in (0, 1)$ is sufficiently small.

We want to show that for any constant $\varepsilon_m > 0$ at $m \geq \varepsilon_m \cdot m^\star$ there is a constant $\varepsilon_P \in (0, 1)$ such that the probability that a randomly generated instance contains a snake of size $t = 2$ is at least $\varepsilon_P > 0$. In that case, the only degree of freedom we have is choosing a constant $\varepsilon_2$ arbitrarily small. Together with our previous results this implies that the probability to generate an unsatisfiable instance is a constant bounded away from zero and one at $m \in \Theta(m^\star)$. However, if we can also choose $\varepsilon_m > 0$ arbitrarily large, we can show that this result holds for any constant $\varepsilon_P \in (0, 1)$. This implies that the probability to generate an unsatisfiable instance approaches one if $m \in \omega(m^\star)$.

In order to derive those results, we first show a lower bound on the expected number of snakes of size $t = 2$. Let us discuss what our lemma is going to state. We assume that there is an $\varepsilon_1 \in (0, 1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ and that we can choose

$\varepsilon_2 \in (0,1)$ arbitrarily small. This setting captures the case $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in o(\sum_{i=2}^n p_i^2)$. We want to show a result for all functions $m \in \Omega(m^\star)$. Thus, we assume $m = \varepsilon_m \cdot m^\star$ for some $\varepsilon_m > 0$. We can show that, given $\varepsilon_1$ and $\varepsilon_m$, we can choose $\varepsilon_2$ small enough such that $\mathbb{E}[\, X_2 \,] > \varepsilon_E$ for any constant $\varepsilon_E < \varepsilon_m^4$. This will imply $\mathbb{E}[\, X_2 \,] \in \Omega(1)$ later. If $\varepsilon_1$ and some $\varepsilon_E > 0$ are given and we can choose $\varepsilon_m$ and $\varepsilon_2$, then we can show that we can choose those values such that $\mathbb{E}[\, X_2 \,] > \varepsilon_E$ for any $\varepsilon_E$ given. This will imply $\mathbb{E}[\, X_2 \,] \in \omega(1)$ later. However, we will show that these results only hold for $m = \varepsilon_m \cdot m^\star$. These are also the values of $m$ for which we will show bounds on the probability to generate unsatisfiable instances. For higher values of $m$, for example for $m \in \omega(m^\star)$, these bounds still hold due to the monotonicity of unsatisfiability in non-uniform random $k$-SAT (c. f. Lemma 3.8).

▶ **Lemma 4.8.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ so that there is a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and let $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. The following statements hold:

1. Given a constant $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ and a constant $\varepsilon_E \in (0,1)$, then we can choose a constant $\varepsilon_2 \in (0,1)$ with $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ such that
$$\mathbb{E}[\, X_2 \,] \geq (1-\varepsilon_E) \cdot \frac{1}{2} \cdot m^4 \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^n p_i^2\right)^2.$$
2. Given a constant $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$, then we can choose a constant $\varepsilon_2 \in (0,1)$ such that $\mathbb{E}[\, X_2 \,] \geq \varepsilon_E$ for any constant $\varepsilon_E \in (0, \frac{1}{2} \cdot \varepsilon_m^4)$.
3. Given a constant $\varepsilon_E > 0$, then we can choose a constant $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ sufficiently large and a constant $\varepsilon_2 \in (0,1)$ sufficiently small such that $\mathbb{E}[\, X_2 \,] \geq \varepsilon_E$.

◀

*Proof.* For the first statement, note that

$$(4 \cdot q_{\max})^{-1} = \frac{1}{4} \cdot \frac{1}{C \cdot p_1 \cdot p_2} = \frac{1}{4} \cdot \frac{\left(\sum_{i=2}^n p_i^2\right)^{1/2}}{p_2} m^\star \geq \frac{1}{4 \cdot \varepsilon_2^{1/2}} \cdot m^\star > \frac{1-\varepsilon_2^{1/2}}{4 \cdot \varepsilon_2^{3/4}} \quad (4.8)$$

due to Lemma 4.7 This means, we can choose $\varepsilon_2$ small enough, such that $t = 2 \leq (2 \cdot q_{\max})^{-1}$. This allows us to use Lemma 4.6 with $t = 2$, which yields

$$\mathbb{E}[\, X_2 \,] \geq \frac{1}{2} \cdot (m-4)^4 \cdot C^4 \cdot (1 - 4 \cdot q_{\max} \cdot m) \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=2}^n p_i^2 - p_2^2\right)^2.$$

We now get

$$\left(\sum_{i=2}^{n} p_i^2 - p_2^2\right)^2 \geq \left(\sum_{i=2}^{n} p_i^2\right)^2 \cdot \left(1 - \frac{p_2^2}{\sum_{i=2}^{n} p_i^2}\right)^2 \geq (1-\varepsilon_2)^2 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2,$$

where we used $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$. Equivalently,

$$(m-4)^4 \geq m^4 \cdot \left(1 - \frac{4}{m}\right)^4 \geq m^4 \cdot \left(1 - \frac{4 \cdot \varepsilon_2^{1/4}}{\varepsilon_m \cdot (1-\varepsilon_2^{1/2})}\right)^4,$$

which holds since $m = \varepsilon_m \cdot m^\star \geq \varepsilon_m \cdot \frac{1-\sqrt{\varepsilon_2}}{\varepsilon_2^{1/4}}$ due to Lemma 4.7. Since $m = \varepsilon_m \cdot m^\star$ and due to equation (4.8) we get

$$1 - 4 \cdot q_{\max} \cdot m \geq 1 - \frac{4 \cdot \varepsilon_2^{1/2} \cdot m}{m^\star} = 1 - 4 \cdot \varepsilon_2^{1/2} \cdot \varepsilon_m.$$

Since $\left(\sum_{i=1}^{n} p_i^4\right) \geq p_1^4$, the expected value now simplifies to

$$\begin{aligned}
&\mathbb{E}[X_2] \\
&\geq \left(1 - \frac{4 \cdot \varepsilon_2^{1/4}}{\varepsilon_m \cdot (1-\varepsilon_2^{1/2})}\right)^4 \cdot (1-\varepsilon_2)^2 \cdot \left(1 - 4 \cdot \varepsilon_2^{1/2} \cdot \varepsilon_m\right) \cdot \frac{m^4}{2} \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2.
\end{aligned}$$

We can see that for any choice of $\varepsilon_m$ and $\varepsilon_1$, the leading factor gets closer to one as $\varepsilon_2$ gets closer to zero. Thus, for any $\varepsilon_m > 0$, $\varepsilon_1 \in (0,1)$, and $\varepsilon_E \in (0,1)$ we can choose a sufficiently small $\varepsilon_2$ to guarantee

$$\mathbb{E}[X_2] \geq (1-\varepsilon_E) \cdot \frac{1}{2} \cdot m^4 \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2.$$

This establishes the first statement.

For the second statement, suppose we are given an $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$. Then,

$$\begin{aligned}
\mathbb{E}[X_2] &\geq (1-\varepsilon) \cdot \frac{1}{2} \cdot m^4 \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2 \\
&= (1-\varepsilon) \cdot \frac{1}{2} \cdot (m/m^\star)^4 = (1-\varepsilon) \cdot \frac{1}{2} \cdot \varepsilon_m^4
\end{aligned}$$

for some constant $\varepsilon$ that decreases with decreasing $\varepsilon_2$. The smaller we choose $\varepsilon_2$, the closer this function gets to $\frac{1}{2} \cdot \varepsilon_m^4$. Thus, for any $\varepsilon_E \in (0, \frac{1}{2} \cdot \varepsilon_m^4)$ we can achieve $\mathbb{E}[X_2] \geq \varepsilon_E$. This establishes the second statement.

For the third statement suppose we are given an $\varepsilon_E > 0$ and we can choose

$\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$. Again, we get

$$\mathbb{E}[\, X_2 \,] \geq (1 - \varepsilon) \cdot \frac{1}{2} \cdot \varepsilon_m^4$$

for some constant $\varepsilon$ that decreases for fixed $\varepsilon_m$ and decreasing $\varepsilon_2$. First, we choose $\varepsilon_m$ such that $\frac{1}{2} \cdot \varepsilon_m^4 > \varepsilon_E$. Now we know that we can make $\varepsilon_2$ small enough so that the expected value is at least $\varepsilon_E$. ■

We are now ready to prove that random formulas are unsatisfiable with some positive constant probability at $m \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1})$. More precisely, we will show that, given $\varepsilon_1$ and $\varepsilon_m$, we can choose an $\varepsilon_2$ sufficiently small such that there is a constant $\varepsilon_P \in (0, 1)$ which bounds the probability to generate unsatisfiable instances from below. Moreover, this value $\varepsilon_P$ depends only on $\varepsilon_1$ and $\varepsilon_m$ and not on $n$. This means, this lower bound does not approach zero or one as $n$ increases.

In the proof we consider $\Pr[X_A = 1 \wedge X_B = 1]$, the probability that both snake $A$ and snake $B$ appear exactly once in a random formula. We distinguish several cases depending on how many clauses $F_A$ and $F_B$ have in common. Then, we analyze the probability of $F_A \cup F_B$ appearing exactly once. In order to do so, we assume that some snake $A$ and the shared clauses of $A$ and $B$ have already been chosen. Then, we construct $B$, incorporating the shared clauses from $A$.

▶ **Lemma 4.9.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n \in \mathbb{N}}$ and a constant $\varepsilon_1 \in (0, 1)$ with $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^{n} p_i^2)$. Let $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$. Then, for any constant $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ and any

$$\varepsilon_P < \frac{\varepsilon_m^4}{\varepsilon_m^4 + 3 \cdot \varepsilon_m^2 \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + 8}$$

we can choose a constant $\varepsilon_2 \in (0, 1)$ with $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$ such that the probability to generate an unsatisfiable formula $\Phi \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ is at least $\varepsilon_P$. ◀

*Proof.* First, we want to show that given $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ and $\varepsilon_1 \in (0, 1)$ with $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^{n} p_i^2)$, there is an $\varepsilon_P \in (0, 1)$ such that

$$\Pr[X_2 > 0] \geq \frac{\mathbb{E}[\, X_2 \,]^2}{\mathbb{E}\left[\, X_2^2 \,\right]} \geq \varepsilon_P.$$

Since Lemma 4.8 gives us a lower bound on $\mathbb{E}[\, X_2 \,]$, we only need to consider $\mathbb{E}\left[\, X_2^2 \,\right]$ now. We use the same approach as Chvátal and Reed [CR92] and split the expected value into two parts as follows

$$\mathbb{E}\left[\, X_2^2 \,\right] = \sum_{A} \sum_{B} \Pr[X_A = 1 \wedge X_B = 1]$$

$$= \sum_A \left( \sum_{B\colon B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] + \sum_{B\colon B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \right),$$

where $B \sim A$ denotes $F_A \cap F_B \neq \emptyset$. We will show that the part for $B \nsim A$ is at most $(1+\varepsilon_E) \cdot \mathbb{E}[X_2]^2$ for some arbitrarily small constant $\varepsilon_E > 0$ and that there is a constant $\varepsilon_F > 0$ such that the other part is at most $\varepsilon_F \cdot \mathbb{E}[X_2]^2$.

First let us consider the part for $B \nsim A$. It holds that

$$\Pr[X_A = 1 \wedge X_B = 1] = \binom{m}{8} \cdot 8! \cdot \left( \prod_{c \in F_A} \Pr[c] \right) \cdot \left( \prod_{c \in F_B} \Pr[c] \right) \cdot \left( 1 - \sum_{c \in F_A \cup F_B} \Pr[c] \right)^{m-8},$$

while

$$\Pr[X_A = 1] = \binom{m}{4} \cdot 4! \cdot \left( \prod_{c \in F_A} \Pr[c] \right) \cdot \left( 1 - \sum_{c \in F_A} \Pr[c] \right)^{m-4}. \tag{4.9}$$

Since $\binom{m}{8} \cdot 8! \leq \left(\binom{m}{4} \cdot 4!\right)^2$ this readily implies

$$\Pr[X_A = 1 \wedge X_B = 1] \leq \Pr[X_A = 1] \cdot \Pr[X_B = 1] \frac{\left(1 - \sum_{c \in F_A \cup F_B} \Pr[c]\right)^{m-8}}{\left(1 - \sum_{c \in F_A} \Pr[c]\right)^{m-4} \left(1 - \sum_{c \in F_B} \Pr[c]\right)^{m-4}}$$

and, due to $\left(1 - \sum_{c \in F_A} \Pr[c]\right) \cdot \left(1 - \sum_{c \in F_B} \Pr[c]\right) \geq 1 - \sum_{c \in F_A \cup F_B} \Pr[c]$, we have

$$\Pr[X_A = 1 \wedge X_B = 1] \leq \Pr[X_A = 1] \cdot \Pr[X_B = 1] \cdot \left(1 - \sum_{c \in F_A} \Pr[c]\right)^{-4} \left(1 - \sum_{c \in F_B} \Pr[c]\right)^{-4}.$$

Again, we can use Bernoulli's inequality to show

$$\begin{aligned} \left(1 - \sum_{c \in F_A} \Pr[c]\right)^4 \left(1 - \sum_{c \in F_B} \Pr[c]\right)^4 &\geq (1 - 4 \cdot q_{\max})^8 \\ &\geq 1 - 32 \cdot q_{\max} \\ &\geq 1 - \frac{32 \cdot \varepsilon_2^{3/4}}{1 - \sqrt{\varepsilon_2}}, \end{aligned}$$

where the last inequality follows with $q_{\max} < \frac{\varepsilon_2^{3/4}}{1-\sqrt{\varepsilon_2}}$ due to equation (4.8). For any fixed $\varepsilon_m$ this expression can be made arbitrarily close to one if we choose a

sufficiently small $\varepsilon_2$. This establishes

$$\begin{aligned}
&\sum_A \sum_{B\colon B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq \frac{1}{1 - 32 \cdot q_{\max}} \sum_A \sum_{B\colon B \nsim A} \Pr[X_A = 1] \cdot \Pr[X_B = 1] \\
&\leq \frac{1 - \sqrt{\varepsilon_2}}{1 - \sqrt{\varepsilon_2} - 32 \cdot \varepsilon_2^{3/4}} \cdot \mathbb{E}[X_2]^2 \\
&= (1 + \varepsilon_E) \cdot \mathbb{E}[X_2]^2
\end{aligned} \tag{4.10}$$

for a constant $\varepsilon_E$ that we can make arbitrarily small by making $\varepsilon_2$ small enough.

Now we turn to the case that $B \sim A$. We want to show that there is a constant $\varepsilon_F > 0$ such that this second sum is at most $\varepsilon_F \cdot \mathbb{E}[X_2]^2$. Let $l = |F_A \cap F_B|$. The first and simplest case is $F_A = F_B$. This obviously happens if $A = B$, but also for three other snakes. So it holds that

$$\sum_A \sum_{\substack{B\colon \\ |F_A \cap F_B| = 4}} \Pr[X_A = 1 \wedge X_B = 1] = 4 \cdot \mathbb{E}[X_2] = \frac{4}{\mathbb{E}[X_2]} \cdot \mathbb{E}[X_2]^2$$

and since we can achieve $\mathbb{E}[X_2] \geq \varepsilon$ for any constant $\varepsilon \in (0, \frac{1}{2} \cdot \varepsilon_m^4)$ due to Corollary 4.10 by making $\varepsilon_2$ sufficiently small, we get

$$\sum_A \sum_{\substack{B\colon \\ |F_A \cap F_B| = 4}} \Pr[X_A = 1 \wedge X_B = 1] \leq \frac{4}{\varepsilon} \cdot \mathbb{E}[X_2]^2 = \varepsilon_F \cdot \mathbb{E}[X_2]^2 \tag{4.11}$$

for any constant $\varepsilon_F > 8/\varepsilon_m^4$. This captures the case $l = 4$.

For $1 \leq l \leq 3$ it holds that

$$\begin{aligned}
&\sum_A \sum_{\substack{B\colon \\ |F_A \cap F_B| = l}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq \binom{m}{8-l} \cdot (8-l)! \cdot \left(1 - \sum_{c \in F_A \cup F_B} \Pr[c]\right)^{m-8+l} \cdot 2^3 \cdot 2! \cdot \left(\frac{C}{2}\right)^4 \cdot \\
&\cdot \left(\sum_{i=1}^{n} \left(p_i^4 \cdot \sum_{\substack{S \subseteq ([n] \setminus \{i\})\colon \\ |S| = 2}} \prod_{s \in S} p_s^2\right)\right) \cdot \sum_{\substack{B\colon \\ |F_A \cap F_B| = l}} \prod_{c \in F_B \setminus F_A} \Pr[c]
\end{aligned} \tag{4.12}$$

where we accounted for the $8 - l$ possible positions of clauses from $F_A \cup F_B$ in $\Phi$, for the $2^3 \cdot 2!$ possibilities to create a snake $A$ from chosen variables if the central variable is determined already, and for the ways to choose those variables. Now

we want to bound the term

$$\sum_{i=1}^{n}\left(p_i^4 \cdot \sum_{\substack{S\subseteq([n]\setminus\{i\}):\\ |S|=2}} \prod_{s\in S} p_s^2\right).$$

In order to do so we distinguish between the cases that $p_1$ appears in the snake as the central variable, a non-central variable or not at all to show the following

$$\begin{aligned}&\sum_{i=1}^{n}\left(p_i^4 \cdot \sum_{\substack{S\subseteq([n]\setminus\{i\}):\\ |S|=2}} \prod_{s\in S} p_s^2\right)\\ &\leq p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2 + \left(\sum_{i=2}^{n} p_i^4\right) \cdot p_1^2 \cdot \left(\sum_{i=2}^{n} p_i^2\right) + \left(\sum_{i=2}^{n} p_i^4\right) \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2.\end{aligned}$$

Again, the monotonicity of vector norms implies $\sum_{i=2}^{n} p_i^4 \leq \left(\sum_{i=2}^{n} p_i^2\right)^2$ and thus

$$\begin{aligned}&\leq p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2 + p_1^2 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^3 + \left(\sum_{i=2}^{n} p_i^2\right)^4\\ &\leq \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2,\end{aligned}$$

where we used the prerequisite $\sum_{i=2}^{n} p_i^2 \leq \sum_{i=1}^{n} p_i^2 \leq p_1^2/\varepsilon_1$. If we plug this into equation (4.12), we get

$$\begin{aligned}&\sum_{A} \sum_{\substack{B:\\ |F_A\cap F_B|=l}} \Pr[X_A = 1 \wedge X_B = 1]\\ &\quad\leq \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot m^{8-l} \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2 \cdot \sum_{\substack{B:\\ |F_A\cap F_B|=l}} \prod_{c\in F_B\setminus F_A} \Pr[c]. \quad (4.13)\end{aligned}$$

Now we consider the cases $l \in \{1, 2, 3\}$. We assume that $A$ is chosen already and that we want to construct all snakes $B$ that contain exactly $l$ clauses from $A$. However, the bounds we derive will be independent of the actual choice of $A$. Thus, we can simply plug them into equation (4.13). Remember that a snake of size 2 contains the four clauses

$$(w_2, w_1), (\overline{w_1}, w_2), (\overline{w_2}, w_3), (\overline{w_3}, \overline{w_2})$$

for literals $w_1, w_2$, and $w_3$ of distinct Boolean variables.

For $l = 1$ we know one shared clause which has to contain $B$'s central variable

$x$ and one of $B$'s non-central variables $y$. Here, we overestimate that for $B$ we choose any two variables from $A$, one as $B$'s central variable and one as a non-central variable. The other non-central variable $z$ only has to be different from $x$ and $y$. If we have a look at the clauses a snake of size 2 contains, we see that the central variable appears 4 times, while the other two variables appear two times each. However, the central and one of the non-central variables already appear in a shared clause from $A$ and no clause of a snake is supposed to appear more than once in the formula. Thus, the central variable $x$ appears an additional 3 times, $y$ appears an additional one time, and $z$ appears an additional two times. Formally, it holds that

$$\sum_{\substack{B: \\ |F_A \cap F_B|=1}} \prod_{c \in F_B \setminus F_A} \Pr[c] \leq \left(\frac{C}{2}\right)^3 \cdot \sum_{x \in (S \cup \{i\})} \left( p_x^3 \cdot \sum_{y \in (S \cup \{i\}) \setminus \{x\}} \left( p_y \cdot \sum_{z \in [n] \setminus \{x,y\}} p_z^2 \right) \right),$$

where $i$ is the central variable and $S$ are the other variables of $A$. Again, we can do a case distinction depending on the appearances of $p_1$. We can see that $p_1$ can appear as one of the three variables only. Also, the variables $S \cup \{i\}$ of $A$ are predetermined and $|S \cup \{i\}| = 3$. That means, if 1 is not part of $S \cup \{i\}$ or not chosen from it, the set contains at most 3 other indices, whose associated variables have probabilities of at most $p_2$ each. We now distinguish 4 cases: $x = 1$, $y = 1$, $z = 1$, and $\{x, y, z\} \cap \{1\} = \emptyset$. The terms of the following expression represent those cases. It holds that

$$\begin{aligned}
&\sum_{x \in (S \cup \{i\})} p_x^3 \cdot \left( \sum_{y \in (S \cup \{i\}) \setminus \{x\}} p_y \cdot \left( \sum_{z \in [n] \setminus \{x,y\}} p_z^2 \right) \right) \\
&\leq p_1^3 \cdot 2p_2 \cdot \sum_{i=2}^{n} p_i^2 + 3p_2^3 \cdot p_1 \cdot \sum_{i=2}^{n} p_i^2 + 3p_2^3 \cdot 2p_2 \cdot p_1^2 + 3p_2^3 \cdot 2p_2 \cdot \sum_{i=2}^{n} p_i^2 \\
&\leq 17 \cdot p_1^3 \cdot p_2 \cdot \sum_{i=2}^{n} p_i^2,
\end{aligned}$$

where we used $p_2 \leq p_1$ and $p_2^2 \leq \sum_{i=2}^{n} p_i^2$. Together with equation (4.13), it now holds that

$$\begin{aligned}
&\sum_{A} \sum_{\substack{B: \\ |F_A \cap F_B|=1}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq \frac{17}{8} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot m^7 \cdot C^7 \cdot p_1^7 \cdot p_2 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^3
\end{aligned}$$

$$= \sqrt{\varepsilon_2} \cdot \frac{17}{8} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot m^7 \cdot C^7 \cdot p_1^7 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{7/2},$$

since $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$. In the first staement of Lemma 4.8 we show that for any given $\varepsilon_1$, $\varepsilon_m$, and $\varepsilon \in (0,1)$, we can choose $\varepsilon_2$ small enough such that

$$\mathbb{E}[X_2] \geq (1-\varepsilon) \cdot \frac{1}{2} \cdot m^4 \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2.$$

This implies

$$\begin{aligned}
&\sum_{A} \sum_{\substack{B: \\ |F_A \cap F_B|=1}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq \frac{\sqrt{\varepsilon_2}}{(1-\varepsilon)^2} \cdot \frac{17}{2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot \frac{\mathbb{E}[X_2]^2}{m \cdot C \cdot p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}.
\end{aligned}$$

Since $m \cdot C \cdot p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2} = m/m^\star = \varepsilon_m$, we get

$$\begin{aligned}
&\sum_{A} \sum_{\substack{B: \\ |F_A \cap F_B|=1}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq \frac{\sqrt{\varepsilon_2}}{(1-\varepsilon)^2} \cdot \frac{17}{2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot \frac{\mathbb{E}[X_2]^2}{\varepsilon_m} \\
&= \varepsilon_F \cdot \mathbb{E}[X_2]^2 \qquad (4.14)
\end{aligned}$$

for any $\varepsilon_F > 0$ if we choose $\varepsilon_2$ small enough ($\varepsilon$ decreases as $\varepsilon_2$ does).

Now we consider $l = 2$. Again, it is helpful to visualize the clauses a snake of size 2 consists of:

$$(w_2, w_1), (\overline{w_1}, w_2), (\overline{w_2}, w_3), (\overline{w_3}, \overline{w_2}).$$

With two shared clauses, two cases can happen. Either all three variables of $A$ appear in the two shared clauses or only two do. In the first case, one variable of $A$ appears in $B$ twice, while the other two appear only once. However, this information is already enough to completely determine how all other clauses of $B$ have to look. It implies that the variable that appears twice is the central variable both in $A$ and in $B$, since only the central variable appears in clauses with both other variables. Moreover, the two shared clauses already imply $A = B$ and thus $l = 4$. This means, this case cannot happen! Thus, we only have to consider the second case, in which two variables from $A$ each appear twice in the shared clauses. Again, the shared clauses already determine that the central variable from $A$ also is the central variable in $B$, since only the literals of the central variable appear with the same sign in both clauses. In $B$ this central variable has

to appear an additional two times and a new variable $x \in ([n] \setminus (S \cup \{i\}))$ has to appear two times as well. The other variable of $B$ does not appear again, since it already appeared two times in shared clauses. More formally,

$$\sum_{\substack{B: \\ |F_A \cap F_B|=2}} \prod_{c \in F_B \setminus F_A} \Pr[c] = \left(\frac{C}{2}\right)^2 \cdot p_i^2 \cdot \sum_{x \in [n] \setminus (S \cup \{i\})}^{n} p_x^2.$$

By considering the possible appearances of $p_1$ again, we get

$$\begin{aligned}\sum_{\substack{B: \\ |F_A \cap F_B|=2}} \prod_{c \in F_B \setminus F_A} \Pr[c] &\leq \left(\frac{C}{2}\right)^2 \cdot \left(p_1^2 \cdot \sum_{i=2}^{n} p_i^2 + p_2^2 \cdot p_1^2 + p_2^2 \cdot \sum_{i=2}^{n} p_i^2\right) \\ &\leq 3 \cdot \left(\frac{C}{2}\right)^2 \cdot p_1^2 \cdot \sum_{i=2}^{n} p_i^2.\end{aligned}$$

Again with equation (4.13), it holds that

$$\sum_{A} \sum_{\substack{B: \\ |F_A \cap F_B|=2}} \Pr[X_A = 1 \wedge X_B = 1] \leq \frac{3}{4} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot m^6 \cdot C^6 \cdot p_1^6 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^3.$$

Since we can choose $\varepsilon_2$ small enough such that $\mathbb{E}[X_2] \geq (1-\varepsilon) \cdot \frac{1}{2} \cdot m^4 \cdot C^4 \cdot p_1^4 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^2$ for any $\varepsilon \in (0,1)$, we get

$$\begin{aligned}&\sum_{A} \sum_{\substack{B: \\ |F_A \cap F_B|=2}} \Pr[X_A = 1 \wedge X_B = 1] && (4.15)\\ &\leq \frac{3}{(1-\varepsilon)^2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot \frac{\mathbb{E}[X_2]^2}{m^2 \cdot C^2 \cdot p_1^2 \cdot \left(\sum_{i=2}^{n} p_i^2\right)} \\ &= \frac{3}{(1-\varepsilon)^2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) \cdot \frac{\mathbb{E}[X_2]^2}{\varepsilon_m^2} && (4.16)\\ &= \varepsilon_F \cdot \mathbb{E}[X_2]^2 && (4.17)\end{aligned}$$

for some constant $\varepsilon_F > \frac{3}{\varepsilon_m^2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right)$.

The last case is $l = 3$. This case can not happen, since 3 shared clauses already fully determine the last clause, which also has to align with one of $A$, i. e. we do not have any degree of freedom to make $F_A \neq F_B$.

Putting equation (4.11), equation (4.14), and equation (4.17) together, establishes that we can choose $\varepsilon_2$ sufficiently small to make

$$\sum_{A} \sum_{B:\, B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \leq \varepsilon_F \cdot \mathbb{E}[X_2]^2$$

for any constant $\varepsilon_F > \frac{3}{\varepsilon_m^2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + \frac{8}{\varepsilon_m^4}$. Together with equation (4.10), this gives us

$$\begin{aligned}\mathbb{E}\left[X_2^2\right] &= \sum_A \left( \sum_{B\colon B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] + \sum_{B\colon B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \right) \\ &\leq (1 + \varepsilon_E + \varepsilon_F) \cdot \mathbb{E}[X_2]^2\end{aligned}$$

for any constant $\varepsilon_E > 0$ and any $\varepsilon_F > \frac{3}{\varepsilon_m^2} \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + \frac{8}{\varepsilon_m^4}$ and implies

$$\Pr[X_2 > 0] \geq \frac{\mathbb{E}[X_2]^2}{\mathbb{E}\left[X_2^2\right]} \geq \frac{1}{1 + \varepsilon_E + \varepsilon_F} = \varepsilon_P$$

for any

$$\varepsilon_P < \frac{\varepsilon_m^4}{\varepsilon_m^4 + 3 \cdot \varepsilon_m^2 \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + 8}.$$

■

The following is a corollary of the former lemma and complements it. It shows that for any $\varepsilon_1 \in (0,1)$ and any $\varepsilon_P \in (0,1)$ non-uniform random 2-SAT formulas are unsatisfiable with probability at least $\varepsilon_P$ if we can choose $\varepsilon_m$ with $m = \varepsilon_m \cdot m^\star$ sufficiently large and $\varepsilon_2$ with $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ sufficiently small. This captures the case $m \in \omega(m^\star)$.

▶ **Corollary 4.10.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ and a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot (\sum_{i=1}^n p_i^2)$. Let $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. For any constant $\varepsilon_P \in (0,1)$ we can choose a constant $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ and a constant $\varepsilon_2 \in (0,1)$ with $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ such that the probability to generate an unsatisfiable formula $\Phi \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ is at least $\varepsilon_P$. ◀

*Proof.* The corollary is a simple application of the former lemma. Suppose we are given $\varepsilon_1$ and $\varepsilon_P$. We can now choose an $\varepsilon_m$ large enough such that

$$\frac{\varepsilon_m^4}{\varepsilon_m^4 + 3 \cdot \varepsilon_m^2 \cdot \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + 8} > \varepsilon_P$$

for the given $\varepsilon_P$. Due to Lemma 4.9 we can then choose $\varepsilon_2$ small enough to generate unsatisfiable instances with probability at least $\varepsilon_P$. ■

The former lemma and corollary together with Lemma 4.4 establish that in the case of $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in o(\sum_{i=2}^n p_i^2)$ the asymptotic threshold is at $m \in \Theta((C \cdot p_1 (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$ and that it is coarse. We will show this formally in Section 4.5.

### 4.3.2 The sharp threshold case

In the last section we analyzed the case $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in o(\sum_{i=2}^n p_i^2)$. Now we tend to the case $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. In this section we are going to show that there is a sharp threshold at $m^\star = (\sum_{i=1}^n p_i^2)^{-1}$. From Lemma 4.3 we already know that formulas are *satisfiable* with probability $1 - o(1)$ if $m \leq \varepsilon_m \cdot m^\star$ for any constant $\varepsilon_m \in (0,1)$. It remains to show that they are *unsatisfiable* with probability $1 - o(1)$ if $m \geq \varepsilon_m \cdot m^\star$ for any constant $\varepsilon_m > 1$. This will establish a sharp threshold at $m^\star = (\sum_{i=1}^n p_i^2)^{-1}$. More generally, we will show that, given $\varepsilon_P \in (0,1)$ and $\varepsilon_m > 1$ so that $m = \varepsilon_m \cdot m^\star$, we can choose $\varepsilon_1$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ sufficiently small so that the probability to generate an unsatisfiable instance is at least $\varepsilon_P$.

We use the same technique as in the last section to prove this result, the second moment method. As before, in order to show a sufficiently large probability for unsatisfiability, we first have to show that the expected number of snakes of a certain size $t$ can be made arbitrarily large. The following lemma establishes this property for the suitably chosen value $t = f^{1/78}$, where $f = (\sum_{i=1}^n p_i^2)/p_1^2 \geq 1/\varepsilon_1$.

▶ **Lemma 4.11.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ and let $t = f^{1/78}$, where $f = (\sum_{i=1}^n p_i^2)/p_1^2$. Let $m^\star = (\sum_{i=1}^n p_i^2)^{-1}$ and let $m = \varepsilon_m \cdot m^\star$ for some given constant $\varepsilon_m > 1$. Given a constant $\varepsilon_E \in (0,1)$, we can choose a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ such that

$$\mathbb{E}[X_t] \geq (1-\varepsilon_E) \cdot \frac{1}{2} \cdot m^{2t} \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=2}^n p_i^2\right)^{2t-2}.$$

Furthermore, for any given $\varepsilon_E > 0$, we can choose $\varepsilon_1 \in (0,1)$ small enough to guarantee $\mathbb{E}[X_t] \geq \varepsilon_E$. ◀

*Proof.* It holds that $p_1 \leq 1$, $\sum_{i=1}^n p_i^2 \leq 1$, and $C = 1 - \sum_{i=1}^n p_i^2 \leq 1$. Thus,

$$t = f^{1/78} = \frac{\left(\sum_{i=1}^n p_i^2\right)^{1/78}}{p_1^{1/39}} \leq p_1^{-1/39} \leq p_1^{-1} \leq 1/(C \cdot p_1 \cdot p_2) = (2 \cdot q_{\max})^{-1}.$$

Also,

$$t = f^{1/78} \geq \frac{1}{\varepsilon_1^{1/78}} \geq 2$$

if $\varepsilon_1$ is sufficiently small. Therefore, we can apply Lemma 4.6 to get

$$\begin{aligned}&\mathbb{E}[X_t] \\ &\geq \frac{1}{2} \cdot (m-2t)^{2t} \cdot C^{2t} \cdot (1 - 2t \cdot q_{\max} \cdot m) \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=2}^n \left(p_i^2 - (2t-3) \cdot p_2^2\right)\right)^{2t-2}.\end{aligned}$$

We are going to show that this is at least

$$(1-\varepsilon_E)\cdot\frac{1}{2}\cdot m^{2t}\cdot\left(\sum_{i=1}^{n}p_i^4\right)\cdot\left(\sum_{i=2}^{n}p_i^2\right)^{2t-2}$$

for some $\varepsilon_E \in (0,1)$ that we can make arbitrarily small by making $\varepsilon_1$ sufficiently small.

First, we see that

$$\begin{aligned}\left(\sum_{i=2}^{n}(p_i^2-(2t-3)\cdot p_2^2)\right)^{2t-2} &\geq \left(\sum_{i=1}^{n}(p_i^2-(2t-2)\cdot p_1^2)\right)^{2t-2}\\ &= \left(\sum_{i=1}^{n}p_i^2\right)^{2t-2}\cdot\left(1-\frac{(2t-2)\cdot p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{2t-2}.\end{aligned}$$

It holds that

$$\begin{aligned}\left(1-\frac{(2t-2)\cdot p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{2t-2} &\geq \left(1-\frac{2\cdot f^{1/78}\cdot p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{2t-2}\\ &= \left(1-2\cdot\left(\frac{p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{77/78}\right)^{2t-2}\end{aligned}$$

where we used our definitions of $t$ and $f$. We can see that $\left(\frac{p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{77/78} \leq \varepsilon_1^{77/78}$. By choosing $\varepsilon_1$ sufficiently small, this allows us to use Bernoulli's inequality and get

$$\begin{aligned}\left(1-\frac{(2t-2)\cdot p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{2t-2} &\geq \left(1-2\cdot\left(\frac{p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{77/78}\cdot 2t\right)\\ &= \left(1-4\cdot\left(\frac{p_1^2}{\sum_{i=1}^{n}p_i^2}\right)^{38/39}\right)\\ &\geq \left(1-4\cdot\varepsilon_1^{38/39}\right).\end{aligned}$$

We can make this factor arbitrarily close to one if we choose $\varepsilon_1$ sufficiently small.

Equivalently,

$$\begin{aligned}(m-2t)^{2t} &= m^{2t}\cdot\left(1-\frac{2t}{m}\right)^{2t}\\ &= m^{2t}\cdot\left(1-2\cdot\frac{f^{1/78}\cdot\left(\sum_{i=1}^{n}p_i^2\right)}{\varepsilon_m}\right)^{2t},\end{aligned}$$

where we used $m = \varepsilon_m / \sum_{i=1}^{n} p_i^2$ and $t = f^{1/78}$. It holds that $f \cdot p_1^2 = \sum_{i=1}^{n} p_i^2 \leq p_1$, which implies $f \leq p_1^{-1}$ and $t \leq p_1^{-1/78}$. Furthermore, we still know that $p_1 \leq \varepsilon_1^{1/2}$, since $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ and $\sum_{i=1}^{n} p_i^2 \leq 1$. This implies

$$\left(1 - 2 \cdot \frac{f^{1/78} \cdot \left(\sum_{i=1}^{n} p_i^2\right)}{\varepsilon_m}\right)^{2t} \geq \left(1 - 2 \cdot \frac{p_1^{77/78}}{\varepsilon_m}\right)^{2t} \geq \left(1 - 2 \cdot \frac{\varepsilon_1^{77/156}}{\varepsilon_m}\right)^{2t}.$$

Again, we can make $\varepsilon_1$ small enough to use Bernoulli's inequality and get

$$\left(1 - 2 \cdot \frac{p_1^{77/78}}{\varepsilon_m}\right)^{2t} \geq \left(1 - 2 \cdot \frac{2t \cdot p_1^{77/78}}{\varepsilon_m}\right) \geq \left(1 - 4 \cdot \frac{p_1^{38/39}}{\varepsilon_m}\right) \geq \left(1 - 4 \cdot \frac{\varepsilon_1^{19/39}}{\varepsilon_m}\right).$$

Thus, for any fixed $\varepsilon_m$ we can make this factor arbitrarily close to one by making $\varepsilon_1$ sufficiently small.

Since we know that $C = (1 - \sum_{i=1}^{n} p_i^2)^{-1}$ and $\sum_{i=1}^{n} p_i^2 \leq p_1 \leq \sqrt{\varepsilon_1}$, this implies $1 \leq C \leq (1 - \sqrt{\varepsilon_1})^{-1}$. We also know that $t = f^{1/78} \geq \varepsilon_1^{-1/78}$, which implies

$$\begin{aligned}
1 - 2t \cdot q_{\max} \cdot m &= 1 - f^{1/78} \cdot C \cdot p_1 \cdot p_2 \cdot \frac{\varepsilon_m}{\sum_{i=1}^{n} p_i^2} \\
&\geq 1 - f^{1/78} \cdot \frac{\varepsilon_m}{1 - \sqrt{\varepsilon_1}} \cdot \frac{p_1^2}{\sum_{i=1}^{n} p_i^2} \\
&= 1 - f^{-77/78} \cdot \frac{\varepsilon_m}{1 - \sqrt{\varepsilon_1}} \\
&\geq 1 - \frac{\varepsilon_m \cdot \varepsilon_1^{77/78}}{1 - \varepsilon_1^{1/2}}.
\end{aligned}$$

Here, we also use $p_2 \leq p_1$ and the definition $f = (\sum_{i=1}^{n} p_i^2)/p_1^2$. Thus, for any given $\varepsilon_m$ we can make this expression arbitrarily close to one by choosing $\varepsilon_1$ sufficiently small.

For any given $\varepsilon_m$ we can now choose an $\varepsilon_E \in (0, 1)$ and get

$$\mathbb{E}[X_t] \geq (1 - \varepsilon_E) \cdot \frac{1}{2} \cdot m^{2t} \cdot \left(\sum_{i=1}^{n} p_i^4\right) \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{2t-2}$$

by making $\varepsilon_1$ sufficiently small. We want to show that for every $\varepsilon_m$ we can make the expected value arbitrarily large by choosing $\varepsilon_1$ small enough. It holds that

$$m^2 \cdot \left(\sum_{i=1}^{n} p_i^4\right) \geq m^2 \cdot p_1^4 = \frac{\varepsilon_m^2 \cdot p_1^4}{\left(\sum_{i=1}^{n} p_i^2\right)^2} = \frac{\varepsilon_m^2}{f^2},$$

where we used $m = \varepsilon_m \cdot m^\star = \varepsilon_m \cdot \left(\sum_{i=1}^n p_i^2\right)^{-1}$. With the same fact it holds that

$$\left(m \cdot \sum_{i=1}^n p_i^2\right)^{2t-2} = \varepsilon_m^{2t-2}.$$

Since we know that $t = f^{1/78}$, it holds that

$$\mathbb{E}[\, X_t \,] \geq (1 - \varepsilon_E) \cdot \frac{1}{2} \cdot \frac{\varepsilon_m^{2 \cdot f^{1/78}}}{f^2}.$$

We can now make this expression as large as we need it if $f$ is sufficiently large, because we assumed $\varepsilon_m > 1$. Since $f \geq 1/\varepsilon_1$ this is the case if $\varepsilon_1$ is sufficiently small. Thus, for any given $\varepsilon_E > 0$ we can choose $\varepsilon_1$ sufficiently small to guarantee $\mathbb{E}[\, X_t \,] \geq \varepsilon_E$. ■

We now turn to the application of the second moment method. Again, we want to show that $\Pr[X_A = 1 \wedge X_B = 1]$ for snakes $A$ and $B$ with shared clauses ($F_A \cap F_B \neq \emptyset$) is relatively small compared to $\mathbb{E}[\, X_t \,]^2$. To this end, we have to consider different possibilities for the shared clauses to influence $\Pr[X_A = 1 \wedge X_B = 1]$. In the proofs of the former case this was rather easy, since we only considered the smallest possible snakes of size $t = 2$. Now the distinction becomes a bit more difficult. We will distinguish several cases: If the number of shared clauses is at least $t$ then $\Pr[X_A = 1 \wedge X_B = 1]$ is by roughly a factor of $\varepsilon_m^t$ smaller than $\mathbb{E}[\, X_t \,]^2$. If the shared clauses form a variable-variable-incidence graph with at least two connected components, then there are enough variable appearances pre-defined for $B$ to make $\Pr[X_A = 1 \wedge X_B = 1]$ sufficiently small. The last case is that the shared clauses form only one connected component, which is a lot smaller than $t - 1$. In that case we have to carefully consider what happens to the central variable of $B$, since this variable appears most times in $B$ and the many appearances take degrees of freedom away from other variables, therefore making $\Pr[X_A = 1 \wedge X_B = 1]$ small. Here, we only show the result for some $\varepsilon_m > 1$ so that $m = \varepsilon_m \cdot \left(\sum_{i=1}^n p_i^2\right)^{-1}$. For $m \in \omega((\sum_{i=1}^n p_i^2)^{-1})$ it follows by the monotonicity of unsatisfiability as we will see later.

▶ **Lemma 4.12.** Given an ensemble of probability distributions $(\vec{p}^{(n)})_{n \in \mathbb{N}}$. Let $m^\star = (\sum_{i=1}^n p_i^2)^{-1}$ and let $m = \varepsilon_m \cdot m^\star$ for some given constant $\varepsilon_m > 1$. Given an $\varepsilon_P \in (0,1)$, we can choose a constant $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ such that the probability that a random formula $\Phi \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ is unsatisfiable is at least $\varepsilon_P$. ◀

*Proof.* Again, we utilize the second moment method. We want to show that for any given $\varepsilon_P \in (0,1)$ we can choose $\varepsilon_1 \in (0,1)$ sufficiently small so that some $F_A$ for a snake $A$ of size $t$ appears in $\Phi$ with probability at least $\varepsilon_P$. This especially implies that we can make this probability arbitrarily close to one. This will hold for $t = f^{1/78}$, where $f = (\sum_{i=1}^n p_i^2)/p_1^2$. We will later see why we chose $t$ this way.

Again, we define $X_A$ as an indicator variable for the event that the formula $F_A$ associated with snake $A$ appears exactly once in $\Phi$ and $X_t = \sum_{\text{snake } A \text{ of size } t} X_A$. As in the proof of Corollary 4.10 we want to show that for any $\varepsilon_E > 0$ we can choose $\varepsilon_1$ small enough so that $\mathbb{E}\left[ X_t^2 \right] \leq (1 + \varepsilon_E) \cdot \mathbb{E}[ X_t ]^2$. Then, the second moment method gives us

$$\Pr[X_t > 0] \geq \frac{\mathbb{E}[ X_t ]^2}{\mathbb{E}\left[ X_t^2 \right]} \geq \frac{1}{1 + \varepsilon_E}.$$

Thus, for any given $\varepsilon_P \in (0, 1)$ we can simply choose $\varepsilon_E = \frac{1}{\varepsilon_P} - 1$ to get the result as desired. We again split the expected value into two sums

$$\begin{aligned}\mathbb{E}\left[ X_t^2 \right] &= \sum_A \sum_B \Pr[X_A = 1 \wedge X_B = 1] \\ &= \sum_{B\colon B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] + \sum_{B\colon B \sim A} \Pr[X_A = 1 \wedge X_B = 1],\end{aligned}$$

where $B \sim A$ denotes $F_A \cap F_B \neq \emptyset$. We will now consider the parts over $B \nsim A$ and $B \sim A$ separately, starting with $B \nsim A$.

As in the proof of Corollary 4.10, we want to show that for any $\varepsilon_E > 0$, we can choose $\varepsilon_1$ such that

$$\sum_A \sum_{B\colon B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] \leq (1 + \varepsilon_E) \cdot \mathbb{E}[ X_t ]^2. \tag{4.18}$$

It holds that

$$\begin{aligned}&\Pr[X_A = 1 \wedge X_B = 1] \\ &\quad = \binom{m}{4t} \cdot (4t)! \cdot \left( \prod_{c \in F_A} \Pr[c] \right) \cdot \left( \prod_{c \in F_B} \Pr[c] \right) \cdot \left( 1 - \sum_{c \in F_A \cup F_B} \Pr[c] \right)^{m-4t},\end{aligned}$$

while

$$\Pr[X_A = 1] = \binom{m}{2t} \cdot (2t)! \cdot \left( \prod_{c \in F_A} \Pr[c] \right) \cdot \left( 1 - \sum_{c \in F_A} \Pr[c] \right)^{m-2t}.$$

This already gives us

$$\begin{aligned}&\Pr[X_A = 1 \wedge X_B = 1] \\ &\leq \Pr[X_A = 1] \cdot \Pr[X_B = 1] \cdot \frac{\left(1 - \sum_{c \in F_A \cup F_B} \Pr[c]\right)^{m-4t}}{\left(1 - \sum_{c \in F_A} \Pr[c]\right)^{m-2t} \left(1 - \sum_{c \in F_B} \Pr[c]\right)^{m-2t}},\end{aligned}$$

since $\binom{m}{4t} \cdot (4t)! \leq \left(\binom{m}{2t} \cdot (2t)!\right)^2$. Due to

$$\left(1 - \sum_{c \in F_A} \Pr[c]\right) \cdot \left(1 - \sum_{c \in F_B} \Pr[c]\right) \geq 1 - \sum_{c \in F_A \cup F_B} \Pr[c],$$

and $\sum_{c \in F_A} \Pr[c], \sum_{c \in F_B} \Pr[c] \leq 2t \cdot q_{\max}$ we have

$$\frac{\left(1 - \sum_{c \in F_A \cup F_B} \Pr[c]\right)^{m-4t}}{\left(1 - \sum_{c \in F_A} \Pr[c]\right)^{m-2t} \left(1 - \sum_{c \in F_B} \Pr[c]\right)^{m-2t}}$$
$$\leq \left(1 - \sum_{c \in F_A} \Pr[c]\right)^{-2t} \cdot \left(1 - \sum_{c \in F_B} \Pr[c]\right)^{-2t} \leq (1 - 2t \cdot q_{\max})^{-4t}.$$

Since we know $t \leq (2 \cdot q_{\max})^{-1}$ from the previous lemma, we can use Bernoulli's inequality to get

$$(1 - 2t \cdot q_{\max})^{-4t} \leq (1 - 8t^2 \cdot q_{\max})^{-1}.$$

We know that $t^2 = f^{1/39} = (\sum_{i=1}^n p_i^2)^{1/39} / p_1^{2/39}$ and that $q_{\max} = \frac{1}{2} \cdot C \cdot p_1 \cdot p_2 = \frac{p_1 \cdot p_2}{2 \cdot (1 - \sum_{i=1}^n p_i^2)}$. Together with $p_2 \leq p_1 \leq \varepsilon_1^{1/2}$ and $\sum_{i=1}^n p_i^2 \leq p_1 \leq \varepsilon_1^{1/2}$ this yields

$$\begin{aligned}(1 - 8t^2 \cdot q_{\max})^{-1} &= \left(1 - 4 \cdot \frac{(\sum_{i=1}^n p_i^2)^{1/39} \cdot p_1 \cdot p_2}{p_1^{2/39} \cdot (1 - \sum_{i=1}^n p_i^2)}\right)^{-1} \\ &\leq \left(1 - 4 \cdot \frac{p_1^{77/39}}{1 - p_1}\right)^{-1} \\ &\leq \left(1 - 4 \cdot \frac{\varepsilon_1^{77/78}}{1 - \varepsilon_1^{1/2}}\right)^{-1} \leq 1 + \varepsilon_E\end{aligned}$$

for any $\varepsilon_E > 0$ if we make $\varepsilon_1$ sufficiently small. We now get

$$\Pr[X_A = 1 \wedge X_B = 1] \leq (1 + \varepsilon_E) \cdot \Pr[X_A = 1] \cdot \Pr[X_B = 1]$$

for $A \nsim B$ and thus

$$\begin{aligned}\sum_A \sum_{B:\, B \nsim A} \Pr[X_A = 1 \wedge X_B = 1] &\leq (1 + \varepsilon_E) \cdot \sum_A \sum_{B:\, B \nsim A} \Pr[X_A = 1] \cdot \Pr[X_B = 1] \\ &\leq (1 + \varepsilon_E) \cdot \mathbb{E}[\, X_2 \,]^2.\end{aligned}$$

Second, we look at snakes $B \sim A$. For those we want to show

$$\sum_A \sum_{B:\, B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \leq \varepsilon_E \cdot \mathbb{E}[\, X_t \,]^2 \tag{4.19}$$

**(a)** If the shared clauses form a forest, the number of connected components $c$ is the difference between the number of different variables in shared clauses $j$ and the number of shared clauses $l$, $c = j - l$.

**(b)** In order for the shared clauses to form a cycle, at least $t$ of the $2t$ clauses have to be shared. In that case the number of connected components is $c = j - l + 1$.

**Figure 4.2:** Visual representation of the variable-variable incidence graph $G_{F_A}$ for a snake $A$. Each node represents a variable of the snake, while each edge represents a clause of $A$ containing those variables. The node of degree 4 represents the central variable of $A$. Shared clauses with snake $B$ are highlighted in red, i. e. red edges represent clauses that appear both in $F_A$ and $F_B$. However, those edges do not necessarily appear at the same position in $G_{F_B}$.

for any $\varepsilon_E > 0$ if we make $\varepsilon_1$ sufficiently small. First, let us consider $F_A = F_B$. As in the case of $t = 2$ it holds that there are exactly 4 snakes with the same set of clauses. Thus,

$$\sum_A \sum_{B\colon F_B = F_A} \Pr[X_A = 1 \wedge X_B = 1] = 4 \cdot \mathbb{E}[\, X_t \,] = \frac{4}{\mathbb{E}[\, X_t \,]} \cdot \mathbb{E}[\, X_t \,]^2.$$

Lemma 4.11 tells us that for any $\varepsilon > 0$ we can choose $\varepsilon_1$ such that $\mathbb{E}[\, X_t \,] \geq \varepsilon$. Therefore, for any $\varepsilon_E > 0$ we can choose $\varepsilon_1$ sufficiently small to get

$$\sum_A \sum_{B\colon F_B = F_A} \Pr[X_A = 1 \wedge X_B = 1] = \frac{4}{\mathbb{E}[\, X_t \,]} \cdot \mathbb{E}[\, X_t \,]^2 \leq \frac{4}{\varepsilon} \cdot \mathbb{E}[\, X_t \,]^2 \leq \varepsilon_E \cdot \mathbb{E}[\, X_t \,]^2.$$

The remaining analysis is a bit more complicated than in the case of $t = 2$, since we can not always surely say how many variables of snake $B$ are predefined by shared clauses. As before, we are classifying snakes $B \sim A$ according to the number $l = |F_A \cap F_B|$ of shared clauses, but also according to the number $j$ of nodes in the variable-variable incidence graph $G_{F_A \cap F_B}$. Note that the number of variables that $F_A$ and $F_B$ have in common (regardless of signs) could be greater! In fact, they could share all their variables without having a single clause in common. However, right now we are only interested in ways to incorporate clauses from $F_A$ as shared clauses into $F_B$. To that end, we only need to consider the variables from these clauses as shared variables. For a representation, see Figure 4.2.

Suppose now that snake $A$ and the shared clauses are fixed. We let $j$ denote the number of variables in shared clauses. We know that there are $2t - 1 - j$ *free* variables in $B$, i. e. variables which are not predetermined to appear in $B$ by shared clauses. Furthermore we can give an upper bound on the number $c$ of connected components of $G_{F_A \cap F_B}$. It is easy to see that $c \leq j - l$ for $l < t$ ($G_{F_A \cap F_B}$ is a forest), $c \leq j - l + 1$ for $t \leq l < 2t$ (we could create one cycle), and

$c = j - l + 2$ for $l = 2t$ ($F_A = F_B$). These cases are also visualized in Figure 4.2. Fixing $l$ and $j$ it holds that

$$\sum_{\substack{\text{snakes } A,\, B:\\ |E(G_{F_A\cap F_B})|=l,\ |V(G_{F_A\cap F_B})|=j}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq \binom{m}{4t-l} \cdot (4t-l)! \cdot \left(\frac{C}{2}\right)^{4t-l} \cdot 2^{2t-1} \cdot (2t-2)! \cdot \left( \sum_{\substack{S_A\subseteq[n]:\\ |S_A|=2t-2}} \prod_{x\in S_A} p(x)^2 \right) \cdot$$

$$\cdot \left( \sum_{y\in[n]} p(y)^4 \right) \cdot 4 \cdot \left( \binom{2t+2}{2(j-l)+2} \right)^2 \cdot c! \cdot 2^c \cdot 2t \cdot (2t-1-j)! \cdot 2^{2t-1-j} \cdot$$

$$\cdot \left( \sum_{\substack{S_B\subseteq[n]:\\ |S_B|=2t-1-j}} \prod_{x\in S_B} p(x)^2 \right) \cdot p_1^{2(j-l+1)} \cdot \left(1 - \sum_{c\in F_A\cup F_B} \Pr[c]\right)^{m-(4t-l)}. \tag{4.20}$$

Before we upper bound this expression even further, let us explain where it comes from. There are $\binom{m}{4t-l} \cdot (4t-l)!$ positions for the $4t-l$ clauses of $F_A \cup F_B$ in the $m$-clause formula $\Phi$. There are at most $2^{2t-2} \cdot (2t-2)!$ possibilities of forming different snakes (signs and positions) from the $2t-2$ variables of $A$, excluding $y = |w_t|$, and two possible signs for $y = |w_t|$. In snake $A$ each variable appears exactly twice, except for $y = |w_t|$, which appears four times. Now we want to count the ways of mapping $G_{F_A\cap F_B}$ to $G_{F_A}$ and $G_{F_B}$. Following the argumentation from [CR92] we can see that there are $2\binom{2t+2}{2j-2l+2}$ possible mappings for $G_{F_A}$ and $G_{F_B}$, respectively. These mappings fix the shared clauses we choose from $A$ as well as the positions where shared clauses can appear in $B$, but not where exactly which clause will appear. This is what we consider next. We know that $G_{F_A\cap F_B}$ contains $c$ connected components. If they are of same length, they can be interchanged in $c!$ ways. Furthermore, each component might be flipped, i. e. the sign of every literal in the component and their order in $B$ can be inverted. For components which are paths, this does not change the set of shared clauses they originate from. Nevertheless, there is still the possibility of having one component which is not a path. For this component there are at most $2t$ ways of mapping it onto its counterpart (if it is a cycle) due to [CR92]. Now we know the shared clauses from $F_A$ and the exact position of these clauses in $F_B$ as well as positions reserved for non-determined variables in snake $B$. The remaining $2t-1-j$ non-determined variables from $B$ can be chosen arbitrarily. Also, there are $2^{2t-1-j} \cdot (2t-1-j)!$ possibilities for them to fill out the blanks of snake $B$. Each of these variables appears at least twice in $B$ only. The remaining at most $2(j-l+1)$ appearances of variables in $F_B$ are determined by the previous choices and give an additional factor of at most $p_1^{2(j-l+1)}$. Note that the case that one of our free variables in $B$ is a central variable is also captured by this upper bound,

since $\sum_{i=1}^n p_i^4 \leq p_1^2 \cdot \sum_{i=1}^n p_i^2$. The other $m - (4t - l)$ clauses of $\Phi$ are supposed to be different from those in $F_A \cup F_B$, so that both $F_A$ and $F_B$ appear exactly once.

Now we want to simplify that expression. It holds that

$$\left(1 - \sum_{c \in F_A \cup F_B} \Pr[c]\right)^{m-(4t-l)} \leq 1$$

and that

$$C^{4t-l} \leq \left(1 + \frac{\sum_{i=1}^n p_i^2}{1 - \sum_{i=1}^n p_i^2}\right)^{4t} \leq \exp\left(4t \cdot \frac{\sum_{i=1}^n p_i^2}{1 - \sum_{i=1}^n p_i^2}\right).$$

We know that $t = f^{1/78} \leq p_1^{-1/78}$ and that $\sum_{i=1}^n p_i^2 \leq p_1 \leq \varepsilon_1^{1/2}$. This implies

$$C^{4t-l} \leq \exp\left(4t \cdot \frac{\sum_{i=1}^n p_i^2}{1 - \sum_{i=1}^n p_i^2}\right) \leq \exp\left(4 \cdot \frac{p_1^{77/78}}{1 - p_1}\right) \leq \exp\left(4 \cdot \frac{\varepsilon_1^{77/156}}{1 - \varepsilon_1^{1/2}}\right).$$

For any $\varepsilon > 0$ we can choose $\varepsilon_1$ small enough such that this expression is at most $1 + \varepsilon$. Again

$$\sum_{\substack{S \subseteq [n]: \\ |S| = x}} \prod_{s \in S} p(s)^2 \leq \frac{1}{x!} \left(\sum_{i=1}^n p_i^2\right)^x$$

according to Lemma 4.1. This step also cancels out the factors $(2t - 2)!$ and $(2t - 1 - j)!$. Also, all factors of 2 that appear cancel out with $c \leq j - l + 2$. We will also use the following estimation

$$\left(\binom{2t+2}{2(j-l)+2}\right)^2 \cdot c! \leq \frac{(2t+2)^{4(j-l+1)}}{(2(j-l+1)!)^2} \cdot (j-l+2)! \leq (2t+2)^{4(j-l+1)} \leq (3t)^{4(j-l+1)}.$$

This holds since $j \geq l - 1$ and $t \geq 2$. However, $j = l - 1$ only happens if $F_A = F_B$. Since we already considered this case, we will further assume $j \geq l$. Plugging everything back into equation (4.20) we get

$$\begin{aligned} &\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})| = l,\ |V(G_{F_A \cap F_B})| = j}} \Pr[X_A = 1 \wedge X_B = 1] \\ &\leq 4 \cdot (1 + \varepsilon) \cdot m^{4t-l} \cdot (3t)^{5(j-l+1)} \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-j-3} \cdot p_1^{2(j-l+1)} \end{aligned} \tag{4.21}$$

for some $\varepsilon > 0$ that decreases as $\varepsilon_1$ does.

We will distinguish three cases now, depending on the value of $j - l$. First $j - l = 0$, then $j - l \geq 2$ and finally $j - l = 1$. For each of these cases we want to

show that for any $\varepsilon_E > 0$ we can choose $\varepsilon_1$ small enough so that

$$\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=j}} \Pr[X_A = 1 \wedge X_B = 1] \leq \varepsilon_E \cdot \frac{\mathbb{E}[X_t]^2}{t^2}.$$

Since $1 \leq l \leq 2t$ and $2 \leq j \leq 2t-1$, we will get an additional factor of $4t^2$ when summing over all snakes $A \sim B$. If we consider all cases, including $F_A = F_B$, this adds up to

$$\sum_A \sum_{B:\, B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \leq 16 \cdot \varepsilon_E \cdot \mathbb{E}[X_t]^2.$$

Still, for any chosen $\varepsilon_E > 0$ we can choose $\varepsilon_1 \in (0,1)$ small enough to make this expression at most $\varepsilon_E \cdot \mathbb{E}[X_t]^2$ as desired.

Now let us consider the first case, $j = l$. This can only happen if $G_{F_A \cap F_B}$ contains a cycle, as we can see in Figure 4.2. However, $G_{F_A \cap F_B}$ can only contain a cycle if $l \geq t$. Due to equation (4.21) it holds that

$$\begin{aligned}&\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l}} \Pr[X_A = 1 \wedge X_B = 1] \\ &\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^5 \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-l-3} \cdot p_1^2\end{aligned}$$

Remember that due to Lemma 4.11 for any $\varepsilon \in (0,1)$ we can choose $\varepsilon_1 \in (0,1)$ small enough so that

$$\mathbb{E}[X_t]^2 \geq (1-\varepsilon) \cdot \frac{1}{4} \cdot m^{4t} \cdot \left(\sum_{i=1}^n p_i^4\right)^2 \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-4}.$$

Thus,

$$\begin{aligned}&\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l}} \Pr[X_A = 1 \wedge X_B = 1] \\ &\leq 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot 3^5 \cdot t^5 \cdot \left(m \cdot \sum_{i=1}^n p_i^2\right)^{-l} \cdot \frac{p_1^2 \cdot \sum_{i=1}^n p_i^2}{\sum_{i=1}^n p_i^4} \cdot \mathbb{E}[X_t]^2,\end{aligned}$$

Here, we can choose $\varepsilon$ arbitrarily small by making $\varepsilon_1$ sufficiently small. Due to $m = \varepsilon_m / \sum_{i=1}^n p_i^2$, $l \geq t$, and $\sum_{i=1}^n p_i^4 \geq p_1^4$ this yields

$$\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot 3^5 \cdot t^5 \cdot \varepsilon_m^{-t} \cdot f \cdot \mathbb{E}[\,X_t\,]^2.$$

Since $t = f^{1/78}$ and $\varepsilon_m > 1$, we can make this expression at most $\varepsilon_E \cdot \mathbb{E}[\,X_t\,]^2$ for any $\varepsilon_E > 0$ by making $f$ sufficiently large. Due to $f \geq 1/\varepsilon_1$, we can also make $\varepsilon_1$ sufficiently small. This gives us the result for the first case as desired.

The second case we consider is $j - l \geq 2$. It holds that

$$\sum_{\substack{\text{snakes } A,\, B:\\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})| \geq l+2}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{5(j-l+1)} \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-j-3} \cdot p_1^{2(j-l+1)}.$$

As before, Lemma 4.11 tells us that for any $\varepsilon \in (0,1)$ we can choose $\varepsilon_1 \in (0,1)$ small enough so that

$$\mathbb{E}[\,X_t\,]^2 \geq (1-\varepsilon) \cdot \frac{1}{4} \cdot m^{4t} \cdot \left(\sum_{i=1}^n p_i^4\right)^2 \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-4}.$$

Thus,

$$\sum_{\substack{\text{snakes } A,\, B:\\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})| \geq l+2}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (3t)^{5(j-l+1)} \cdot m^{-l} \cdot \left(\sum_{i=1}^n p_i^2\right)^{-j+1} \cdot \left(\sum_{i=1}^n p_i^4\right)^{-1} \cdot p_1^{2(j-l+1)} \cdot \mathbb{E}[\,X_t\,]^2$$

and since $\sum_{i=1}^n p_i^4 \geq p_1^4$, we get

$$\leq 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (3t)^{5(j-l+1)} \cdot \left(m \cdot \left(\sum_{i=1}^n p_i^2\right)\right)^{-l} \cdot \frac{p_1^{2(j-l+1)}}{p_1^4 \cdot \left(\sum_{i=1}^n p_i^2\right)^{j-l-1}} \cdot \mathbb{E}[\,X_t\,]^2.$$

Again, we can use $m = \varepsilon_m \cdot \left(\sum_{i=1}^n p_i^2\right)^{-1}$ to get

$$= 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (3t)^{5(j-l+1)} \cdot \varepsilon_m^{-l} \cdot \frac{p_1^{2(j-l-1)}}{\left(\sum_{i=1}^n p_i^2\right)^{j-l-1}} \cdot \mathbb{E}[\,X_t\,]^2$$

and $f = p_1^2 / \left(\sum_{i=1}^n p_i^2\right)$, which yields

$$= 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot (3t)^{5(j-l+1)} \cdot \varepsilon_m^{-l} \cdot f^{-(j-l-1)} \cdot \mathbb{E}[\,X_t\,]^2.$$

Since we know that $t = f^{1/78}$ we get

$$= 16 \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot \varepsilon_m^{-l} \cdot (3t)^{10} \cdot \left(\frac{(3t)^5}{t^{78}}\right)^{j-l-1} \cdot \mathbb{E}[X_t]^2.$$

Since we know that $j - l \geq 2$ and $\varepsilon_m > 1$, we can make this expression at most $\varepsilon_E \cdot \mathbb{E}[X_t]^2/t^2$ for any $\varepsilon_E > 0$ by making $t$ sufficiently large. The same holds if we make $\varepsilon_1$ sufficiently small, because $t = f^{1/78} \geq {\varepsilon_1}^{-1/78}$. As we do so, $\varepsilon$ decreases as well.

The last case we consider is $j - l = 1$. This happens if we either only have one connected component in $G_{F_A \cap F_B}$ that does not form a cycle or if $G_{F_A \cap F_B}$ contains a cycle and one other connected component. In the latter case, equation (4.21) gives us

$$\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1 \\ \text{cycle in } G_{F_A \cap F_B}}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{5(j-l+1)} \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-j-3} \cdot p_1^{2(j-l+1)}$$

$$= 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{10} \cdot \left(\sum_{i=1}^n p_i^4\right) \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-l-4} \cdot p_1^4,$$

where we can choose the value of $\varepsilon \in (0,1)$ by making $\varepsilon_1$ sufficiently small. As before, we can use the estimate

$$\mathbb{E}[X_t]^2 \geq (1-\varepsilon) \cdot \frac{1}{4} \cdot m^{4t} \cdot \left(\sum_{i=1}^n p_i^4\right)^2 \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-4}.$$

from Lemma 4.11 to achieve an upper bound of

$$\leq 16 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot t^{10} \cdot \left(m \cdot \sum_{i=1}^n p_i^2\right)^{-l} \cdot \frac{p_1^4}{\sum_{i=1}^n p_i^4} \cdot \mathbb{E}[X_t]^2.$$

Since a cycle can only exist for $l \geq t$, due to the requirement $m = \varepsilon_m / \sum_{i=1}^n p_i^2$ for some $\varepsilon_m > 1$, and with $p_1^4 \leq \sum_{i=1}^n p_i^4$ it holds that this is

$$\leq 16 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot t^{10} \cdot \varepsilon_m^{-t} \cdot \mathbb{E}[X_t]^2.$$

As in the case of $j - l = 0$, where $G_{F_A \cap F_B}$ also contained a cycle, we see that for any $\varepsilon_E > 0$ we can bound this expression by $\varepsilon_E \cdot \mathbb{E}[X_t]^2/t^2$ as desired by making $\varepsilon_1$ sufficiently small, which makes $t$ sufficiently large.

If $j - l = 1$ and $G_{F_A \cap F_B}$ does not contain a cycle, we have to look a bit more closely, since we cannot guarantee a large enough $l$ to make the expression sufficiently small. Instead, we will consider different cases for mapping the central variable of $B$. These cases will result in slightly better bounds than the one in equation (4.21).

First, we assume that $B$'s central variable is a free variable, i. e. the central variable of $B$ does not appear in *any* shared clauses of $A$ and $B$. This means, we can actually choose $B$'s central variable freely and it will appear at least 4 times in $F_A \cup F_B$. In equation (4.21) we assumed that each of our free variables only contributed $\sum_{i=1}^{n} p_i^2$. However, in the current case, one of them (the central one) contributes $\sum_{i=1}^{n} p_i^4$. Thus, we can substitute a factor of $(\sum_{i=1}^{n} p_i^2) \cdot p_1^2$ in equation (4.20) with $\sum_{i=1}^{n} p_i^4$ to get

$$\begin{aligned}
&\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1, \\ \text{central of } B \text{ is free}}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{10} \cdot \left(\sum_{i=1}^{n} p_i^4\right)^2 \cdot \left(\sum_{i=1}^{n} p_i^2\right)^{4t-l-5} \cdot p_1^2. \qquad (4.22)
\end{aligned}$$

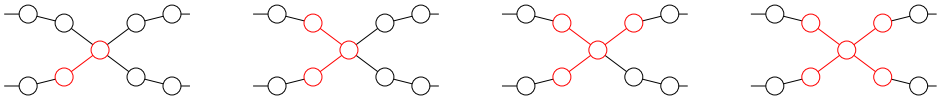
As in the cases before, we use the lower bound on $\mathbb{E}[X_t]$ from Lemma 4.11 and our definition $(\sum_{i=1}^{n} p_i^2)/p_1^2 = f = t^{78}$ to get

$$\begin{aligned}
&\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1, \\ \text{central of } B \text{ not in } F_A \cap F_B}} \Pr[X_A = 1 \wedge X_B = 1] \\
&\leq 16 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot t^{10} \cdot \left(m \cdot \sum_{i=1}^{n} p_i^2\right)^{-l} \cdot \frac{p_1^2}{\sum_{i=1}^{n} p_i^2} \cdot \mathbb{E}[X_t]^2 \\
&= 4 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot \varepsilon_m^{-l} \cdot t^{10} \cdot f^{-1} \cdot \mathbb{E}[X_t]^2 \\
&= 4 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot \varepsilon_m^{-l} \cdot t^{-68} \cdot \mathbb{E}[X_t]^2.
\end{aligned}$$

It is obvious that for any $\varepsilon_E > 0$ this is at most $\varepsilon_E \cdot \mathbb{E}[X_t]^2/t^2$ as desired if we choose $t \geq {\varepsilon_1}^{-1}$ sufficiently large or, conversely, $\varepsilon_1$ sufficiently small.

Now we assume that the central variable in $B$ is not free. What could happen? It could coincide with a non-central variable from $A$ or with the central variable from $A$. Thus, the central variable of $B$ could already appear once or twice in shared clauses in the first and one to four times in the second case.

Let us start with the case that it coincides with a non-central variable in $A$. Then, one of the variables that appears twice in $A$ appears an additional (not in shared clauses) 2 or 3 times as the central node in $B$, depending on the number of shared clauses it already appears in. In total it either appears 4 times or 5



**(a)** The central variable of $B$ appears in one shared clause.

**(b)** The central variable of $B$ appears in two shared clauses.

**(c)** The central variable of $B$ appears in three shared clauses.

**(d)** The central variable of $B$ appears in four shared clauses.

**Figure 4.3:** Snapshot of $B$'s central variable in $G_{F_B}$. Shared clauses of $F_A$ and $F_B$ are highlighted in red. If the central variable appears in $x$ shared clauses, then there are $x$ variables that appear exactly once in shared clauses. Then, $B$'s central variable appears an additional $4 - x$ times in $B$ and the variables that appear only once in shared clauses, each appear one additional time in $B$.

times in $F_A \cup F_B$. For a representation of those two cases, see Figure 4.3 (a) and Figure 4.3 (b).

Thus, we can replace the two appearances of a variable in $A$ and 2 resp. 3 appearances of unfree variables in $B$ with 4 resp. 5 appearances of a variable in total ($A$ and $B$). That is, we multiply the expression from equation (4.21) with $(\sum_{i=1}^n p_i^4)/(p_1^2 \cdot \sum_{i=1}^n p_i^2)$ resp. $(\sum_{i=1}^n p_i^5)/(p_1^3 \cdot \sum_{i=1}^n p_i^2)$. Since $\sum_{i=1}^n p_i^5 \leq p_1 \sum_{i=1}^n p_i^4$, the former case gives us an upper bound. We get

$$\sum_{\substack{\text{snakes } A,\, B:\\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1,\\ \text{central of } B \text{ not free and not central of } A}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{10} \cdot \left(\sum_{i=1}^n p_i^4\right)^2 \cdot \left(\sum_{i=1}^n p_i^2\right)^{4t-l-5} \cdot p_1^2.$$

This is the same upper bound we had in the previous case, equation (4.22), when the central variable of $B$ was free. Thus, we already know that for any $\varepsilon_E > 0$ we can choose $\varepsilon_1$ small enough to get a bound of at most $\varepsilon_E \cdot \mathbb{E}[X_t]^2/t^2$.

The last case is that the central variable of $B$ coincides with the central variable of $A$. Then, the variable that appears 4 times in $A$ might appear 0 to 3 additional times (i. e. not in shared clauses) in $B$, depending on the number of shared clauses it already appears in. It cannot appear an additional 4 times, since the central variable of $A$ must appear in a shared clause at least once for the variable to not be free. Remember that we are in the case where $G_{F_A \cap F_B}$ only contains one connected component that is not a cycle. This means, we have at most 4 variables in shared clauses that each appear one additional time in $B$. See Figure 4.3 for a visual representation of those cases. Let $x \in \{1, 2, 3, 4\}$ be the number of times that the central variable of $A$ appears in shared clauses. Then, it appears an additional $4 - x$ times in $B$. In addition to the central variable, there are now $x$ other unfree variables that each appear one additional time in $B$. Each of these variables actually appears 3 times in $A$ and $B$ together

instead of 2 times in $A$ and once as a single predetermined variable in $B$. As before, we can substitute their appearances by multiplying equation (4.21) with a factor of $(\sum_{i=1}^{n} p_i^3)/(p_1 \cdot \sum_{i=1}^{n} p_i^2)$ for each of them. By handling the shared central variable of $A$ and $B$ in the same way, we get an additional factor of $(\sum_{i=1}^{n} p_i^{8-x})/(p_1^{4-x} \sum_{i=1}^{n} p_i^4)$. We now get

$$\sum_{\substack{\text{snakes } A,\, B:\\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1,\\ \text{central of } B \text{ is central of } A,\ \text{appears in } x \text{ shared clauses}}} \Pr[X_A = 1 \wedge X_B = 1]$$

$$\leq 4 \cdot (1+\varepsilon) \cdot m^{4t-l} \cdot (3t)^{10} \cdot \left(\sum_{i=1}^{n} p_i^{8-x}\right) \cdot \left(\sum_{i=1}^{n} p_i^2\right)^{4t-l-4-x} \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}.$$

Again, we can use the lower bound on $\mathbb{E}[X_t]$ from Lemma 4.11 to get

$$\leq 16 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot t^{10} \cdot \left(m \cdot \sum_{i=1}^{n} p_i^2\right)^{-l} \cdot \frac{\left(\sum_{i=1}^{n} p_i^{8-x}\right) \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}}{\left(\sum_{i=1}^{n} p_i^4\right)^2 \left(\sum_{i=1}^{n} p_i^2\right)^{x}} \cdot \mathbb{E}[X_t]^2$$

and $m \geq 1/\sum_{i=1}^{n} p_i^2$ implies

$$\leq 16 \cdot 3^{10} \cdot \frac{1+\varepsilon}{1-\varepsilon} \cdot t^{10} \cdot \frac{\left(\sum_{i=1}^{n} p_i^{8-x}\right) \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}}{\left(\sum_{i=1}^{n} p_i^4\right)^2 \left(\sum_{i=1}^{n} p_i^2\right)^{x}} \cdot \mathbb{E}[X_t]^2.$$

It remains to show that for any $\varepsilon_E > 0$ we can choose $\varepsilon_1$ small enough so that

$$t^{10} \cdot \frac{\left(\sum_{i=1}^{n} p_i^{8-x}\right) \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}}{\left(\sum_{i=1}^{n} p_i^4\right)^2 \left(\sum_{i=1}^{n} p_i^2\right)^{x}} \leq \frac{\varepsilon_E}{t^2}.$$

First, note that $\sum_{i=1}^{n} p_i^{8-x} \leq p_1^{4-x} \cdot \sum_{i=1}^{n} p_i^4$ and thus

$$t^{10} \cdot \frac{\left(\sum_{i=1}^{n} p_i^{8-x}\right) \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}}{\left(\sum_{i=1}^{n} p_i^4\right)^2 \left(\sum_{i=1}^{n} p_i^2\right)^{x}} \leq t^{10} \cdot \frac{p_1^{4-x} \cdot \left(\sum_{i=1}^{n} p_i^3\right)^{x}}{\left(\sum_{i=1}^{n} p_i^4\right) \left(\sum_{i=1}^{n} p_i^2\right)^{x}}$$

In order to further bound this expression, we consider the probability vector $\vec{p}^{(n)} = p_1, p_2, \ldots, p_n$. We now split the probabilities into those with $p_i \geq p_1/f^{1/6}$ and those with $p_i < p_1/f^{1/6}$. Let $N = |\{i \in [n] \mid p_i \geq p_1/f^{1/6}\}|$ be the number of probabilities in $\vec{p}^{(n)}$ larger than the bound we set. We now distinguish two cases: $N \geq f^{5/6}$ and $N < f^{5/6}$.

Assume the first case, $N \geq f^{5/6}$. It holds that

$$\sum_{i=1}^{n} p_i^4 \geq N \cdot \left(\frac{p_1}{f^{1/6}}\right)^4 = p_1^4 \cdot f^{1/6}.$$

Together with $\sum_{i=1}^{n} p_i^3 \leq p_1 \cdot \sum_{i=1}^{n} p_i^2$ this implies

$$t^{10} \cdot \frac{p_1^{4-x} \cdot \left(\sum_{i=1}^{n} p_i^3\right)^x}{\left(\sum_{i=1}^{n} p_i^4\right)\left(\sum_{i=1}^{n} p_i^2\right)^x} \leq t^{10} \cdot \frac{p_1^4 \cdot \left(\sum_{i=1}^{n} p_i^2\right)^x}{p_1^4 \cdot f^{1/6}\left(\sum_{i=1}^{n} p_i^2\right)^x}$$
$$= t^{10} \cdot f^{-1/6} = t^{-3} \leq \varepsilon_1^{1/78} \cdot t^{-2}$$

as desired, due to our choice $t = f^{1/78}$ and since we can make $\varepsilon_1$ as small as necessary.

Now assume $N < f^{5/6}$. It holds that

$$\sum_{i=1}^{n} p_i^3 < N \cdot p_1^3 + \frac{p_1}{f^{1/6}} \cdot \sum_{i=1}^{n} p_i^2$$
$$\leq p_1^3 \cdot f^{5/6} + p_1^3 \cdot f^{5/6} = 2 \cdot p_1^3 \cdot f^{5/6},$$

where we used $\sum_{i=1}^{n} p_i^2 = f \cdot p_1^2$. With $\sum_{i=1}^{n} p_i^4 \geq p_1^4$ and $\sum_{i=1}^{n} p_i^2 = f \cdot p_1^2$ this readily implies

$$t^{10} \cdot \frac{p_1^{4-x} \cdot \left(\sum_{i=1}^{n} p_i^3\right)^x}{\left(\sum_{i=1}^{n} p_i^4\right)\left(\sum_{i=1}^{n} p_i^2\right)^x} \leq t^{10} \cdot \frac{p_1^{4-x} \cdot \left(2 \cdot p_1^3 \cdot f^{5/6}\right)^x}{p_1^4 \cdot \left(p_1^2 \cdot f\right)^x}$$
$$\leq t^{10} \cdot 2^4 \cdot f^{-x/6} \leq t^{10} \cdot 2^4 \cdot f^{-1/6} \leq 16 \cdot \varepsilon_1^{1/78} \cdot t^{-2}.$$

Again we can make this as small as any $\varepsilon_E / t^2$ if we choose $\varepsilon_1$ sufficiently small.

Finally, we took care of all the cases for $j - l = 1$ and showed

$$\sum_{\substack{\text{snakes } A, B: \\ |E(G_{F_A \cap F_B})|=l,\ |V(G_{F_A \cap F_B})|=l+1}} \Pr[X_A = 1 \wedge X_B = 1] \leq \varepsilon_E \cdot \frac{\mathbb{E}[X_t]^2}{t^2}$$

as desired. This implies

$$\sum_{A} \sum_{B:\, B \sim A} \Pr[X_A = 1 \wedge X_B = 1] \leq \varepsilon_E \cdot \mathbb{E}[X_t]^2$$

and concludes the proof. ∎

Lemma 4.12 and Lemma 4.3 now establish the existence of a sharp threshold at $m = \left(\sum_{i=1}^{n} p_i^2\right)^{-1}$ as we will see in Section 4.5. However, we first have to show an upper bound for $p_1^2 \in \Theta(\sum_{i=1}^{n} p_i^2)$ and $p_2^2 \in \Theta(\sum_{i=2}^{n} p_i^2)$, or more generally, for the case that we are given constants $\varepsilon_1, \varepsilon_2 \in (0, 1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ and $p_2^2 \geq \varepsilon_2 \cdot \sum_{i=2}^{n} p_i^2$. This case will be handled in the next section.

## 4.4 A Simple Upper Bound on the Satisfiability Threshold

This section handles the case that there are constants $\varepsilon_1, \varepsilon_2 \in (0,1)$ with $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and $p_2^2 \geq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$. This especially includes $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in \Theta(\sum_{i=2}^n p_i^2)$. This case is particularly easy, since it implies $(C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1} \in \Theta(q_{\max}^{-1})$. That means, the probability for a formula to be unsatisfiable is dominated by the highest clause probability.

We are going to show that there is a coarse threshold at $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1} \in \Theta(q_{\max}^{-1})$. Note that Lemma 4.4 only assumes $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. Thus, the lemma already handles $m < \varepsilon_m \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ for sufficiently small constants $\varepsilon_m \in (0,1)$. Now we only have to see what happens for $m \in \Omega(m^\star)$.

In the following lemma we give a lower bound on the probability to generate an unsatisfiable instance by showing the existence of an unsatisfiable sub-formula consisting only of clauses with the highest clause probability. These are the clauses consisting of the two most-probable Boolean variables. The lemma generally holds for $k \geq 2$, but it especially serves our purpose of considering this easy case.

▶ **Lemma 4.13.** Let $\Phi \sim \mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ be a non-uniform random k-SAT formula and let $q_{\max}$ denote the maximum clause probability. Then, $\Phi$ is unsatisfiable with probability at least

$$(1 - e^{-q_{\max}\cdot m})^{2^k} - q_{\max}^2 \cdot 2^{2k} \cdot m \cdot (1 + e^{-q_{\max}\cdot m})^{2^k}.$$

◀

*Proof.* Let $c$ be the clause with maximum probability. Since the signs of literals are chosen with probability $1/2$ independently at random, it holds that each clause with the same variables as $c$ has the same probability. Our lower bound is now just a lower bound on the probability of having each of the $2^k$ clauses with these variables, which constitute an unsatisfiable sub-formula. Let us enumerate the different clauses $c_1, \ldots, c_{2^k}$ with variables $X_1, \ldots, X_k$ in an arbitrary order. Now let $\overline{A_j}$ denote the event that $c_j$ is *not* appearing in $\Phi$ and let $\overline{A} = \bigcup_{j\in[2^k]} \overline{A_j}$ denote the event that at least one of these clauses does not appear. Due to the principle of inclusion and exclusion it holds that

$$\Pr\left[\overline{A}\right] = \sum_{l=1}^{2^k} (-1)^{l+1} \sum_{J\subseteq[2^k]\colon |J|=l} \Pr\left[\bigcap_{j\in J} \overline{A_j}\right] = \sum_{l=1}^{2^k} (-1)^{l+1} \left(\binom{2^k}{l} \cdot (1 - l \cdot q_{\max})^m\right),$$

because the clauses $c_1, \ldots, c_{2^k}$ have the same probability $q_{\max}$ of appearing and

all clauses are drawn independently at random. It now holds that

$$\Pr[\Phi \text{ unsat}] \geq \Pr[A] = 1 - \left( \sum_{l=1}^{2^k} \binom{2^k}{l} \cdot (-1)^{l+1} \cdot (1 - l \cdot q_{\max})^m \right)$$
$$= \sum_{l=0}^{2^k} \left( \binom{2^k}{l} \cdot (-1)^l \cdot (1 - l \cdot q_{\max})^m \right).$$

We can now estimate

$$-(1 - q_{\max} \cdot l)^m \geq -e^{-q_{\max} \cdot l \cdot m}$$

and, due to [MR99, Proposition B.3],

$$(1 - q_{\max} \cdot l)^m \geq e^{-q_{\max} \cdot l \cdot m} \cdot \left(1 - q_{\max}^2 \cdot l^2 \cdot m\right) \geq e^{-q_{\max} \cdot l \cdot m} \cdot \left(1 - q_{\max}^2 \cdot 2^{2k} \cdot m\right).$$

In total, we get

$$\begin{aligned} &\Pr[\Phi \text{ unsat}] \\ &\geq \sum_{l=0}^{2^k} \left( \binom{2^k}{l} \cdot (-1)^l \cdot e^{-q_{\max} \cdot l \cdot m} - \binom{2^k}{l} \cdot q_{\max}^2 \cdot 2^{2k} \cdot m \cdot e^{-q_{\max} \cdot l \cdot m} \right) \\ &= (1 - e^{-q_{\max} \cdot m})^{2^k} - q_{\max}^2 \cdot 2^{2k} \cdot m \cdot (1 + e^{-q_{\max} \cdot m})^{2^k}. \end{aligned}$$

■

Note that the former lemma implies the statement we want only if $q_{\max} \in o(1)$. Since

$$q_{\max} = \frac{\prod_{i=1}^{k} p_i}{2^k \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j},$$

it is also a function in $n$. For $k = 2$ the expression simplifies to $(p_1 \cdot p_2)/(2 \cdot (1 - \sum_{i=1}^{n} p_i^2))$. More generally than $q_{\max} \in o(1)$, we will now assume that we can choose an $\varepsilon_q \in (0, 1/2^k)$ so that $q_{\max} \leq \varepsilon_q$. We will handle the case $q_{\max} \notin o(1)$ afterward. The former lemma now yields the following corollary.

▶ **Corollary 4.14.** Let $(\vec{p}^{(n)})_{n \in \mathbb{N}}$ be an ensemble of probability distributions. Let $\Phi \sim \mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ be a non-uniform random $k$-SAT formula. Then,

1. for any $\varepsilon_P \in (0, (1 - e^{-\varepsilon_m})^{2^k})$ and for any $\varepsilon_m > 0$ so that $m = \varepsilon_m / q_{\max}$ we can choose $\varepsilon_q \in (0, 1/2^k)$ with $q_{\max} \leq \varepsilon_q$ sufficiently small so that $\Pr[\Phi \text{ unsatisfiable}] \geq \varepsilon_P$.

2. for any $\varepsilon_P \in (0, 1)$, we can choose $\varepsilon_m > 0$ with $m = \varepsilon_m / q_{\max}$ sufficiently large and $\varepsilon_q \in (0, 1/2^k)$ with $q_{\max} \leq \varepsilon_q$ sufficiently small so that $\Pr[\Phi \text{ unsatisfiable}] \geq \varepsilon_P$.

◀

*Proof.* Let $m^\star = q_{\max}^{-1}$ and fix a constant $\varepsilon_m > 0$ so that $m = \varepsilon_m \cdot m^\star$ and a constant $\varepsilon_P \in (0, (1-e^{-\varepsilon_m})^{2^k})$. Lemma 4.13 tells us that

$$\begin{aligned}\Pr[\Phi \text{ unsatisfiable}] &\geq (1-e^{-q_{\max}\cdot m})^{2^k} - q_{\max}^2 \cdot 2^{2k} \cdot m \cdot (1+e^{-q_{\max}\cdot m})^{2^k} \\ &= (1-e^{-\varepsilon_m})^{2^k} - q_{\max} \cdot 2^{2k} \cdot \varepsilon_m \cdot (1+e^{-\varepsilon_m})^{2^k} \\ &\geq (1-e^{-\varepsilon_m})^{2^k} - \varepsilon_q \cdot 2^{2k} \cdot \varepsilon_m \cdot (1+e^{-\varepsilon_m})^{2^k}.\end{aligned}$$

If we choose $\varepsilon_q$ sufficiently small, we can reach any value $\varepsilon_P < (1-e^{-\varepsilon_m})^{2^k}$ as desired.

Now we turn to the case that we are only given $\varepsilon_P \in (0,1)$. It still holds that

$$\Pr[\Phi \text{ unsatisfiable}] \geq (1-e^{-\varepsilon_m})^{2^k} - \varepsilon_q \cdot 2^{2k} \cdot \varepsilon_m \cdot (1+e^{-\varepsilon_m})^{2^k}.$$

We can see that if we choose $\varepsilon_m$ sufficiently large and $\varepsilon_q$ sufficiently small, we can make this expression at least $\varepsilon_P$. ■

This lemma already captures the case $q_{\max} \in o(1)$. Let us now assume that there is some $\varepsilon_q \in (0, 1/2^k)$ so that $q_{\max} \geq \varepsilon_q$. It then holds that $m^\star = q_{\max}^{-1} \leq 1/\varepsilon_q$. Remember that $q_{\max} \leq 1/2^k$ also still holds. This means, the threshold function is bounded by a constant. It is easy to see that for $\Phi \sim \mathcal{D}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m\big)$ and a constant $m \geq 2^k$ it holds that $\Pr[\Phi \text{ unsatisfiable}] \geq q_{\max}^m \geq \varepsilon_q^m$, since this is the probability of an unsatisfiable instance, where the most probable clause appears with all $2^k$ combinations of signs and then one of these clauses appears an additional $m - 2^k$ times. Similarly, $\Pr[\Phi \text{ satisfiable}] \geq q_{\max}^m \geq \varepsilon_q^m$, as this is the probability of a satisfiable instance, where the same most probable clause appears $m$ times with the same sign. Since $0 < q_{\max} \leq 1/2^k$ is a constant, the probability is a constant bounded away from zero and one.

It remains to show that $\Phi$ is unsatisfiable with probability $1-o(1)$ for $m \in \omega(1)$. More generally, we want to show that for any $\varepsilon_P \in (0,1)$ we can choose an $\varepsilon_m > 0$ with $m = \varepsilon_m \cdot m^\star$ large enough so that $\Phi$ is unsatisfiable with probability at least $\varepsilon_P$. The following lemma implies this. Again, this lemma also holds for $k \geq 2$ in general and without assuming anything for $q_{\max}$.

▶ **Lemma 4.15.** Consider a non-uniform random k-SAT formula $\Phi$. Then $\Phi$ is unsatisfiable with probability at least

$$2 - (1+\exp(-q_{\max}\cdot m))^{2^k}.$$

◀

*Proof.* As in Lemma 4.13, it holds that

$$\Pr[\Phi \text{ unsat}] \geq \sum_{l=0}^{2^k} \left( \binom{2^k}{l} (-1)^l (1 - l \cdot q_{\max})^m \right).$$

We can now estimate

$$\begin{aligned} \sum_{l=0}^{2^k} \left( \binom{2^k}{l} (-1)^l (1 - l \cdot q_{\max})^m \right) &\geq 1 - \sum_{l=1}^{2^k} \left( \binom{2^k}{l} (1 - l \cdot q_{\max})^m \right) \\ &\geq 1 - \sum_{l=1}^{2^k} \left( \binom{2^k}{l} \exp(-m \cdot l \cdot q_{\max}) \right) \\ &= 2 - (1 + \exp(-m \cdot q_{\max}))^{2^k} \end{aligned}$$

■

We can now see that our desired statement holds. The former implies it if we can choose $\varepsilon_m$ large enough.

▶ **Corollary 4.16.** Let $(\vec{p}^{(n)})_{n \in \mathbb{N}}$ be an ensemble of probability distributions. For any constant $\varepsilon_P \in (0, 1)$ we can choose a constant $\varepsilon_m > 0$ with $m = \varepsilon_m / q_{\max}$ sufficiently large so that the probability to generate an unsatisfiable formula $\Phi \sim \mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, m)$ is at least $\varepsilon_P$. ◀

*Proof.* Lemma 4.15 tells us

$$\Pr[\Phi \text{ unsatisfiable}] \geq 2 - (1 + \exp(-m \cdot q_{\max}))^{2^k} = 2 - (1 + \exp(-\varepsilon_m))^{2^k},$$

since $m \cdot q_{\max} = \varepsilon_m$. We can now simply make $\varepsilon_m$ large enough to make this expression at least $\varepsilon_P$. ■

## 4.5 Putting it All Together

In this section we put the upper and lower bounds of the previous sections together. This will show our main result of this chapter, the existence and sharpness of a satisfiability threshold for non-uniform random 2-SAT depending on the ensemble of probability distributions $(\vec{p}^{(n)})_{n \in \mathbb{N}}$. We could have unified the proof to capture all cases with $p_1^2 \notin o(\sum_{i=1}^n p_i^2)$. However, the current setting has the advantage that we can also state reasons for the threshold being coarse in the second and third case. In the second case it is due to the emergence of a snake of size 2, i. e. an unsatisfiable sub-formula that looks like this

$$(w_2, w_1), (\overline{w_1}, w_2), (\overline{w_2}, w_3), (\overline{w_3}, \overline{w_2})$$

for literals $w_1$, $w_2$, and $w_3$ of distinct Boolean variables. In the third case, the coarseness comes from the emergence of an unsatisfiable sub-formula, where the clause with the two most probable variables appears with all four combinations of signs.

▶ **Theorem 4.17.** Given an ensemble of probability distributions $\left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}}$.

1. If $p_1^2 \in o(\sum_{i=1}^n p_i^2)$, then non-uniform random 2-SAT has a sharp satisfiability threshold at $m^\star = 1/\sum_{i=1}^n p_i^2$.
2. If $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in o\left(\sum_{i=2}^n p_i^2\right)$, then non-uniform random 2-SAT has a coarse satisfiability threshold at $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. Furthermore, for any large enough $n$ there is a range of size $\Theta(m^\star)$ around the threshold, where the probability to generate satisfiable instances is bounded away from zero and one.
3. If $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in \Theta\left(\sum_{i=2}^n p_i^2\right)$, then non-uniform random 2-SAT has a coarse satisfiability threshold at $m^\star = (q_{\max}^{-1}) \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$. Furthermore, for any large enough $n$ there is a range of size $\Theta(m^\star)$ around the threshold, where the probability to generate satisfiable instances is bounded away from zero and one.
4. Otherwise, non-uniform random 2-SAT has a coarse satisfiability threshold at $m^\star = (C \cdot \sum_{i=2}^n p_i^2 + C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$.

◀

*Proof.* Remember our discussion in Section 4.1. We want to show that $m^\star$ is an asymptotic threshold function of non-uniform random 2-SAT with respect to parameter $m$. This means:

1. for any function $m\colon \mathbb{N} \to \mathbb{R}^+$ with $m \in o(m^\star)$ and any $\varepsilon_P \in (0,1)$ there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ the probability to generate a satisfiable instance is at least $\varepsilon_P$.
2. and for all $m\colon \mathbb{N} \to \mathbb{R}^+$ with $m \in \omega(m^\star)$ and any $\varepsilon_P \in (0,1)$ there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ the probability to generate an unsatisfiable instance is at least $\varepsilon_P$.

If we want to show a sharp threshold, we have to certify that:

1. for any given constant $\varepsilon_m \in (0,1)$, any function $m\colon \mathbb{N} \to \mathbb{R}^+$ with $m \leq \varepsilon_m \cdot m^\star$, and any $\varepsilon_P \in (0,1)$ there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ the probability to generate a satisfiable instance is at least $\varepsilon_P$.
2. and for any given constant $\varepsilon_m > 1$, all $m\colon \mathbb{N} \to \mathbb{R}^+$ with $m \geq \varepsilon_m \cdot m^\star$, and any $\varepsilon_P \in (0,1)$ there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ the probability to generate an unsatisfiable instance is at least $\varepsilon_P$.

**Case 1: $p_1^2 \in o(\sum_{i=1}^n p_i^2)$** The first case we consider is $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. We want to show a sharp threshold at $m^\star = 1/\sum_{i=1}^n p_i^2$. The requirement $p_1^2 \in o(\sum_{i=1}^n p_i^2)$ implies that we can choose any $\varepsilon_1 \in (0,1)$ and for some $n_0 \in \mathbb{N}$ it holds that $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ for all $n \geq n_0$. Thus, Lemma 4.3 directly implies the first requirement for sharpness and Lemma 4.12 directly implies the second requirement.

**Case 2: $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in o\left(\sum_{i=2}^n p_i^2\right)$** The second case we consider is $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in o\left(\sum_{i=2}^n p_i^2\right)$. We want to show that $m^\star = (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ is a coarse satisfiability threshold. The requirements imply that there is some $\varepsilon_1 \in (0,1)$ and that we can choose an $\varepsilon_2 \in (0,1)$ so that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ hold simultaneously for all sufficiently large $n$. If $m \in o(m^\star)$, then for any $\varepsilon_m \in (0,1)$ there is an $n_0 \in \mathbb{N}$ so that $m \leq \varepsilon_m \cdot m^\star$ for all $n \geq n_0$. Thus, we can apply Lemma 4.4 to certify the first condition on $m^\star$ being an asymptotic threshold function. Equivalently, if $m \in \omega(m^\star)$, then for any $\varepsilon_m > 1$, there is an $n_0 \in \mathbb{N}$ so that $m \geq \varepsilon_m \cdot m^\star$ for all $n \geq n_0$. We can now apply Corollary 4.10. Note that the lemma assumes $m = \varepsilon_m \cdot m^\star$. However, since the probability to generate satisfiable instances is non-increasing in $m$ (c. f. Lemma 3.8), it suffices to consider $m' = \varepsilon_m \cdot m$. The probability to generate satisfiable (unsatisfiable) instances at the actual number of clauses $m \geq m'$ can only be smaller (larger). Thus, Corollary 4.10 implies the second condition on $m^\star$ being an asymptotic threshold function.

It remains to show that the threshold is not sharp. Essentially, we are going to show that there is a non-empty range of $\varepsilon_m \in [\varepsilon_m^{(1)}, \varepsilon_m^{(2)}]$ for which the probability to generate satisfiable instances at $m = \varepsilon_m \cdot m^\star$ is bounded away from zero and one by a constant. If the threshold was sharp, at least one of the probabilities at positions $m^{(1)}$ and $m^{(2)}$ that are a constant factor apart would approach zero or one in the limit. Since in our case neither the probability at $m^{(1)} = \varepsilon_m^{(1)} \cdot m^\star$ nor the one at $m^{(2)} = \varepsilon_m^{(2)} \cdot m^\star$ does, the threshold must be coarse. First, Lemma 4.4 states that for any $\varepsilon_P \in (0,1)$ we can choose $\varepsilon_m \in (0,1)$ small enough so that the probability to generate a satisfiable instance at $m = \varepsilon_m \cdot m^\star$ is at least $\varepsilon_P$. We can now choose $\varepsilon_P^{(1)} > \varepsilon_P^{(2)}$. This will result in some $\varepsilon_m^{(1)} < \varepsilon_m^{(2)}$ so that the probability to generate a satisfiable instance is at least $\varepsilon_P^{(1)}$ at $m = \varepsilon_m^{(1)} \cdot m^\star$ and $\varepsilon_P^{(2)}$ at $m = \varepsilon_m^{(2)} \cdot m^\star$. However, Lemma 4.9 states that for the same values of $\varepsilon_m$ we can choose $\varepsilon_2$ with $p^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ small enough so that the probability to generate an unsatisfiable instance at $m = \varepsilon_m \cdot m^\star$ is at least $\varepsilon_P$ for any constant

$$\varepsilon_P < \frac{\varepsilon_m^4}{\varepsilon_m^4 + 3 \cdot \varepsilon_m^2 \left(1 + \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_1^2}\right) + 8}.$$

This requirement on $\varepsilon_2$ holds for all sufficiently large $n$, since $p_2^2 \in o(\sum_{i=2}^n p_i^2)$. Thus, for $\varepsilon_m^{(1)}$ and $\varepsilon_m^{(2)}$ both the probability to generate a satisfiable and the probability to generate an unsatisfiable instance are at least some constant

depending only on $\varepsilon_1$ and $\varepsilon_m$ if $n$ is large enough. Since both $\varepsilon_m$ and $\varepsilon_1$ are fixed, these probabilities cannot approach zero or one in the limit. This implies coarseness of the threshold as desired.

**Case 3: $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in \Theta\left(\sum_{i=2}^n p_i^2\right)$** The third case we consider is $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in \Theta\left(\sum_{i=2}^n p_i^2\right)$. We want to show a coarse satisfiability threshold at $m^\star = q_{\max}^{-1}$, where $q_{\max} = (C \cdot p_1 \cdot p_2)/2$ is the maximum clause probability. Note that in this case, $m^\star \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$. As before, we can apply Lemma 4.4 to certify the first condition on $m^\star$ being an asymptotic threshold function. The second condition is implied by our results in Section 4.4. The second statement of Corollary 4.14 certifies the second condition if $\varepsilon_q \in (0,1)$ with $q_{\max} \leq \varepsilon_q$ is sufficiently small and $\varepsilon_m > 0$ with $m = \varepsilon_m / q_{\max}$ is sufficiently large. If $q_{\max} \in o(1)$ and $m \in \omega(m^\star) = \omega(q_{\max}^{-1})$ both conditions hold for all sufficiently large values of $n$. If $q_{\max} \notin o(1)$, the second condition holds as follows. According to the second condition we are given an $m \in \omega(m^\star)$ and an $\varepsilon_P \in (0,1)$. We choose $\varepsilon_m$ sufficiently large and $\varepsilon_q$ sufficiently small so that we generate an unsatisfiable instance with probability at least $\varepsilon_P$ according to Corollary 4.14. Then, we fix that value of $\varepsilon_q$ and choose an $\varepsilon_m$ sufficiently large so that we generate an unsatisfiable instance with probability at least $\varepsilon_P$ according to Corollary 4.16. Since $m \in \omega(m^\star)$, we know that $m \geq \varepsilon_m / q_{\max}$ holds for both values of $\varepsilon_m$ we chose as soon as $n$ is sufficiently large. For all such values of $n$ we either have $q_{\max} \leq \varepsilon_q$ or $q_{\max} > \varepsilon_q$. Thus, the second condition holds either according to Corollary 4.14 or according to Corollary 4.16.

As in the previous case we have to rule out that the threshold is sharp. Again, we will show that there is a range of $m \in \Theta(m^\star)$ where the probability to generate satisfiable instances is bounded away from zero and one by constants. However, depending on whether or not $q_{\max} \in o(1)$, this range can be at different positions in $\Theta(m^\star)$. This is due to the fact that, if $q_{\max} \in \Omega(1)$, then $m^\star \in O(1)$. However, in order to have an unsatisfiable instance we need $m \geq 4$. At the same time Lemma 4.4 might require us to choose an $\varepsilon_m$ so small that this is not guaranteed anymore. Thus, in the case that $q_{\max} \in \Omega(1)$ we choose a different range of $\varepsilon_m$ with $m = \varepsilon_m \cdot m^\star$.

We start with $q_{\max} \in o(1)$. Now, note that $m^\star = q_{\max}^{-1} \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$. Thus, there are constants ${\varepsilon_l}^{(1)}, {\varepsilon_l}^{(2)} > 0$ such that ${\varepsilon_l}^{(1)} \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1} \leq m^\star \leq {\varepsilon_l}^{(2)} \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ for all sufficiently large values of $n$. We now choose $m^{(1)} = {\varepsilon_m}^{(1)} \cdot m^\star$ and $m^{(2)} = {\varepsilon_m}^{(2)} \cdot m^\star$ with ${\varepsilon_m}^{(1)} < {\varepsilon_m}^{(2)}$. It holds that $m^{(1)} \leq {\varepsilon_m}^{(1)} \cdot {\varepsilon_l}^{(2)} \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ and $m^{(2)} \leq {\varepsilon_m}^{(2)} \cdot {\varepsilon_l}^{(2)} \cdot (C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. Since Lemma 4.4 only requires $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$, we can now use it equivalently to the second case. That means, if we choose the constants ${\varepsilon_m}^{(1)}$ and ${\varepsilon_m}^{(2)}$ small enough, the probability to generate satisfiable instances at both number of clauses is at least a constant depending only on $\varepsilon_1$, ${\varepsilon_l}^{(2)}$ and $\varepsilon_m$, all of which are constant for sufficiently large $n$. For the same values of $m$ we want to have a constant lower bound on the probability to generate unsatisfiable instances. Again, our results from section 4.4 provide

us with these lower bounds. According to Corollary 4.14 it holds for both $m^{(1)}$ and $m^{(2)}$ that the probability to generate unsatisfiable instances can be lower bounded by a constant that only depends on $\varepsilon_m$ as soon as $\varepsilon_q \in (0, 1/2^k)$ with $q \leq \varepsilon_q$ is small enough. This holds for all sufficiently large $n$, since we assumed $q_{\max} \in o(1)$. Since $\varepsilon_m^{(1)}$ and $\varepsilon_m^{(2)}$ are fixed constants, the resulting probability is constant as well. This gives us the desired result if $q_{\max} \in o(1)$.

Now we consider $q_{\max} \in \Omega(1)$. It holds that $m^\star = 1/q_{\max} \in O(1)$. Then, we can simply choose any two constants $\varepsilon_m^{(1)}, \varepsilon_m^{(2)} > 1$ that are sufficiently far apart for $m^{(1)} = \varepsilon_m^{(1)} \cdot m^\star$ and $m^{(2)} = \varepsilon_m^{(1)} \cdot m^\star$ to be different integers. Both the probability to generate a satisfiable and an unsatisfiable instance are at least $q_{\max}^m$. Since $q_{\max} \in \Omega(1)$, $q_{\max}$ is lower-bounded by a constant for all sufficiently large $n$. The same holds for $m = \varepsilon_m \cdot m^\star$, since $q_{\max} \leq 1/2^k$ and $\varepsilon_m$ is some fixed constant as well. Thus, the probabilities to generate satisfiable and unsatisfiable instances at $m^{(1)}$ and $m^{(2)}$ are bounded away from zero and one as desired.

Last, we consider $q_{\max} \notin o(1)$ and $q_{\max} \notin \Omega(1)$. First, we choose ${\varepsilon_m}^{(1)}, {\varepsilon_m}^{(2)} > 0$ as before and $\varepsilon_q$ small enough so that the same bounds hold as in the case of $q_{\max} \in o(1)$. Then, we assume $q_{\max} \geq \varepsilon_q$ and choose $\varepsilon_m^{(1)}, \varepsilon_m^{(2)} > 1$ as in the case of $q_{\max} \in \Omega(1)$. This implies probabilities of at least $\varepsilon_q^{\varepsilon_m/\varepsilon_q}$ to generate a satisfiable/unsatisfiable instance. For all sufficiently large $n$ we either have $q_{\max} \leq \varepsilon_q$ or $q_{\max} > \varepsilon_q$. Thus, the threshold is coarse either way.

**Case 4: Otherwise** The last case we consider is that none of the three other cases hold. We are going to show that there is a coarse threshold at $m^\star = (C \cdot \sum_{i=2}^n p_i^2 + C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. The threshold function is chosen such that, depending on $p_1^2$ and $p_2^2$, either the first or the second term dominates. That means, if $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \sum_{i=1}^n p_i^2$ is small enough, then $m^\star \in \Theta((C \cdot \sum_{i=1}^n p_i^2)^{-1})$. In that case, we have an asymptotic threshold as if $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. Otherwise, $m^\star \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$. Then, we have an asymptotic threshold as if $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$. To make things easier, let us investigate $m \in o(m^\star)$ and $m \in \omega(m^\star)$ separately.

Let us start with $m \in o(m^\star)$. We are given an $\varepsilon_P \in (0,1)$ and have to assure that the probability to generate a satisfiable instance at $m$ is at least $\varepsilon_P$. Thus, we first choose some $\varepsilon_m \in (0,1)$ with $m = \varepsilon_m \cdot m^\star$. Furthermore, we assume that we can choose an $\varepsilon_1$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. We now want to apply Lemma 4.3. However, the lemma is stated with respect to the threshold function $(\sum_{i=1}^n p_i^2)^{-1}$. Thus, we first have to relate our $m^\star$ to this function, assuming that we can choose $\varepsilon_m$ and $\varepsilon_1$ arbitrarily small. It holds that

$$\sum_{i=1}^n p_i^2 = p_1^2 + \sum_{i=2}^n p_i^2 \leq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2 + \sum_{i=2}^n p_i^2.$$

Thus, $\sum_{i=2}^{n} p_i^2 \geq (1-\varepsilon_1) \cdot \sum_{i=1}^{n} p_i^2$ and therefore

$$m^\star = \left( C \cdot \sum_{i=2}^{n} p_i^2 + C \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{-1} \leq \frac{1}{1-\varepsilon_1} \cdot \left( C \cdot \sum_{i=1}^{n} p_i^2 \right)^{-1}.$$

Note that $C = 1/(1-\sum_{i=1}^{n} p_i^2) \geq 1$. For any fixed $\varepsilon_1$ this especially means $m \in o(m^\star)$ implies $m \in o((\sum_{i=1}^{n} p_i^2)^{-1})$. This allows us to apply Lemma 4.3 with $\varepsilon_m \in (0,1)$ and an $\varepsilon_1 \in (0,1)$ small enough to give us a probability of at least $\varepsilon_P$. The requirement $m \in o((\sum_{i=1}^{n} p_i^2)^{-1})$ guarantees that the condition on $\varepsilon_m$ is fulfilled. However, we can not guarantee that $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ holds. Thus, we now assume $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$. This is what we also assumed in the case $p_1^2 \in \Theta(\sum_{i=1}^{n} p_i^2)$ and in fact, we can use the same results now. That is, we can use Lemma 4.4. Again, we have to relate $m^\star$ to the threshold function $(C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$ the lemma uses. However, we can easily see that $m^\star \leq (C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$. Thus, any function $m \in o(m^\star)$ is also in $o((C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1})$. Lemma 4.4 states that for the value of $\varepsilon_1 \in (0,1)$ we have chosen before and the given value $\varepsilon_P$, we can now choose $\varepsilon_m \in (0,1)$ with $m \leq \varepsilon_m \cdot m^\star \leq \varepsilon_m \cdot (C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$ sufficiently small so that the probability to generate a satisfiable instance is at least $\varepsilon_P$. Thus, for all large enough values of $n$, both $m \leq \varepsilon_m \cdot (\sum_{i=1}^{n} p_i^2)^{-1}$ and $m \leq \varepsilon_m \cdot (C \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1}$ hold. Then, the probability of at least $\varepsilon_P$ at $m$ is guaranteed either by Lemma 4.3 if $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ or by Lemma 4.4 if $p_1^2 > \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$.

Let us now turn to $m \in \omega(m^\star)$. Given an $\varepsilon_P \in (0,1)$ we want to show that the probability to generate an unsatisfiable instance is at least $\varepsilon_P$ at $m$. Again, we assume that we can choose an $\varepsilon_1$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$. Then, we can apply Lemma 4.12. However, we first have to compare $m^\star$ to $(\sum_{i=1}^{n} p_i^2)^{-1}$ again. First, it holds that $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2 \leq \varepsilon_1 \cdot p_1$ and thus $p_1 \leq \varepsilon_1$. This implies $C = 1/(1-\sum_{i=1}^{n} p_i^2) \leq 1/(1-\varepsilon_1)$. It also implies

$$m^\star = \left( C \cdot \sum_{i=2}^{n} p_i^2 + C \cdot p_1 \cdot \left( \sum_{i=2}^{n} p_i^2 \right)^{1/2} \right)^{-1} \geq \left( 2 \cdot C \cdot \sum_{i=1}^{n} p_i^2 \right)^{-1} \geq \frac{1-\varepsilon_1}{2 \cdot \sum_{i=1}^{n} p_i^2},$$

since $p_1 \leq (\sum_{i=1}^{n} p_i^2)^{1/2}$ and $\sum_{i=2}^{n} p_i^2 \leq \sum_{i=1}^{n} p_i^2$. This means, $m \in \omega(m^\star)$ implies $m \in \omega(1/\sum_{i=1}^{n} p_i^2)$. We can now choose some $\varepsilon_m > 1$ with $m \geq \varepsilon_m / \sum_{i=1}^{n} p_i^2$ and apply Lemma 4.12 to show that the probability to generate an unsatisfiable instance at $\varepsilon_m / \sum_{i=1}^{n} p_i^2$ is at least $\varepsilon_P$ if $\varepsilon_1 \in (0,1)$ with $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ is small enough. Note that this probability only holds at $\varepsilon_m / \sum_{i=1}^{n} p_i^2$. However, due to the monotonicity of the probability function in our model (c. f. Lemma 3.8), it also holds for all $m \geq \varepsilon_m / \sum_{i=1}^{n} p_i^2$. Since $m \in \omega(1/\sum_{i=1}^{n} p_i^2)$, $m \geq \varepsilon_m / \sum_{i=1}^{n} p_i^2$ holds for all sufficiently large values of $n$. Up to this point we assumed $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ for the value $\varepsilon_1$ we needed in Lemma 4.12. Now we assume $p_1^2 > \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ for that same value $\varepsilon_1$. However, we have to make another distinction depending

on $p_2^2$. First, we assume $p_2^2 \leq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ for some $\varepsilon_2 \in (0,1)$ of our choice. We want to use Corollary 4.10 to show the bound we need. Again, we have to show that $m^\star = (C \cdot \sum_{i=2}^n p_i^2 + C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$ is large enough compared to $(C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1}$. It holds that $\sum_{i=2}^n p_i^2 \leq \sum_{i=1}^n p_i^2 \leq p_1^2/\varepsilon_1$. Thus, $C \cdot \sum_{i=2}^n p_i^2 \leq C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2}/\sqrt{\varepsilon_1}$ and

$$m^\star \geq \frac{1}{1+1/\sqrt{\varepsilon_1}} \cdot \left( C \cdot p_1 \cdot \left( \sum_{i=2}^n p_i^2 \right)^{1/2} \right)^{-1}.$$

Thus, for our fixed value $\varepsilon_1$ it holds that $m \in \omega(m^\star)$ implies $m \in \omega((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$. We can now apply Corollary 4.10 for some sufficiently large $\varepsilon_m > 0$ and some sufficiently small $\varepsilon_2 \in (0,1)$ to have a probability of at least $\varepsilon_P$ for generating an unsatisfiable instance. As with $p_1^2$, we now assume the contrary for $p_2^2$, i. e. $p_2^2 > \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ for the value $\varepsilon_2$ we just chose. We want to use Corollary 4.16 to show a probability of at least $\varepsilon_P$ for generating an unsatisfiable instance. The lemma holds if we have $m \geq \varepsilon_m/q_{\max}$ for an $\varepsilon_m > 0$ large enough. Under our assumptions $p_1^2 > \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ and $p_2^2 > \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$ it holds that

$$m^\star = \left( C \cdot \sum_{i=2}^n p_i^2 + C \cdot p_1 \cdot \left( \sum_{i=2}^n p_i^2 \right)^{1/2} \right)^{-1} \geq \left( \left( \frac{1}{\sqrt{\varepsilon_1 \cdot \varepsilon_2}} + \frac{1}{\sqrt{\varepsilon_2}} \right) \cdot 2 \cdot q_{\max} \right)^{-1},$$

because $\sum_{i=2}^n p_i^2 \leq (\sum_{i=1}^n p_i^2)^{1/2} \cdot (\sum_{i=2}^n p_i^2)^{1/2}$ and $q_{\max} = C \cdot p_1 \cdot p_2/2$. Therefore, $m \in \omega(m^\star)$ implies $m \in \omega(q_{\max}^{-1})$ in this case. Thus, for any $\varepsilon_m > 0$ it holds that $m \geq \varepsilon_m/q_{\max}$ for all sufficiently large $n$ and Corollary 4.16 gives us a probability of at least $\varepsilon_P$ as desired. Note that, depending on the lemma or corollary we used, we made different choices for $\varepsilon_m$. However, all these choices are satisfied for all sufficiently large $n$, since $m$ always grows asymptotically faster than the respective threshold function in all three cases. From this point, either Lemma 4.12, or Corollary 4.10, or Corollary 4.16 guarantees that the probability to generate an unsatisfiable instance is at least $\varepsilon_P$.

It remains to show that the threshold is not sharp in the last case. If none of the first three cases hold, then either

1. $p_1^2 \notin \Theta(\sum_{i=1}^n p_i^2)$ and $p_1^2 \notin o(\sum_{i=1}^n p_i^2)$ or
2. $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$, but $p_2^2 \notin \Theta(\sum_{i=2}^n p_i^2)$ and $p_2^2 \notin o(\sum_{i=2}^n p_i^2)$.

If $p_1^2 \notin o(\sum_{i=1}^n p_i^2)$, then there is a constant $\varepsilon_1$ so that for any $n_0 \in \mathbb{N}$ there must be an $n \geq n_0$ such that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$. We now consider only the values of $n$, where $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^n p_i^2$ holds. Essentially, we treat this as an ensemble with $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$. Now we either have $p_2^2 \in o(\sum_{i=2}^n p_i^2)$, or $p_2^2 \in \Theta(\sum_{i=2}^n p_i^2)$, or neither of the two. For $p_2^2 \in o(\sum_{i=2}^n p_i^2)$ we know from case 2 that there is a range of $m$ of size $\Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$, where the probability function approaches

neither zero nor one. For $p_2^2 \in \Theta(\sum_{i=2}^n p_i^2)$ the same holds for $\Theta(q_{\max}^{-1})$ due to case 3. From proving that $m^\star$ is an asymptotic threshold function we know that, depending on $p_1^2$ and $p_2^2$, $m^\star \in \Theta((C \cdot p_1 \cdot (\sum_{i=2}^n p_i^2)^{1/2})^{-1})$ or $m^\star \in \Theta(q_{\max}^{-1})$, respectively. Thus for all sufficiently large values of $n$ we consider, there are ranges of $m \in \Theta(m^\star)$, where the probability function approaches neither zero nor one. If we considered all values of $n$ now, the ones we selected before prevent our probability function from approaching zero or one in the chosen ranges. Thus, the threshold cannot be sharp. If $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$, or if we have infinitely many values of $n$ where this holds as for $p_1^2 \notin o(\sum_{i=1}^n p_i^2)$, and $p_2^2 \notin o(\sum_{i=2}^n p_i^2)$ a similar argumentation holds for the chosen values of $n$ with $p_2^2 \geq \varepsilon_2 \cdot \sum_{i=2}^n p_i^2$, i. e. we can assume $p_1^2 \in \Theta(\sum_{i=1}^n p_i^2)$ and $p_2^2 \in \Theta(\sum_{i=2}^n p_i^2)$ for those values. ■

## 4.6 Examples

We now apply Theorem 4.17 to determine the satisfiability threshold behavior of non-uniform random 2-SAT with different ensembles of probability distributions.

### 4.6.1 Random 2-SAT

For random 2-SAT the probability distribution for $n \in \mathbb{N}$ is $\vec{p}^{(n)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$. This means $p_1^2 = \frac{1}{n^2}$ and $\sum_{i=1}^n p_i^2 = \frac{1}{n}$. We see that $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. The first case of our theorem now tells us that there is a sharp threshold at $m^\star = (\sum_{i=1}^n p_i^2)^{-1} = n$. This is exactly what Chvátal and Reed [CR92] found out as well.

### 4.6.2 Power-law Random 2-SAT

Theorem 4.17 implies the following corollary.

▶ **Corollary 4.18.** For power-law random 2-SAT, if

- $\beta < 3$, then the threshold is coarse at $m^\star \in \Theta\left(q_{\max}^{-1}\right) \in \Theta\left(n^{2(\beta-2)/(\beta-1)}\right)$.
- $\beta = 3$, then the threshold is sharp at $m^\star = 4 \cdot \frac{n}{\ln n}$.
- $\beta > 3$, then the threshold is sharp at $m^\star = \frac{(\beta-1)\cdot(\beta-3)}{(\beta-2)^2} \cdot n$.

◀

*Proof.* For power-law random 2-SAT we assume some fixed $\beta > 2$. Then for $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{(n/i)^{\frac{1}{\beta-1}}}{\sum_{j=1}^n (n/j)^{\frac{1}{\beta-1}}}.$$

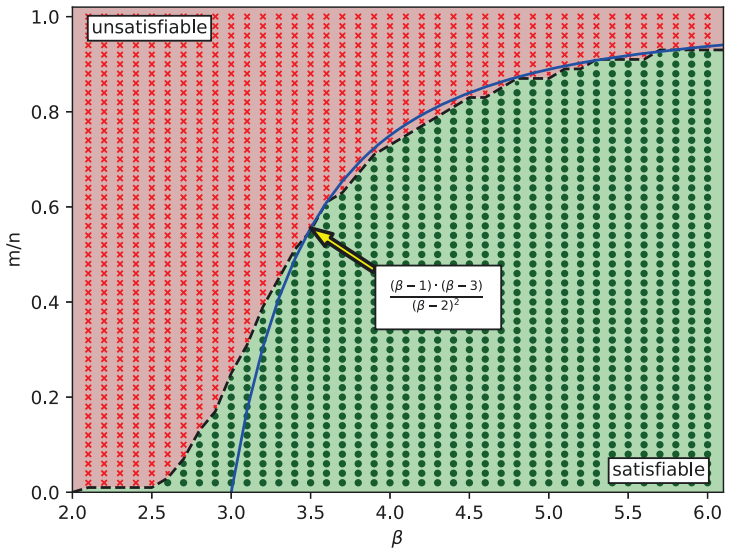



**Figure 4.4:** Phase diagram for power-law random 2-SAT formulas with $n = 10^7$ variables. Each point is a sample of 100 random instances at the given parameter combination. We drew a red cross if all instances were unsatisfiable and a green dot if at least one instance was satisfiable with the size of the dot scaling with the fraction of satisfiable instances. We empirically observe a sharp phase transition (– –), which closely matches the theoretical bound of Theorem 4.17 (——).

It holds that $p_1 \geq p_2 \geq \ldots \geq p_n$. Lemma 3.12 tells us that

$$\begin{aligned} p_1^2 &= (1 \pm o(1)) \cdot \left(\frac{\beta-2}{\beta-1}\right)^2 \cdot n^{-2\frac{\beta-2}{\beta-1}}, \\ p_2^2 &= (1 \pm o(1)) \cdot \left(\frac{\beta-2}{\beta-1}\right)^2 \cdot 2^{-\frac{1}{\beta-1}} \cdot n^{-2\frac{\beta-2}{\beta-1}}, \text{ and} \\ \sum_{i=1}^{n} p_i^2 &= \begin{cases} \Theta\left(n^{-2\frac{\beta-2}{\beta-1}}\right) & \text{for } \beta < 3 \\ (1 \pm o(1)) \cdot \frac{1}{4} \cdot \frac{\ln n}{n} & \text{for } \beta = 3 \\ (1 \pm o(1)) \cdot \frac{(\beta-2)^2}{(\beta-3)\cdot(\beta-1)} \cdot n^{-1} & \text{for } \beta > 3. \end{cases} \end{aligned}$$

For $\beta < 3$ it holds that $p_1^2 \in \Theta\left(\sum_{i=1}^n p_i^2\right)$ and $p_2^2 \in \Theta(\sum_{i=2}^n p_i^2)$. Thus, there is a coarse threshold at $m^\star = q_{\max}^{-1} \in \Theta(n^{2(\beta-2)/(\beta-1)})$, since $q_{\max} = C \cdot p_1 \cdot p_2/2$, $p_1, p_2 \in \Theta\left(n^{-(\beta-2)/(\beta-1)}\right)$, and $C = 1/(1 - \sum_{i=1}^n p_i^2) = 1 + o(1)$.

For $\beta = 3$ it holds that $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. Thus, there is a sharp satisfiability threshold at $m^\star = 4 \cdot \frac{n}{\ln n}$.

For $\beta > 3$ it also holds that $p_1^2 \in o(\sum_{i=1}^n p_i^2)$. Thus, there is a sharp satisfiability threshold at $m^\star = \frac{(\beta-1)\cdot(\beta-3)}{(\beta-2)^2} \cdot n$. ■

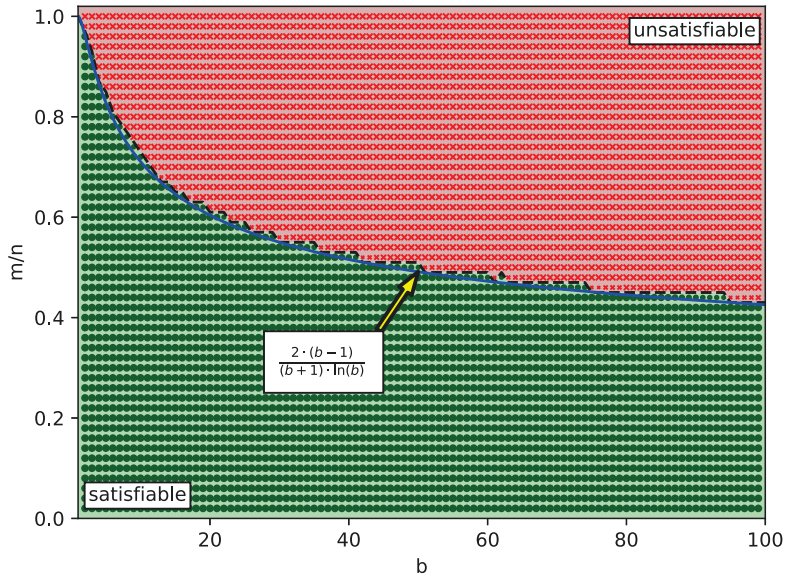



**Figure 4.5:** Phase diagram for geometric random 2-SAT formulas with $n = 10^6$ variables. Each point is a sample of 100 random instances at the given parameter combination. We drew a red cross if all instances were unsatisfiable and a green dot if at least one instance was satisfiable with the size of the dot scaling with the fraction of satisfiable instances. We empirically observe a sharp phase transition (– –), which closely matches the theoretical bound of Theorem 4.17 (—).

Figure 4.4 visualizes the empirical threshold position compared to the theoretical position according to Corollary 4.18.

### 4.6.3 Geometric Random 2-SAT

Theorem 4.17 implies the following corollary.

▶ **Corollary 4.19.** For geometric random 2-SAT with base $b > 1$ there is a sharp threshold at $m^\star = \frac{2\cdot(b-1)}{(b+1)\cdot\ln b} \cdot n$. ◀

*Proof.* We assume some fixed $b > 1$. Then for $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{b \cdot (1 - b^{-1/n})}{b - 1} \cdot b^{-(i-1)/n}.$$

Again, it holds that $p_1 \geq p_2 \geq \ldots \geq p_n$.

Lemma 3.13 tells us

$$p_1^2 = (1 - o(1)) \cdot \left(\frac{b \cdot \ln b}{b-1}\right)^2 \cdot n^{-2}.$$

and

$$\sum_{i=1}^{n} p_i^2 = (1 \pm o(1)) \cdot \frac{b+1}{b-1} \cdot \frac{\ln b}{2} \cdot n^{-1}.$$

Since $p_1^2 \in o(\sum_{i=1}^{n} p_i^2)$, the threshold is sharp at $m^\star = \frac{2\cdot(b-1)}{(b+1)\cdot \ln b} \cdot n$. ■

Again, Figure 4.5 visualizes the empirical threshold position compared to the theoretical one from Corollary 4.19.

# 5 Asymptotic Threshold in Non-Uniform Random k-SAT

*The content of this chapter is based on the publication [Fri+17a], which is joint work with Tobias Friedrich, Anton Krohmer, Thomas Sauerwald, and Andrew M. Sutton. The results from that paper have been generalized to encompass non-uniform random k-SAT with arbitrary ensembles of probability distributions instead of only power-law random k-SAT.*

In this chapter we analyze the behavior of the satisfiability threshold in non-uniform random $k$-SAT for $k \geq 3$ with regard to the number of clauses $m$. In the last chapter we did the same for $k = 2$. However, instances with $k \geq 3$ are more difficult to analyze, since the structures that result in unsatisfiability are not as simple as snakes and bicycles anymore. Thus, it is harder to certify satisfiability or unsatisfiability by showing the existence or absence of such structures.

To show that formulas are unsatisfiable above the threshold, we use the first moment method to bound the number of satisfying assignments. However, the bound we get from this approach is rather high. Hence, we show a better bound using the Single Flip Method [Kir+98]. The method improves upon the first moment bound by considering only a subset of satisfying assignments. Alternatively, Corollary 4.16 from the last chapter gives a bound depending on the maximum clause probability.

To show that formulas are satisfiable below the threshold, we reduce $k$-SAT instances to 2-SAT instances. Given a Boolean formula in $k$-CNF one can pick two literals from each clause to get a formula in 2-CNF. If this 2-SAT formula is satisfiable, so is the original formula. We will prove that, if the two literals from each clause are picked in a suitable way, we almost surely get a satisfiable 2-SAT formula as soon as the number of clauses is small enough. In order to prove this, we will use our results on non-uniform random 2-SAT from Chapter 4.

Before we can show these results, let us repeat some basics of non-uniform random $k$-SAT that will be crucial in this chapter. The probability to draw a clause $c = (\ell_1 \vee \ell_2 \vee \ldots \vee \ell_k)$ is

$$q_c = C \cdot \frac{k!}{2^k} \prod_{\ell \in c} p(|\ell|). \tag{5.1}$$

with

$$C = \left( k! \cdot \sum_{J \in \mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j \in J} p_j \right)^{-1}.$$

We want to estimate $C$ to make the factor more manageable. Note that $C$ is a normalization factor which describes the probability that all $k$ Boolean

variables drawn in a clause are different. Thus, to get an upper bound on $C = 1/\Pr[\text{all } k \text{ variables different}]$, it suffices to have an upper bound on the probability that we draw a variable twice. The following lemma gives us exactly that bound.

▶ **Lemma 5.1 (Non-Uniform Birthday Paradox [ASV15]).** Let $\vec{p} = (p_1, \ldots, p_n)$ be any probability distribution on $n$ items. Assume we sample $t$ items from $\vec{p}$. Let $\mathcal{E}(t)$ be the event that there is a collision, i. e. that at least 2 of $t$ items are equal. Then,

$$\Pr[\mathcal{E}(t)] \leq \frac{t \cdot (t-1)}{2} \sum_{i=1}^{n} p_i^2.$$

◀

The lemma directly yields

$$C \leq \left(1 - \frac{k \cdot (k-1)}{2} \cdot \sum_{i=1}^{n} p_i^2\right)^{-1}. \tag{5.2}$$

Since we will consider $C$ for different values of $k$, we also denote it as $C_k$.

Remember that in order for a function $m^\star \colon \mathbb{N} \to \mathbb{R}$ to be an asymptotic threshold function for satisfiability it has to hold that

$$\lim_{n\to\infty} \Pr_{\Phi \sim \mathcal{D}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},m)} [\,\Phi \text{ satisfiable}\,] = \begin{cases} 1, & \text{if } m \in o_n(m^\star) \\ 0, & \text{if } m \in \omega_n(m^\star). \end{cases}$$

We will consider $m \in \omega_n(m^\star)$ in Section 5.1 and $m \in o_n(m^\star)$ in Section 5.2.

## 5.1 Unsatisfiability

It is a well-known result [CR92] that random k-SAT on any probability distribution will result in unsatisfiable formulas if the clause-variable ratio is high. This follows from the first moment method: The expected number of assignments that satisfy a formula is $2^n(1-2^{-k})^m$. This is independent of the variable distribution as long as each variable is negated with probability $1/2$. Hence, if the clause-variable ratio exceeds $\ln(2)/\ln(2^k/(2^k-1))$, the resulting formula will be unsatisfiable with high probability. This constant is rather large, however: In the case of $k = 3$ this yields an upper bound on the clause-variable ratio of $\approx 5.191$. Nevertheless, it certifies that the satisfiability threshold of non-uniform random $k$-SAT can be at most $m^\star \in O(n)$, independently of the ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$.

Obviously, there are ensembles of probability distributions which yield a much smaller threshold function as we have seen in the case of power-law random 2-SAT with exponent $\beta \leq 3$ (c. f. Section 4.6.2). One reason for formulas to be unsatisfiable could be that the $k$ variables with highest probabilities appear

together in clauses too often, thus appearing with all $2^k$ possible combinations of signs and making the formula unsatisfiable. Corollary 4.16 from the last chapter tells us that non-uniform random $k$-SAT instances with $m \in \omega(1/q_{\max})$ clauses are a. a. s. unsatisfiable, where $q_{\max}$ is the maximum clause probability. In some cases this gives a better bound than the first moment method.

Alternatively, we can try to improve the bound from the first moment method. The Single Flip Method was introduced by Kirousis et al. [Kir+98]. It improves the first moment bound by only considering a subset of satisfying assignments. Since the number of those is smaller, so is the resulting probability bound. Thus, fewer clauses are needed to make the probability approach zero.

The satisfying assignments the method considers have the following property.

▶ **Definition 5.2 (Single-Flip Property [Kir+98]).** For a formula $\Phi$ a truth assignment $\alpha$ has the *single-flip property* iff $\alpha$ satisfies $\Phi$ and every assignment $\alpha'$ obtained from $\alpha$ by flipping exactly one zero to one does *not* satisfy $\Phi$. ◀

If $\Phi$ is satisfiable, then such an assignment exists due to [Kir+98]. This is intuitively clear, if we consider any satisfying assignment $\alpha$. Either it has the property or we can flip a zero to one to get a different satisfying assignment that we can consider instead. We can repeat this until we either find an assignment with the property or reach the assignment $\alpha = 1^n$. If we reach $1^n$, it must be satisfying and thus has the single-flip property by definition.

Note that in this section we annotate our probabilities with $c$ and $\Phi$ respectively to differentiate between drawing a single random clause $c$ and drawing a random formula $\Phi$ consisting of $m$ randomly drawn clauses. We let the random variable $N_{SF}$ denote the number of assignments with the single-flip property. Markov's inequality now tells us that $\Pr_\Phi[\,\Phi \text{ satisfiable}\,] \leq \mathbb{E}_\Phi[\,N_{SF}\,]$. In the following, we derive a bound on $\mathbb{E}_\Phi[\,N_{SF}\,]$. To bound the number of assignments with the single-flip property, we use a result by Kirousis et al. [Kir+98].

▶ **Lemma 5.3 ([Kir+98]).** The expected number of assignments with the single-flip property is

$$\mathop{\mathbb{E}}_{\Phi}[\,N_{SF}\,] = \left(1 - \tfrac{1}{2^k}\right)^m \sum_{\text{assignment } \alpha} \Pr_\Phi[\,\alpha \text{ single-flip} \mid \alpha \text{ satisfying}\,].$$

◀

Now we bound the probability that a satisfying assignment $\alpha$ has the single-flip property.

▶ **Lemma 5.4.** For an assignment $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \{0,1\}^n$ it holds that

$$\Pr_\Phi[\,\alpha \text{ single-flip} \mid \alpha \text{ satisfying}\,] \leq \prod_{i\colon \alpha_i = 0} \left(1 - \left(1 - C_k \cdot \frac{k \cdot p_i}{2^k - 1}\right)^m\right).$$

◀

*Proof.* For a satisfying assignment $\alpha$ to have the single-flip property, all assignments $\alpha^{(i)}$ obtained by flipping a bit $\alpha_i = 0$ of $\alpha$ must not satisfy $\Phi$. To fulfill this property for $\alpha^{(i)}$, we have to choose at least one clause which contains $\overline{X_i}$ and $k-1$ other variables with appropriate signs so that $\alpha^{(i)}$ does not satisfy the clause. Let $S^{(i)}$ denote the event that we draw a clause $c$ that is satisfied by $\alpha$, but not by $\alpha^{(i)}$. Then, it holds that

$$\Pr_c\left[\, S^{(i)} \,\right] = C_k \cdot \frac{k!}{2^k} \cdot p_i \cdot \sum_{J \in \mathcal{P}_{k-1}([n]\setminus\{i\})} \prod_{j \in J} p_j \leq C_k \cdot \frac{k \cdot p_i}{2^k},$$

since $\sum_{J \in \mathcal{P}_{k-1}([n]\setminus\{i\})} \prod_{j \in J} p_j \leq \frac{1}{(k-1)!}$ according to Lemma 4.1. The probability of choosing a clause not satisfied by $\alpha^{(i)}$ under the condition that we draw a clause that $\alpha$ satisfies is then

$$\Pr_c\left[\, S^{(i)} \mid \alpha \text{ sat} \,\right] \leq C_k \cdot \frac{k \cdot p_i}{2^k - 1}$$

as the probability of choosing a clause which is satisfied by any fixed assignment is exactly $(2^k - 1)/2^k$. For a fixed assignment $\alpha^{(i)}$ we conclude

$$\begin{aligned} \Pr_\Phi\left[\, \alpha^{(i)} \text{ unsat} \mid \alpha \text{ sat} \,\right] &= 1 - \left(1 - \Pr_c\left[\, S^{(i)} \mid \alpha \text{ satisfies } c \,\right]\right)^m \\ &\leq 1 - \left(1 - C_k \cdot \frac{k \cdot p_i}{2^k - 1}\right)^m. \end{aligned} \tag{5.3}$$

It remains to find the joint probability that all single-flipped assignments $\alpha^{(i)}$ for $1 \leq i \leq n$ with $\alpha_i = 0$ are not satisfying. We show this using a correlation inequality by Farr [McD92]. The sets of clauses which are not satisfied by the $\alpha^{(i)}$'s are pairwise disjoint as each clause in the set for $\alpha^{(i)}$ has to contain $\overline{X_i}$, whereas each clause in the set for $\alpha^{(j)}$ $(j \neq i)$ can not contain $\overline{X_i}$, since $\alpha^{(j)}$ differs from $\alpha$ only in $X_j$ and thus satisfies ${\alpha_i}^{(j)} = 0$ as does $\alpha$. In the context of the correlation inequality from [McD92] we set $V = \{1, 2, \ldots, m\}$, $I = \{i \in \{1, 2, \ldots, n\} \mid \alpha_i = 0\}$, $X_v = i$ iff the $v$-th clause is satisfied by $\alpha$, but not by $\alpha^{(i)}$, and $\mathcal{F}_i$ the "increasing" collection of non-empty subsets of $V$. The application of the Theorem then directly yields

$$\begin{aligned} \Pr_\Phi[\, \alpha \text{ single-flip} \mid \alpha \text{ sat} \,] &= \Pr_\Phi\left[\, \bigwedge_{i\colon \alpha_i = 0} \alpha^{(i)} \text{ unsat} \mid \alpha \text{ sat} \,\right] \\ &\leq \prod_{i\colon \alpha_i = 0} \Pr_\Phi\left[\, \alpha^{(i)} \text{ unsat} \mid \alpha \text{ sat} \,\right]. \\ &\leq \prod_{i\colon \alpha_i = 0} \left[1 - \left(1 - C_k \cdot \frac{k \cdot p_i}{2^k - 1}\right)^m\right]. \end{aligned}$$

■

Combining Lemma 5.3 and Lemma 5.4 we get the following result.

▶ **Corollary 5.5.** Let $\Phi \sim \mathcal{D}^N\left(n, k, \left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}}, m\right)$ be a non-uniform random $k$-SAT formula. The expected number of assignments with single-flip property is at most

$$\mathop{\mathbb{E}}_{\Phi\sim\mathcal{D}^N}[\, N_{SF}\,] \leq \left(1-\tfrac{1}{2^k}\right)^m \prod_{i=1}^{n}\left[2-\left(1-C_k\cdot\frac{k\cdot p_i}{2^k-1}\right)^m\right].$$

◀

*Proof.* Plugging Lemma 5.4 into Lemma 5.3 we get

$$\begin{aligned}\mathop{\mathbb{E}}_{\Phi\sim\mathcal{D}^N}[\, N_{SF}\,] &\leq \left(1-\tfrac{1}{2^k}\right)^m \sum_{I\subseteq\{1,2,\ldots,n\}}\prod_{i\in I}\left[1-\left(1-C_k\cdot\frac{k\cdot p_i}{2^k-1}\right)^m\right]\\ &= \left(1-\tfrac{1}{2^k}\right)^m \prod_{i=1}^{n}\left[2-\left(1-C_k\cdot\frac{k\cdot p_i}{2^k-1}\right)^m\right].\end{aligned}$$

■

If we can guarantee this expected value to be $o(1)$, the same holds for the probability to generate satisfiable instances, since

$$\Pr_{\Phi}[\,\Phi \text{ satisfiable}\,] = \Pr_{\Phi}[\,\Phi \text{ has a single-flip assignment}\,] \leq \mathop{\mathbb{E}}_{\Phi}[\, N_{SF}\,]$$

as we explained at the beginning of this section. We can now use Corollary 5.5 to derive upper bounds on the satisfiability threshold that improve on the bound

$$m^\star \leq \frac{\ln(2)}{\ln(2^k/(2^k-1))}\cdot n \tag{5.4}$$

derived by Chvátal and Reed [CR92]. Our new bounds are tailored to the non-uniform input distributions we consider. However, it is difficult to derive closed expressions for our example distributions. Furthermore, at least for the example distributions we consider, Corollary 5.5 does not yield asymptotically smaller upper bounds than the ones we get from Corollary 4.16 or equation (5.4).

Nevertheless, our result can prove useful in bounding the leading constant of a sharp satisfiability threshold. Some first steps into that direction have been made in the paper [Fri+17a] this chapter is based on. There, we showed upper bounds on the satisfiability threshold for power-law random $k$-SAT with exponents $\beta > \frac{2k-1}{k-1}$. Those bounds were are even smaller than known lower bounds on the threshold for random $k$-SAT. This proved that for certain exponents $\beta$ the satisfiability threshold of power-law random $k$-SAT is smaller than the one for random $k$-SAT. However, in this thesis we only study the asymptotic threshold position and sharpness of the threshold for non-uniform random $k$-SAT with $k \geq 3$. We leave deriving leading constants of sharp thresholds to future work.

## 5.2 Satisfiability

In order to show satisfiability, we reduce an instance of non-uniform random $k$-SAT to a Boolean formula in 2-CNF by picking two literals from each clause. Any satisfying assignment for the resulting formula is also a satisfying assignment for the original formula. If we pick two literals from each clause uniformly at random, we will get an instance of non-uniform random 2-SAT with exactly the same ensemble of probability distributions $(\vec{p}^{(n)})_{n\in\mathbb{N}}$. Thus, the following corollary holds.

▶ **Corollary 5.6.** Let $\Phi \sim \mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ be a non-uniform random $k$-SAT formula. Then, the probability that $\Phi$ is satisfiable is at least as high as the probability that $\Phi' \sim \mathcal{D}^N(n, 2, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ is satisfiable. Especially, $\Phi$ is satisfiable a. a. s. if $m \in o(m^\star)$, where

$$m^\star = \frac{1 - \sum_{i=1}^n p_i^2}{\sum_{i=2}^n p_i^2 + p_1 \cdot \left(\sum_{i=2}^n p_i^2\right)^{1/2}}$$

is the asymptotic threshold function of $\mathcal{D}^N\left(n, 2, \left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}}, m\right)$ with respect to $m$. ◀

However, we can get a better bound by picking from each clause the two literals with smallest probability. This results in the following probability distribution on 2-clauses. For $i, j \in [n]$ and $\ell_1, \ell_2$ literals over $X_i$ and $X_j$ it holds that

$$\Pr[c = (\ell_1 \vee \ell_2)] \leq C_k \frac{k \cdot (k-1)}{4} \cdot p_i \cdot F(i-1)^{(k-2)/2} \cdot p_j \cdot F(j-1)^{(k-2)/2} \approx p_i' \cdot p_j',$$

where $F(i) = \sum_{l=1}^i p_l$. We can now plug these upper bounds on the clause and variable probabilities into a relaxed version of Lemma 4.4. The relaxation is in the sense that now we are only concerned with asymptotic threshold functions and that we only need upper bounds on clause probabilities. The lemma derives an upper bound on the expected number of bicycles in a Boolean formula in 2-CNF. If we choose a number of clauses $m$ so that this expected number is $o(1)$, there is a. a. s. no bicycle in the formula. Thus, it is a. a. s. satisfiable according to Chvátal and Reed [CR92] (see also section 4.2). As mentioned before, this approach also works if only upper bounds for the clause and variable probabilities are known.

Why can the bounds we get from picking the two least-probable literals be better than the ones we get from picking literals at random? This might be due to the non-uniformity of the variable probability distributions. We conjecture that for non-uniform random $k$-SAT the uniform distribution (random $k$-SAT) constitutes an extreme case in the sense that the satisfiability threshold is as high as it could possibly be. By picking the two least-probable literals from each clause we make the probability distribution $\vec{p}'$ more uniform, since $p_i' \in \Theta(p_i \cdot F(i-1)^{k/2-1})$ and $p_i$ decreases, while $F(i)$ increases with increasing $i$. Due to our conjecture this more uniform probability distribution results in a higher

bound on the satisfiability threshold. While picking literals at random is easier and yields the same asymptotic bound in some cases, we will also see examples where picking the two least-probable literals yields a much better bound (c. f. Section 5.3.2).

We will now show the necessary results rigorously. The relaxed lemma states the following.

▶ **Lemma 5.7.** Let $\mathcal{M}$ be a random 2-SAT model over $n$ variables and with $m$ clauses drawn independently at random. If the Boolean variables $X_i$ can be assigned functions $p_i \colon \mathbb{N} \to \mathbb{R}^+$ such that the probability to draw each clause $c$ is at most

$$\Pr[c = (\ell_1 \vee \ell_2)] \leq \alpha \cdot p(|\ell_1|) \cdot p(|\ell_2|),$$

then a random formula $\Phi \sim \mathcal{M}$ is a. a. s. satisfiable if $m \in o(m^\star)$, where

$$m^\star = \left(2 \cdot \alpha \cdot \sum_{i=2}^{n} p_i^2 + 2 \cdot \alpha \cdot p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}\right)^{-1}.$$

◀

*Proof.* The probability that a *specific* bicycle $B$ of size $t$ appears in $\Phi$ is

$$\Pr[B \text{ in } \Phi] = \underbrace{\binom{m}{t+1} \cdot (t+1)!}_{\text{positions of } B \text{ in } \Phi} \cdot \Pr[(u \vee w_1)] \cdot \Pr[(\overline{w_t} \vee v)] \cdot \prod_{h=1}^{t-1} \Pr[(\overline{w_h} \vee w_{h+1})].$$

Thus, for a set $S \in \mathcal{P}_t([n])$ of variables the probability that *any* bicycle with the variables from $S$ appears in $\Phi$ is at most

$$\Pr[S\text{-bicycle in } \Phi] \leq m^{t+1} \cdot t! \cdot 2^t \cdot \alpha^{t+1} \cdot \prod_{i \in S} p_i^2 \cdot \left(2 \cdot \sum_{i \in S} p_i\right)^2,$$

where the last factor accounts for the possibilities to choose $u$ and $v$. Here, we used the requirement $\Pr[c = (\ell_1 \vee \ell_2)] \leq \alpha \cdot p(|\ell_1|) \cdot p(|\ell_2|)$. It now holds that

$$\Pr[\Phi \text{ contains a bicycle}] \leq \sum_{t=2}^{n} \sum_{S \in \mathcal{P}_t([n])} \left(m^{t+1} \cdot t! \cdot 2^t \cdot \alpha^{t+1} 2^2 \cdot \prod_{i \in S} p_i^2 \left(\sum_{i \in S} p_i\right)^2\right).$$

Depending on the relation of $p_1^2$ to $\sum_{i=1}^n p_i^2$, we get different bounds. It holds that

$$\Pr[\Phi \text{ contains a bicycle}] \leq 2 \cdot \frac{p_1^2}{\sum_{i=1}^n p_i^2} \cdot \sum_{t=2}^{n} \left(\left(2 \cdot \alpha \cdot m \cdot \sum_{i=1}^{n} p_i^2\right)^{t+1} \cdot t^2\right)$$

$$\leq 4 \cdot \alpha \cdot m \cdot \left(\sum_{i=1}^{n} p_i^2\right) \cdot \sum_{t=2}^{n} \left(\left(2 \cdot \alpha \cdot m \cdot \sum_{i=1}^{n} p_i^2\right)^t \cdot t^2\right).$$

Thus, for $m \in o((2 \cdot \alpha \cdot \sum_{i=1}^{n} p_i^2)^{-1})$ this is at most $o(1)$. However, if $p_1^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$ for some $\varepsilon_1 \in (0,1)$, it holds that

$$\sum_{i=2}^{n} p_i^2 = \sum_{i=1}^{n} p_i^2 - p_1^2 \geq (1-\varepsilon_1) \cdot \sum_{i=1}^{n} p_i^2.$$

In that case $m \in o((2 \cdot \alpha \cdot \sum_{i=2}^{n} p_i^2)^{-1})$ is sufficient. However, if we can choose an $\varepsilon_1 \in (0,1)$ so that $p_1^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i^2$, then we can achieve a better bound as we have seen in Lemma 4.4. It then holds that

$$\Pr[\Phi \text{ unsat}] \leq 2 \cdot (1 + 1/\varepsilon_1) \sum_{t=2}^{n} \left(\left(2 \cdot \alpha \cdot m \cdot \frac{1}{\sqrt{\varepsilon_1}} \cdot p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}\right)^{t+1} \cdot t^3\right).$$

This is $o(1)$ if $m \in o((2 \cdot \alpha \cdot p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2})^{-1})$.

As in Section 4.5 we can use these two cases to derive the result as desired. First we choose some fixed $\varepsilon_1 \in (0,1)$. For all $n \in \mathbb{N}$ it either holds that $p_1(n)^2 \leq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i(n)^2$ or $p_1(n)^2 \geq \varepsilon_1 \cdot \sum_{i=1}^{n} p_i(n)^2$. In the former case $m \in o(m^\star)$ implies $m \in o((2 \cdot \alpha \cdot \sum_{i=2}^{n} p_i^2)^{-1})$. In the latter case $m \in o(m^\star)$ implies $m \in o((2 \cdot \alpha \cdot p_1 \cdot (\sum_{i=2}^{n} p_i^2)^{1/2})^{-1})$. Thus, either way we have a probability of $1 - o(1)$ that a random instance is satisfiable. ■

It remains to bound the probability to sample 2-clauses from non-uniform random $k$-SAT by picking the two least-probable literals.

▶ **Lemma 5.8.** Let $c$ be a clause drawn with non-uniform random $k$-SAT with probability ensemble $(\vec{p}^{(n)})_{n \in \mathbb{N}}$. Let $c'$ be the 2-clause consisting of the two literals with the smallest variable probability from $c$. For $i, j \in [n]$ and $\ell_1, \ell_2$ literals over $X_i$ and $X_j$ it holds that

$$\Pr[c = (\ell_1 \vee \ell_2)] \leq C_k \cdot \frac{k \cdot (k-1)}{4} \cdot p_i \cdot F(i-1)^{k/2-1} \cdot p_j \cdot F(j-1)^{k/2-1},$$

where $F(i) = \sum_{l=1}^{i} p_l$. Furthermore $\Pr\left[c = (\ell_i \vee \ell_j)\right] = 0$ if $i \leq k-2$ or $j \leq k-2$. ◀

*Proof.* We create a random clause with non-uniform random $k$-SAT with probability ensemble $(\vec{p}^{(n)})_{n \in \mathbb{N}}$ and pick the two variables with smallest probability. It holds that $p_1 \geq p_2 \geq \ldots \geq p_n$. Let $i, j \in [n]$ with $i < j$ and $\ell_1, \ell_2$ literals over $X_i$ and $X_j$. For $X_i$ and $X_j$ to be the two variables with smallest probabilities, the $k-2$ other variables have to have an index of at most $i$. Thus, the probability is

zero if one of the two indices is at most $k-2$. Furthermore, it holds that

$$\begin{aligned}\Pr[c = (\ell_1 \vee \ell_2)] &\leq C_k \cdot \frac{k!}{2^k} \cdot p_i \cdot p_j \cdot 2^{k-2} \sum_{S \in \mathcal{P}_{k-2}([i-1])} \prod_{s \in S} p_s \\ &\leq C_k \cdot \frac{k!}{4} \cdot p_i \cdot p_j \cdot \frac{1}{(k-2)!} \left( \sum_{l=1}^{i-1} p_l \right)^{k-2}\end{aligned}$$

by Lemma 4.1. Since $F(i-1) = \sum_{l=1}^{i-1} p_l \leq \sum_{l=1}^{j-1} p_l = F(j-1)$, we get

$$\Pr[c = (\ell_1 \vee \ell_2)] \leq C_k \cdot \frac{k \cdot (k-1)}{4} \cdot p_i \cdot p_j \cdot F(i-1)^{k/2-1} \cdot F(j-1)^{k/2-1}.$$

■

Lemma 5.7 and Lemma 5.8 imply the following theorem.

▶ **Theorem 5.9.** Let $\Phi \sim \mathcal{D}^N\left(n, k, \left(\vec{p}^{(n)}\right)_{n \in \mathbb{N}}, m\right)$ be a non-uniform random $k$-SAT formula. Let

$$m^\star = \left( \frac{C_k \cdot k \cdot (k-1)}{2} \cdot \sum_{i=2}^{n-k+2} p_i'^2 + \frac{C_k \cdot k \cdot (k-1)}{2} \cdot p_1' \cdot \left( \sum_{i=2}^{n-k+2} p_i'^2 \right)^{1/2} \right)^{-1},$$

where $p_i'(n)$ is the $i$-th largest value in the set

$$\left\{ p_j(n) \cdot \left( \sum_{l=1}^{j-1} p_l(n) \right)^{k/2-1} \,\middle|\, j \in \{k-1, k, \ldots, n\} \right\}$$

and $p_i' \colon \mathbb{N} \to \mathbb{R}^+$ is the function of those values depending on $n$. Then, $\Phi$ is a. a. s. satisfiable if $m \in o(m^\star)$. ◀

## 5.3 Examples

We can use our results to determine the asymptotic threshold functions of non-uniform random $k$-SAT with given ensembles of probability distributions. As before, we consider three such ensembles and their corresponding models as examples: random $k$-SAT, power-law random $k$-SAT, and geometric random $k$-SAT. Since we need to know the asymptotic threshold function to derive sharpness of the satisfiability threshold, these results will be necessary when considering the same examples in the next chapter.

### 5.3.1 Random k-SAT

For random $k$-SAT the probability ensemble is

$$\forall n \in \mathbb{N}\colon \vec{p}^{(n)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

It holds that

$$C_k \leq \frac{1}{1 - \frac{k^2}{2} \cdot \sum_{i=1}^{n} p_i^2} = \frac{1}{1 - \frac{k^2}{2 \cdot n}} = 1 + o(1).$$

Thus, $q_{\max} \in \Theta(n^{-k})$ and Corollary 4.16 tells us that instances become a. a. s. unsatisfiable for $m \in \omega(n^k)$. However, the first moment method gives us a better bound of $m \in \omega(n)$ for instance to be a. a. s. unsatisfiable.

Now we reduce random $k$-SAT to random 2-SAT by picking two literals from each clause uniformly at random. We know that for random 2-SAT, instances are a. a. s. satisfiable for $m \in o(n)$. However, we can also use the more general bound of Corollary 5.6, which tells us that instances are a. a. s. satisfiable if

$$m \in o\left(\frac{1 - \sum_{i=1}^{n} p_i^2}{\sum_{i=2}^{n} p_i^2 + p_1 \cdot \left(\sum_{i=2}^{n} p_i^2\right)^{1/2}}\right) \in o(n),$$

since $p_1 = \frac{1}{n}$ and $\sum_{i=2}^{n} p_i^2 \in \Theta(\frac{1}{n})$. Thus, the asymptotic satisfiability threshold for random $k$-SAT is $m^\star \in \Theta(n)$ as Chvátal and Reed [CR92] already showed.

### 5.3.2 Power-Law Random k-SAT

We could use Corollary 5.6 and Corollary 4.18 to derive a bound on satisfiability. The corollary tells us that instances are a. a. s. satisfiable if $m \in o(n^{2\cdot(\beta-2)/(\beta-1)})$ for $\beta < 3$, $m \in o(n/\ln n)$ for $\beta = 3$, and $m \in o(n)$ for $\beta > 3$. However, Theorem 5.9 yields a better bound as we will see in the following corollary.

▶ **Corollary 5.10.** For power-law random $k$-SAT, if

- $\beta < \frac{2k-1}{k-1}$, then the threshold is coarse at $m^\star \in \Theta\left(q_{\max}^{-1}\right) \in \Theta\left(n^{k(\beta-2)/(\beta-1)}\right)$.
- $\beta = \frac{2k-1}{k-1}$, then formulas with $m^\star \in o\left(\frac{n}{\ln n}\right)$ are a. a. s. satisfiable and formulas with $m^\star \in \omega(n)$ are a. a. s. unsatisfiable.
- $\beta > \frac{2k-1}{k-1}$, then the asymptotic threshold is at $m^\star \in \Theta(n)$.

◀

*Proof.* For power-law random $k$-SAT we assume some fixed $\beta > 2$. Then, for $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{(n/i)^{\frac{1}{\beta-1}}}{\sum_{j=1}^{n} (n/j)^{\frac{1}{\beta-1}}}.$$

It holds that $p_1 \geq p_2 \geq \ldots \geq p_n$ and Lemma 3.12 tells us

$$p_1 = (1 \pm o(1)) \cdot \left(\frac{\beta - 2}{\beta - 1}\right) \cdot n^{-\frac{\beta-2}{\beta-1}},$$

and

$$\sum_{i=1}^{n} p_i^2 = \begin{cases} \Theta\left(n^{-2\frac{\beta-2}{\beta-1}}\right) & \text{for } \beta < 3 \\ (1 \pm o(1)) \cdot \frac{1}{4} \cdot \frac{\ln n}{n} & \text{for } \beta = 3 \\ (1 \pm o(1)) \cdot \frac{(\beta-2)^2}{(\beta-3)\cdot(\beta-1)} \cdot n^{-1} & \text{for } \beta > 3. \end{cases}$$

It holds that $q_{\max} \in \Theta(n^{-k\cdot(\beta-2)/(\beta-1)})$. Thus, instances are a. a. s. unsatisfiable for $m \in \omega(n^{k\cdot(\beta-2)/(\beta-1)})$ due to Corollary 4.16. For $\beta < \frac{2k-1}{k-1}$ this is smaller than the bound we get from the first moment method. For $\beta \geq \frac{2k-1}{k-1}$ the first moment method gives us a smaller bound of $m \in \omega(n)$.

Again due to Lemma 3.12 it holds that

$$F(i) = \sum_{j=1}^{n} p_j \leq (1 + o(1)) \cdot \left(\frac{i}{n}\right)^{\frac{\beta-2}{\beta-1}}.$$

Thus,

$$p_i \cdot F(i-1)^{k/2-1} \leq (1 + o(1)) \cdot \frac{1}{n} \cdot \frac{\beta - 2}{\beta - 1} \cdot \left(\frac{i}{n}\right)^{(k/2)\cdot\frac{\beta-2}{\beta-1}-1}.$$

We can substitute $(k/2) \cdot \frac{\beta-2}{\beta-1} - 1$ with $-1/(\beta' - 1)$, where $\beta' = 1 + 1/(1 - \frac{k}{2} \cdot \frac{\beta-2}{\beta-1})$. Thus, these probabilities are power-law distributed with exponent $\beta'$, i. e.

$$p_i \cdot F(i-1)^{k/2-1} \in \Theta\left(\frac{\left(\frac{n}{i}\right)^{\frac{1}{\beta'-1}}}{n}\right).$$

It holds that $\beta' > 2$ iff $\beta > 2$ and that $\beta' < 3$ iff $\beta < \frac{2k-1}{k-1}$. The functions $p'_i$ from Theorem 5.9 are now

$$p'_i \in \Theta\left(\frac{\left(\frac{n}{i+k-2}\right)^{\frac{1}{\beta'-1}}}{n}\right).$$

Thus,

$$p'_1 \in \Theta\left(\frac{\left(\frac{n}{k-1}\right)^{\frac{1}{\beta'-1}}}{n}\right) \in \Theta\left(n^{-\frac{\beta'-2}{\beta'-1}}\right)$$

and

$$\sum_{i=2}^{n-k+2} p_i'^2 \in \begin{cases} \Theta\left(n^{-2\frac{\beta'-2}{\beta'-1}}\right) & \text{for } \beta' < 3 \\ \Theta\left(\frac{\ln n}{n}\right) & \text{for } \beta' = 3 \\ \Theta\left(n^{-1}\right) & \text{for } \beta' > 3. \end{cases}$$

From this we get that instances are a. a. s. satisfiable if $m \in o(n^{2\cdot(\beta'-2)/(\beta'-1)})$ for $\beta' < 3$, $m \in o(n/\ln n)$ for $\beta' = 3$, and $m \in o(n)$ for $\beta' > 3$. If we substitute $\beta'$ again, this yields

$$m \in \begin{cases} o(n^{k\cdot(\beta-2)/(\beta-1)}) & \text{for } \beta < \frac{2k-1}{k-1}, \\ o(n/\ln n) & \text{for } \beta = \frac{2k-1}{k-1}, \\ o(n) & \text{for } \beta > \frac{2k-1}{k-1}. \end{cases}$$

Thus, for $\beta < \frac{2k-1}{k-1}$ the asymptotic threshold function is $m^\star \in \Theta\left(n^{k\cdot(\beta-2)/(\beta-1)}\right)$ and for $\beta > \frac{2k-1}{k-1}$ the asymptotic threshold function is $m^\star \in \Theta(n)$. For $\beta = \frac{2k-1}{k-1}$ we do not know if there is an asymptotic threshold function, but if it exists, it is somewhere between $n/\ln n$ and $n$. More discussion on this case can be found in Section 5.4.

If we look a bit closer, we can even derive that the threshold is coarse for $\beta < \frac{2k-1}{k-1}$. Lemma 5.7 yields a constant upper bound bounded away from one on the probability to generate unsatisfiable instances if $m = \varepsilon_m \cdot n^{k\cdot(\beta-2)/(\beta-1)}$ for small enough constants $\varepsilon_m > 0$. Analogously, Corollary 4.14 yields a constant lower bound bounded away from zero for $m = \varepsilon_m \cdot n^{k\cdot(\beta-2)/(\beta-1)}$ with any constant $\varepsilon_m > 0$ if $q_{\max} \in o(1)$. This is the case, since $q_{\max} \in \Theta(n^{-k\cdot(\beta-2)/(\beta-1)}) \in o(1)$. Both results together give us a range of constants $0 < \varepsilon_1 < \varepsilon_2$, where the probability is bounded away from zero and one in the limit. This implies a coarse threshold. ■

### 5.3.3 Geometric Random k-SAT

Corollary 5.6 implies the following corollary.

▶ **Corollary 5.11.** For geometric random k-SAT with some constant base $b > 1$, the asymptotic satisfiability threshold is at $m^\star \in \Theta(n)$. ◀

*Proof.* For $n \in \mathbb{N}$ the distribution of geometric random k-SAT is $\vec{p}^{(n)} = (p_1^{(n)}, p_2^{(n)}, \ldots, p_n^{(n)})$ with

$$p_i^{(n)} = \frac{b \cdot (1 - b^{-1/n})}{b - 1} \cdot b^{-(i-1)/n}.$$

As always, the first moment method tells us that instances are a. a. s. unsatisfiable for $m \in \omega(n)$. Additionally, we can use Corollary 5.6, which tells us that instances are a. a. s. satisfiable if $m \leq \varepsilon_m \cdot m^\star$ for constants $\varepsilon_m \in (0, 1)$, where $m^\star = \frac{2\cdot(b-1)}{(b+1)\cdot\ln b} \cdot n$ is the threshold for geometric random 2-SAT with base $b$. Thus, instances are a. a. s. satisfiable for $m \in o(n)$. This implies that the asymptotic threshold for geometric random $k$-SAT is $m^\star \in \Theta(n)$. ■

## 5.4 Remarks

Note that the toolkit we provide is not exhaustive. It suffices to derive the asymptotic threshold function for some ensembles of probability distributions, but not for all of them. This is the case for power-law random $k$-SAT with exponent $\beta = \frac{2k-1}{k-1}$. If we draw a connection to non-uniform random 2-SAT we might have an explanation for this phenomenon. The first moment method asserts that the asymptotic threshold function is $O(n)$. This matches the lower bound for power-law random $k$-SAT with $\beta > \frac{2k-1}{k-1}$. Corollary 4.16 assumes that the largest clause probability dominates, which holds for power-law random $k$-SAT with $\beta < \frac{2k-1}{k-1}$. However, we do not have a result that works well if the largest clause probability does not dominate and the threshold is $o(n)$. This is the case for power-law random 2-SAT with exponent $\beta = 3$ and we suspect to also be the case for power-law random $k$-SAT with exponent $\beta = \frac{2k-1}{k-1}$ in general.

# 6 Sharpness in Non-Uniform Random k-SAT

*This chapter is based on the publication [FR18], which is joint work with Tobias Friedrich. The results from that paper have been heavily reworked. We had to correct Lemma 6.1 and Lemma 6.22. The new versions of those lemmas have slightly stronger prerequisites, which also required us to adjust our main theorems.*

In this chapter we show that the threshold is sharp if the ensemble of probability distributions satisfies certain requirements. Friedgut [Fri99] showed that random $k$-SAT has a sharp threshold for all $k \geq 3$. Surprisingly, his result only requires knowledge of the asymptotic threshold function and does not imply an exact function or leading constants for the satisfiability threshold. We will generalize Friedgut's result to non-uniform random $k$-SAT. To that end we use the proof framework provided by Friedgut [Fri99] as well as the sharp threshold theorem by Friedgut and Bourgain [Fri99] in the version found in O'Donnell's book "The analysis of Boolean Functions" [ODo14]. We will go into more detail about the proof in Section 6.4. However, there is one big problem if we want to show sharpness for the non-uniform clause-drawing model. The Sharp Threshold Theorem and thus the whole sharpness proof only holds on product probability spaces. That means it holds for the clause-flipping model $\mathcal{F}^N$, but not for the clause-drawing model $\mathcal{D}^N$. Therefore, we relate the two models in Section 6.1. We show under which requirements their asymptotic satisfiability thresholds coincide. Furthermore, we show that under the same assumptions sharpness of the clause-flipping model implies sharpness for its clause-drawing equivalent. We will use these results as follows.

1. Show asymptotic threshold function for $\mathcal{D}^N$. See Chapter 5.
2. The same asymptotic threshold function holds for $\mathcal{F}^N$. See Lemma 6.4.
3. The asymptotic threshold function can be used to show sharpness of that threshold in $\mathcal{F}^N$. See Theorem 6.12.
4. Sharpness in $\mathcal{F}^N$ implies sharpness in $\mathcal{D}^N$. See Lemma 6.5.

Thus, our framework allows us to show sharpness of the satisfiability threshold for the clause-drawing model if only the asymptotic threshold function for that model is known.

## 6.1 Relation of Clause Flipping and Clause Drawing

We start this half of the chapter by relating our two models for non-uniform random $k$-SAT. Up to this point, we concentrated on the clause-drawing model

$\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$. In that model $m$ clauses are drawn with repetition according to some probability distribution, which is derived from the probability distribution $\vec{p}$ of the Boolean variables. In the clause-flipping model $\mathcal{F}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s)$ we flip a coin for each of the $\binom{n}{k} \cdot 2^k$ possible clauses and add it to the formula if the coin flip is a success. We can choose the probability for each clause in $\mathcal{F}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s)$ to be the same as the probability to draw the clause in $\mathcal{D}^N(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ times the number of clauses $m$ (c. f. Definition 3.7) by setting $s = m$. Then the two models exhibit a similar probability to generate satisfiable instances for $s, m \in o(q_{\max}^{-1/2})$, where $q_{\max}$ is the maximum clause probability of $\mathcal{D}^N$. If the satisfiability threshold is in this region of the scaling parameters $s$ and $m$, the two models also exhibit similar threshold behavior. This is what we are going to show in this first section.

However, we first want to generalize our models a bit to encompass arbitrary clause probabilities. We are given a number of variables $n$ and a clause size $k$ and let $N = \binom{n}{k} \cdot 2^k$ denote the total number of different clauses. We assume that these clauses are in some fixed order, so we can identify them by indices $i \in [N]$. Thus, we can also encode formulas as sets of clause indices and let $\Phi_I$ denote the formula which contains the $i$-th clause iff $i \in I$. To that end, we let $\overline{I} = [N] \setminus I$.

For our general clause-flipping model $\mathcal{F}(n, k, (\vec{q}^{(n)})_{n\in\mathbb{N}}, s)$ we assume to be given a number of variables $n$, a clause length $k$, an ensemble of normalized clause probability distributions $(\vec{q}^{(n)})_{n\in\mathbb{N}} = \left({q_1}^{(n)}, \ldots, {q_N}^{(n)}\right)_{n\in\mathbb{N}}$, i. e. $\sum_{i\in[N]} {q_i}^{(n)} = 1$ for all $n \in \mathbb{N}$, and a scaling factor $s \in [0, 1/\min_{i\in[N]}({q_i}^{(n)})]$. The probability to flip a clause is now ${q_i}^{(n)}(s) = \min\left(s \cdot {q_i}^{(n)}, 1\right)$ and the probability to generate formula $\Phi_I$ is

$$\Pr_{\Phi\sim\mathcal{F}}[\,\Phi = \Phi_I\,] = \prod_{i\in I} {q_i}^{(n)}(s) \cdot \prod_{i\in\overline{I}} (1 - {q_i}^{(n)}(s)). \tag{6.1}$$

We also define a general clause-drawing model $\mathcal{D}(n, k, (\vec{q}^{(n)})_{n\in\mathbb{N}}, m)$, where $m$ $k$-clauses over $n$ variables are drawn with repetition according to a normalized probability distribution from a given ensemble $(\vec{q}^{(n)})_{n\in\mathbb{N}} = \left({q_1}^{(n)}, \ldots, {q_N}^{(n)}\right)_{n\in\mathbb{N}}$. As for the non-uniform random $k$-SAT models, we interpret all parameters (including $N$) as functions in $n$ and omit the input parameter $n$ for the sake of simplicity. Note that the probabilities in $\vec{q}$ are not necessarily proportional to the products of given variable probabilities, but can be chosen arbitrarily. The non-uniform random $k$-SAT models we consider are special cases of these models, where the clause probabilities are derived from products of variable probabilities, i. e.

$$q_i = C\frac{k!}{2^k} \prod_{\ell\in c_i} p(|\ell|)$$

with

$$C = \left(k! \cdot \sum_{J\in\mathcal{P}_k(\{1,2,\ldots,n\})} \prod_{j\in J} p_j\right)^{-1}.$$

Note that in $\mathcal{D}$ clauses are drawn with repetition, i. e. it could happen that we draw fewer than $m$ clauses. In $\mathcal{F}$ we could draw any number of clauses, although the expected number is equal to the scaling factor $s$. This makes the comparison of these models difficult. However, if we condition on a certain number of clauses being drawn, the conditional probabilities for both models on the same input parameters are very close. More precisely, the total variation distance of the conditional probabilities in both models is $o(1)$ if $s \cdot m \in o(q_{\max}^{-1})$. This is what the following simple lemma shows.

▶ **Lemma 6.1.** Let $\mathcal{D}(n, k, (\vec{q}^{(n)})_{n\in\mathbb{N}}, m)$ be a clause-drawing model and let $\mathcal{F}(n, k, (\vec{q}^{(n)})_{n\in\mathbb{N}}, s)$ be a clause-flipping model with the same ensemble of clause probabilities $(\vec{q}^{(n)})_{n\in\mathbb{N}}$. Then, for all events $\mathcal{E}$ and all functions $s, m$ such that $s \cdot m \in o({q_{\max}}^{-1})$ it holds that

$$\left| \Pr_{\Phi\sim\mathcal{F}(s)} [\, \mathcal{E} \mid \{m \text{ clauses flipped}\} \,] - \Pr_{\Phi\sim\mathcal{D}(m)} [\, \mathcal{E} \mid \{\text{no duplicates}\} \,] \right| \in o(1).$$

◀

*Proof.* Let $D_m$ denote the event that exactly $m$ different clauses are drawn in $\mathcal{D}$ and let $F_m$ denote the event that exactly $m$ clauses are flipped in $\mathcal{F}$. Due to these conditions, the elementary events of the conditional probability spaces are the formulas with exactly $m$ clauses. Let $S \in \mathcal{P}_m([N])$ encode a formula $\Phi_S$ with exactly $m$ clauses. Due to the requirement $s \cdot m \in o({q_{\max}}^{-1})$, it also holds that $s \cdot q_{\max} < 1$ for all sufficiently large $n$. Thus our clause probabilities will not exceed one and we can simply write $q_i(s) = q_i \cdot s$. It holds that

$$\begin{aligned}
\Pr_{\Phi\sim\mathcal{F}(s)} [\, \Phi = \Phi_S \mid F_m \,] &= \frac{\prod_{i\in S} s \cdot q_i \cdot \prod_{i\in\overline{S}} (1 - s \cdot q_i)}{\sum_{S'\in\mathcal{P}_m([N])} \left( \prod_{j\in S'} s \cdot q_j \cdot \prod_{j\in\overline{S'}} (1 - s \cdot q_j) \right)} \\
&= \frac{\prod_{i\in S} \frac{s\cdot q_i}{1-s\cdot q_i} \cdot \prod_{i\in[N]} (1 - s \cdot q_i)}{\sum_{S'\in\mathcal{P}_m([N])} \left( \prod_{j\in S'} \frac{s\cdot q_j}{1-s\cdot q_j} \cdot \prod_{j\in[N]} (1 - s \cdot q_j) \right)} \\
&= \frac{\prod_{i\in S} \frac{q_i}{1-s\cdot q_i}}{\sum_{S'\in\mathcal{P}_m([N])} \left( \prod_{j\in S'} \frac{q_j}{1-s\cdot q_j} \right)}. \qquad (6.2)
\end{aligned}$$

It also holds that

$$\Pr_{\Phi\sim\mathcal{D}(m)} [\, \Phi = \Phi_S \mid D_m \,] = \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])} \left( \prod_{j\in S'} q_j \right)}, \qquad (6.3)$$

since we assume the clause probabilities $\vec{q}$ to be normalized. We can now see that

$$\Pr_{\Phi\sim\mathcal{F}(s)} [\, \Phi = \Phi_S \mid F_m \,] = \frac{\prod_{i\in S} \frac{q_i}{1-s\cdot q_i}}{\sum_{S'\in\mathcal{P}_m([N])} \left( \prod_{j\in S'} \frac{q_j}{1-s\cdot q_j} \right)}$$

$$\geq \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)} \cdot (1 - s\cdot q_{\max})^m$$
$$\geq \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)} \cdot (1 - s\cdot q_{\max}\cdot m)$$
$$= (1-o(1))\cdot \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)}$$

due to the requirement $s\cdot m \in o(q_{\max}^{-1})$. Furthermore

$$\Pr_{\Phi\sim\mathcal{F}(s)}[\,\Phi = \Phi_S \mid F_m\,] = \frac{\prod_{i\in S}\frac{q_i}{1-s\cdot q_i}}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'}\frac{q_j}{1-s\cdot q_j}\right)}$$
$$\leq \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)} \cdot \left(\frac{1}{1-s\cdot q_{\max}}\right)^m$$
$$\leq \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)} \cdot \exp\left(m\cdot\frac{s\cdot q_{\max}}{1-s\cdot q_{\max}}\right)$$
$$= (1+o(1))\cdot \frac{\prod_{i\in S} q_i}{\sum_{S'\in\mathcal{P}_m([N])}\left(\prod_{j\in S'} q_j\right)}$$

due to the same requirement $s\cdot m \in o(q_{\max}^{-1})$. This establishes the result as desired. ■

For the values of $m$ and $s$ we consider this will result in a small enough total variation distance to compare the threshold behavior. We also need that the conditional probability for a monotone property to hold conditioned on the number of clauses flipped is non-decreasing in the number of clauses we condition on. This is shown in the following lemma. The lemma actually holds in a much more general setting, as long as the clause probabilities are fixed. In the context of $\mathcal{F}$ it holds as long as the scaling parameter $s$ is fixed, i. e. we can incorporated the scaling factor $s$ into the clause probabilities $(\vec{q}_i)_{i\in[N]}$.

▶ **Lemma 6.2.** Let $\mathcal{F}$ be a clause-flipping model with an arbitrary ensemble of clause probability distributions $(\vec{q}^{(n)})_{n\in\mathbb{N}}$ and let $P$ be a monotone property. For $i, j \in [N]$ with $i \leq j$ it holds that

$$\Pr_{\Phi\sim\mathcal{F}}[\,P(\Phi)=1 \mid |\Phi| = i\,] \leq \Pr_{\Phi\sim\mathcal{F}}[\,P(\Phi)=1 \mid |\Phi| = j\,].$$

◀

*Proof.* We consider the random process described in Algorithm 1 and show that it generates each Boolean formula $\Phi$ with the same probability as $\mathcal{F}$. The process imposes an artificial order on the clauses of formulas $\Phi\sim\mathcal{F}$. This results in a probability space of $(\Omega', \pi')$ with

$$\Omega' = \{(i_1, i_2, \ldots, i_l) \mid l\in[N],\ \{i_1,\ldots,i_l\}\in\mathcal{P}_l([N])\},$$

**Algorithm 1:** Random process generating $\Phi \sim \mathcal{F}$ with artificial order imposed on positions of clauses in $\Phi$

```
l := 0;
I := ∅;
while true do
    b := Ber( Pr_{Φ∼F} [ |Φ| = l | |Φ| ≥ l ] );
    if b = 1 then
        return Φ with c_i ∈ Φ ↔ i ∈ I;
    else
        l := l + 1;
        choose i ∉ I with probability β(i, I);
        I := I ∪ {i};
```

i.e. $\Omega'$ contains tuples of indices from $[N]$ of size 0 to $N$ with no repetitions. These correspond to the clauses in $\Phi$ and the order in which they were flipped.

Let $b_i = \frac{q'_i}{1-q'_i}$. We now choose

$$\beta(i,I) = \frac{\Pr[|\Phi| = |I|]}{\Pr[|\Phi| = |I|+1]} \cdot b_i \cdot \\ \cdot \frac{\sum_{S \subseteq I} \left( \frac{1}{|I|+1-|S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|I|-|S|}([N] \setminus (I \cup \{i\}))} \prod_{s \in S'} b_s \right) \right)}{\sum_{S \in \mathcal{P}_{|I|}([N])} \prod_{s \in S} b_s}.$$

To make sure that $\beta(i,I)$ is a legal probability distribution for each $I \subseteq [N]$, it has to hold that $\sum_{i \notin I} \beta(i,I) = 1$. We will see that this is the case by showing

$$\sum_{i \notin I} \left( b_i \sum_{S \subseteq I} \left( \frac{1}{|I|+1-|S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|I|-|S|}([N] \setminus (I \cup \{i\}))} \prod_{s \in S'} b_s \right) \right) \right) \\ = \sum_{S \in \mathcal{P}_{|I|+1}([N])} \prod_{s \in S} b_s. \quad (6.4)$$

This implies

$$\begin{aligned} \sum_{i \notin I} \beta(i,I) &= \frac{\Pr[|\Phi| = |I|]}{\Pr[|\Phi| = |I|+1]} \cdot \frac{\sum_{S \in \mathcal{P}_{|I|+1}([N])} \prod_{s \in S} b_s}{\sum_{S \in \mathcal{P}_{|I|}([N])} \prod_{s \in S} b_s} \\ &= \frac{\Pr[|\Phi| = |I|]}{\Pr[|\Phi| = |I|+1]} \cdot \frac{\sum_{S \in \mathcal{P}_{|I|+1}([N])} \prod_{s \in S} b_s}{\sum_{S \in \mathcal{P}_{|I|}([N])} \prod_{s \in S} b_s} \cdot \frac{\prod_{i \in [N]} (1-q_i)}{\prod_{i \in [N]} (1-q_i)} \\ &= \frac{\Pr[|\Phi| = |I|]}{\Pr[|\Phi| = |I|+1]} \cdot \frac{\sum_{S \in \mathcal{P}_{|I|+1}([N])} \prod_{s \in S} q_s \prod_{s \notin S} (1-q_s)}{\sum_{S \in \mathcal{P}_{|I|}([N])} \prod_{s \in S} q_s \prod_{s \notin S} (1-q_s)} \end{aligned}$$

$$= \frac{\Pr[|\Phi| = |I|]}{\Pr[|\Phi| = |I| + 1]} \cdot \frac{\Pr[|\Phi| = |I| + 1]}{\Pr[|\Phi| = |I|]} = 1,$$

since

$$\Pr[|\Phi| = |I|] = \sum_{S \in \mathcal{P}_{|I|}([N])} \left( \prod_{s \in S} q_i \cdot \prod_{i \notin S} (1 - q_i) \right)$$

due to equation (6.1). In order to prove equation (6.4) we have to count how often each $\prod_{s \in K} b_s$ appears on the left-hand side of the equation for some fixed $K \subseteq [N]$ with $|K| = |I| + 1$. It holds that for each $i \in K \setminus I$ we have to choose exactly $S = I \cap K$ and $S' = K \setminus (I \cup \{i\})$. There are $|K \setminus I| \geq 1$ elements which generate $\prod_{s \in K} b_s$ with a factor of $\frac{1}{|I|+1-|S|} = \frac{1}{|I|+1-|I \cap K|} = \frac{1}{|K \setminus I|}$ each. Therefore, their appearances sum up to exactly $\prod_{s \in K} b_s$. Since this holds for all $K \subseteq [N]$ with $|K| = |I| + 1$ and no different-sized subsets of $[N]$ can be generated, equation (6.4) holds.

Now we want to show that this random process generates each formula $\Phi_J$ containing exactly the clauses indexed by $J \subseteq [N]$ with the same probability as $\mathcal{F}^N$. We use induction over $l$ to show that for all $J \subseteq [N]$ with $|J| = l$ the random process generates $I = J$ (and therefore a formula $\Phi_J$) with probability $\prod_{i \in J} q_i \prod_{i \notin J} (1 - q_i)$. The base case is $J = \emptyset$, i.e. $\Phi_\emptyset$ being the empty formula. The probability to generate this formula with our random process is

$$\Pr_{\Phi \sim \mathcal{F}}[\, |\Phi| = 0 \mid |\Phi| \geq 0 \,] = \Pr_{\Phi \sim \mathcal{F}}[\, |\Phi| = 0 \,] = \Pr_{\Phi \sim \mathcal{F}}[\, \Phi = \Phi_\emptyset \,],$$

which means, the induction hypothesis holds for this case. Now we want to go to $J \subseteq [N]$ with $|J| = l + 1$. The probability that our process generates $J$ is

$$\begin{aligned} \Pr[I = J] \quad = \quad & \sum_{i \in J} \Pr[I = J \setminus \{i\}] \cdot \beta(i, J \setminus \{i\}) \cdot \frac{\Pr[|\Phi| = l + 1 \mid |\Phi| \geq l + 1]}{\Pr[|\Phi| = l \mid |\Phi| \geq l]} \cdot \\ & \cdot (1 - \Pr[|\Phi| = l \mid |\Phi| \geq l]). \end{aligned}$$

This expression consists of the probability of choosing $J \setminus \{i\}$ in the first $l$ steps, but, instead of stopping after step $l$, continuing, choosing $i$ with probability $\beta(i, J \setminus \{i\})$, and then stopping after step $l + 1$. It holds that

$$\begin{aligned} & \Pr[I = J] \\ & = \sum_{i \in J} \Pr[I = J \setminus \{i\}] \cdot \beta(i, J \setminus \{i\}) \cdot \frac{\Pr[|\Phi| = l + 1 \mid |\Phi| \geq l + 1]}{\Pr[|\Phi| = l \mid |\Phi| \geq l]} \cdot \\ & \cdot (1 - \Pr[|\Phi| = l \mid |\Phi| \geq l]) \\ & = \sum_{i \in J} \Pr[I = J \setminus \{i\}] \cdot \beta(i, J \setminus \{i\}) \cdot \frac{\Pr[|\Phi| = l + 1 \mid |\Phi| \geq l + 1]}{\Pr[|\Phi| = l \mid |\Phi| \geq l]} \cdot \\ & \cdot \frac{\Pr[|\Phi| \geq l + 1]}{\Pr[|\Phi| \geq l]} \end{aligned}$$

$$= \sum_{i \in J} \Pr[I = J \setminus \{i\}] \cdot b_i \cdot \frac{\sum_{S \subseteq J \setminus \{i\}} \frac{1}{|J \setminus \{i\}|+1-|S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|J \setminus \{i\}|-|S|}([N] \setminus J)} \prod_{s \in S'} b_s \right)}{\sum_{S \in \mathcal{P}_{|J \setminus \{i\}|}([N])} \prod_{s \in S} b_s}$$

$$= \sum_{i \in J} \Pr[I = J \setminus \{i\}] \cdot b_i \cdot \frac{\sum_{S \subseteq J \setminus \{i\}} \frac{1}{|J|-|S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|J|-|S|-1}([N] \setminus J)} \prod_{s \in S'} b_s \right)}{\sum_{S \in \mathcal{P}_{|J|-1}([N])} \prod_{s \in S} b_s}$$

$$= \prod_{s \in J} q_s \prod_{s \notin J} (1 - q_s) \cdot \sum_{i \in J} \frac{\sum_{S \subseteq J \setminus \{i\}} \frac{1}{|J|-|S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|J|-|S|-1}([N] \setminus J)} \prod_{s \in S'} b_s \right)}{\sum_{S \in \mathcal{P}_{|J|-1}([N])} \prod_{s \in S} b_s}, \quad (6.5)$$

where we used the induction hypothesis and the definition $b_i = \frac{q_i}{1-q_i}$ in the last line. Again, it suffices to count the number of appearances of $\prod_{s \in K} q_s$ for each $K \subseteq [N]$ with $|K| = |J| - 1$ in the numerator. Since $i$ does never appear in the resulting products, we have to choose an $i \in J \setminus K$. There are $|J \setminus K| \geq 1$ such elements, since $|J| = |K| + 1$. Now $K$ can only be chosen if $S = J \cap K$ and $S' = K \setminus J$. The product then appears with a factor of $\frac{1}{|J|-|S|} = \frac{1}{|J|-|J \cap K|} = \frac{1}{|J \setminus K|}$ for each $i \in J \setminus K$. Since only products of $|J| - 1$ factors are created, it holds that

$$\sum_{i \in J} \sum_{S \subseteq J \setminus \{i\}} \frac{1}{|J| - |S|} \prod_{s \in S} b_s \cdot \left( \sum_{S' \in \mathcal{P}_{|J|-|S|-1}([N] \setminus J)} \prod_{s \in S'} b_s \right) = \sum_{S \in \mathcal{P}_{|J|-1}([N])} \prod_{s \in S} b_s.$$

This implies $\Pr[I = J] = \prod_{s \in J} q_s \prod_{s \notin J} (1 - q_s)$, because the sum at the right-hand-side of equation (6.5) equals one.

Now that we know that our random process creates $\Phi_J$ with the same probability as $\mathcal{F}^N$, we can show the result of the theorem. For an $A \in \Omega'$ let $W_A$ denote the event that the random process chooses at least $|A|$ elements and that the first $|A|$ elements it chooses are given by $A$, i. e. for $A = (a_1, a_2, \ldots, a_{|A|})$ the random process chooses $a_i$ in the $i$-th round. It holds that

$$\Pr[P(\Phi) = 1 \mid |\Phi| = l] = \sum_{\substack{A \in \Omega' : \\ |A| = l-1}} \Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l] \cdot \Pr[W_A \mid |\Phi| = l]. \quad (6.6)$$

We can now imagine executing the random process and having completed iteration $l - 1$. Since $W_A$ with $|A| = l - 1$ is independent of everything that happens after that iteration, it then holds that

$$\Pr[W_A \mid |\Phi| = l] = \frac{\Pr[W_A \wedge |\Phi| = l]}{\Pr[|\Phi| = l]}$$

$$
\begin{aligned}
&= \Pr[W_A] \cdot \frac{1 - \Pr[|\Phi| = l-1 \mid |\Phi| \geq l-1]}{\Pr[|\Phi| = l]} \cdot \Pr[|\Phi| = l \mid |\Phi| \geq l] \\
&= \Pr[W_A] \cdot \frac{\Pr[|\Phi| \geq l]}{\Pr[|\Phi| = l] \cdot \Pr[|\Phi| \geq l-1]} \cdot \Pr[|\Phi| = l \mid |\Phi| \geq l] \\
&= \Pr[W_A] \cdot \frac{\Pr[|\Phi| = l]}{\Pr[|\Phi| = l] \cdot \Pr[|\Phi| \geq l-1]} \\
&= \frac{\Pr[W_A]}{\Pr[|\Phi| \geq l-1]} \\
&= \frac{\Pr[W_A] \cdot \Pr[|\Phi| = l-1]}{\Pr[|\Phi| \geq l-1] \cdot \Pr[|\Phi| = l-1]} \\
&= \frac{\Pr[W_A] \cdot \Pr[|\Phi| = l-1 \mid |\Phi| \geq l-1]}{\Pr[|\Phi| = l-1]} \\
&= \Pr[W_A \mid |\Phi| = l-1]. \qquad (6.7)
\end{aligned}
$$

Furthermore, if an $A \in \Omega'$ already implies that the corresponding formula $\Phi$ satisfies $P(\Phi) = 1$, the same holds for all $\Phi'$ which contain all the clauses of $\Phi$ due to the monotonicity of $P$. This means, in that case

$$\Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l-1] = 1 = \Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l].$$

Otherwise, it holds that

$$\Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l-1] = 0 \leq \Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l].$$

We can plug these inequalities and equation (6.7) into equation (6.6) to get

$$
\begin{aligned}
&\Pr[P(\Phi) = 1 \mid |\Phi| = l] \\
&= \sum_{\substack{A \in \Omega' : \\ |A| = l-1}} \Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l] \cdot \Pr[W_A \mid |\Phi| = l] \\
&\geq \sum_{\substack{A \in \Omega' : \\ |A| = l-1}} \Pr[P(\Phi) = 1 \mid W_A, |\Phi| = l-1] \cdot \Pr[W_A \mid |\Phi| = l-1] \\
&= \Pr[P(\Phi) = 1 \mid |\Phi| = l-1].
\end{aligned}
$$

By iteratively using this inequality we get the result as desired. ■

Lemma 6.2 also holds if we condition on the number of clauses in pairwise disjoint subsets of $[N]$. This also holds as long as the clause probabilities are fixed.

▶ **Lemma 6.3.** Let $\mathcal{F}$ be a clause-flipping model with clause probability ensemble $(\vec{q}^{(n)})_{n \in \mathbb{N}}$ and let $P$ be a monotone property. Let $S_1, S_2 \ldots, S_t \subseteq N$ be pairwise disjoint. For each $(i_1, \ldots, i_t),(j_1, \ldots, j_t) \in [|S_1|] \times [|S_2|] \times \ldots \times [|S_t|]$

with $(i_1, \ldots, i_t) \leq (j_1, \ldots, j_t)$ coordinate-wise it holds that

$$\Pr_{\Phi \sim \mathcal{F}}\left[ P(\Phi) = 1 \;\middle|\; \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right] \leq \Pr_{\Phi \sim \mathcal{F}}\left[ P(\Phi) = 1 \;\middle|\; \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = j_l \right],$$

where $|\Phi \cap \Phi_{S_l}|$ denotes the number of clause indices in $S_l$ that $\Phi$ contains.

◀

*Proof.* The proof of this lemma follows the same lines as the one for Lemma 6.2. We use Algorithm 1 to flip the clauses with indices in each $S_l$ separately. Let $R = [N] \setminus \bigcup_{l=1}^{t} S_l$ be the set of coordinates belonging to none of the subsets. We still assume that the clauses in $R$ are generated according to the original product probability, i.e. we impose no order on its clauses. As in the proof of Lemma 6.2, $\Omega'$ denotes the sample space we get from imposing a sampling order on clauses as defined by the random process. $A_1, \ldots, A_t$ represent the order of sampled clause indices from $S_1, \ldots, S_t$, respectively. $W_A$ denotes the event that at least $|A|$ clauses are sampled and the first $|A|$ of them in the order given by $A$. We will only argue on the new probability space created by the process described in Algorithm 1 applied to the subsets $A_1, \ldots, A_t$. Thus, we will omit the probability space from our probabilities. In the end we will see that the result holds for the original probability space as well. Now let $j \in [t]$ be arbitrary but fixed. It holds that

$$\begin{aligned}
&\Pr\left[ P(\Phi) = 1 \;\middle|\; \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right] \\
&= \sum_{I_R \subseteq R} \sum_{\substack{A_1, \ldots, A_t \in \Omega'_1 \times \ldots \times \Omega'_t: \\ |A_j| = i_j - 1 \;\wedge \\ \wedge \forall l \in [t] \setminus \{j\}:\; |A_l| = i_l}} \Pr\left[ P(\Phi) = 1 \;\middle|\; \Phi \cap \Phi_R = \Phi_{I_R} \bigwedge_{l=1}^{t} W_{A_l} \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right] \cdot \\
&\cdot \frac{\Pr\left[ \Phi \cap \Phi_R = \Phi_{I_R} \bigwedge_{l=1}^{t} W_{A_l} \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right]}{\Pr\left[ \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right]}.
\end{aligned}$$

Since the clauses in all the subsets we consider are flipped independently, it holds that

$$\begin{aligned}
&\frac{\Pr\left[ \Phi \cap \Phi_R = \Phi_{I_R} \bigwedge_{l=1}^{t} W_{A_l} \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right]}{\Pr\left[ \bigwedge_{l=1}^{t} |\Phi \cap \Phi_{S_l}| = i_l \right]} \\
&\qquad = \Pr\left[ \Phi \cap \Phi_R = \Phi_{I_R} \right] \cdot \frac{\prod_{l=1}^{t} \Pr\left[ W_{A_l} \wedge |\Phi \cap \Phi_{S_l}| = i_l \right]}{\prod_{l=1}^{t} \Pr\left[ |\Phi \cap \Phi_{S_l}| = i_l \right]}.
\end{aligned}$$

With the same argumentation for $W_{A_j}$ as for $W_A$ in the proof of Lemma 6.2 it

now holds that

$$\Pr\left[P(\Phi)=1 \,\middle|\, \Phi\cap\Phi_R=\Phi_{I_R}\bigwedge_{l=1}^{t}W_{A_l}\bigwedge_{l=1}^{t}|\Phi\cap\Phi_{S_l}|=i_l\right]$$
$$\geq \Pr\left[P(\Phi)=1 \,\middle|\, \Phi\cap\Phi_R=\Phi_{I_R}\bigwedge_{l=1}^{t}W_{A_l}\bigwedge_{\substack{l=1\\l\neq j}}^{t}|\Phi\cap\Phi_{S_l}|=i_l\wedge|\Phi\cap\Phi_{S_j}|=i_j-1\right]$$

and that

$$\Pr\left[W_{A_j} \mid |\Phi\cap\Phi_{S_j}|=i_j\right]=\Pr\left[W_{A_j} \mid |\Phi\cap\Phi_{S_j}|=i_j-1\right],$$

which gives us

$$\Pr\left[P(\Phi)=1 \,\middle|\, \bigwedge_{l=1}^{t}|\Phi\cap\Phi_{S_l}|=i_l\right]$$
$$\geq \Pr\left[P(\Phi)=1 \,\middle|\, \bigwedge_{l=1}^{t}|\Phi\cap\Phi_{S_l}|=i_l\wedge|\Phi\cap\Phi_{S_j}|=i_j-1\right].$$

Using this inequality iteratively, we get the result as desired. It is easy to see that any formula is created with the same probability in $\mathcal{F}$ and by using our random process for each of the subsets $S_1,\ldots,S_t$. We can simply view the model $\mathcal{F}$ as a product of models $\mathcal{F}_{S_l}$, which only sample clauses with indices in $S_l$ instead of clauses with all indices in $[N]$. According to our results in Lemma 6.2, the process on each of these models creates each subformula on $S_l$ with the same probabilities as $\mathcal{F}_{S_l}$. Thus the product of those probabilities is the same as the original sampling probability of $\mathcal{F}$. ■

The following lemma shows under which conditions the asymptotic thresholds of clause-drawing and clause-flipping non-uniform random $k$-SAT with the same ensemble of probability distributions coincide. Note that we show the result for monotone functions in general, i. e. in the context of the lemma the thresholds go from probabilities approaching zero to probabilities approaching one, whereas the satisfiability threshold goes from one to zero. In the context of satisfiability the monotone property would be unsatisfiability, i. e. the probability to generate an unsatisfiable formula increases with the scaling parameter $s$.

▶ **Lemma 6.4.** Let $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ be an ensemble of variable probability distributions and let $m^\star\in\omega(1)$ be an asymptotic threshold for a monotone property $P$ on $\mathcal{D}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},m)$ with respect to $m$. If $m^\star\in o(q_{\max}^{-1/2})$, then $s^\star=m^\star$ is an asymptotic threshold for $P$ on $\mathcal{F}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},s)$ with respect to $s$. ◀

*Proof.* For the sake of simplicity, we will use the shorthand notations $\mathcal{F}^N(s)$ for $\mathcal{F}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},s)$ and $\mathcal{D}^N(m)$ for $\mathcal{D}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},m)$. We want to

show that $s^\star = m^\star$ is an asymptotic threshold function for $P$ on $\mathcal{F}^N(s)$ with respect to $s$.

Let us consider any fixed $s \in o(s^\star)$. We need to show that

$$\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \,] \in o(1).$$

First, we show that a. a. s. formulas $\Phi \sim \mathcal{F}^N(s)$ consist of at most $s' = s + s^{2/3}$ clauses. This holds due to a simple Chernoff bound. The model is defined in such a way that $\mathbb{E}[\, |\Phi| \,] = s$. Since each clause is flipped independently at random Theorem 2.6 tells us that

$$\begin{aligned}\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| > s' \,] &= \Pr_{\Phi \sim \mathcal{F}^N(s)}\left[\, |\Phi| > (1 + s^{-1/3}) \cdot \mathbb{E}[\, |\Phi| \,] \,\right] \\ &\leq \exp\left(-\frac{s^{-2/3} \cdot s}{3}\right) \in o(1)\end{aligned}$$

for $s \in \omega(1)$. It holds that

$$\begin{aligned}&\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \,] \\ &= \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| \leq s' \,] + \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| > s' \,] \\ &\leq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| \leq s' \,] + o(1).\end{aligned}$$

We can further bound this probability as follows

$$\begin{aligned}&\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| \leq s' \,] \\ &= \sum_{j=0}^{s'} \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| = j \,] \\ &= \sum_{j=0}^{s'} \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = j \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| = j \,] \\ &\leq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \cdot \sum_{j=0}^{s'} \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| = j \,] \\ &= \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| \leq s' \,] \\ &\leq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,],\end{aligned}$$

since

$$\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = j \,] \leq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,]$$

due to Lemma 6.2. Lemma 6.1 now yields

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \mid |\Phi|=s'\,] \leq \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1 \mid |\Phi|=s'\,] + o(1),$$

since $s\cdot m = s\cdot s' \in o(q_{\max}^{-1})$. It remains to bound $\Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\,]$ related to this latter conditional probability. Thus, we will now bound the probability that a clause occurs twice in $\mathcal{D}^N(s')$.

$$\begin{aligned}\Pr_{\Phi\sim\mathcal{D}^N(s')}[\,|\Phi|<s'\,] &\leq \sum_{i,j\in[s']} \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,i\text{-th and } j\text{-th clause identical}\,]\\ &\leq \binom{s'}{2}\cdot q_{\max} \in o(1),\end{aligned}$$

since $s'^2\cdot q_{\max} \in \Theta(s^2\cdot q_{\max}) \in o(1)$ due to $s\in o(m^\star)$ and $m^\star \in o(q_{\max}^{-1/2})$. Thus,

$$\begin{aligned}\Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1 \mid |\Phi|=s'\,] &= \frac{\Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\wedge|\Phi|=s'\,]}{\Pr_{\Phi\sim\mathcal{D}^N(s')}[\,|\Phi|=s'\,]}\\ &\leq \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\wedge|\Phi|=s'\,]+o(1)\\ &\leq \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\,]+o(1).\end{aligned}$$

Thus, we get

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] \leq \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\,]+o(1).$$

Since $m^\star$ is an asymptotic threshold function and $s'\in o(m^\star)$, this probability is $o(1)$ as desired.

Now let us consider an $s\in\omega(s^\star)$. The argument is similar to the case $s\in o(s^\star)$. However, we have to make sure that $s\in o(q_{\max}^{-1/2})$ still holds in order to be able to use Lemma 6.1. Thus, we define a new function $s_2$ so that $s_2\in\omega(s^\star)$ and $s_2\in o(q_{\max}^{-1/2})$. This is possible, since we assume $s^\star = m^\star \in o(q_{\max}^{-1/2})$. One possible function with these properties could be

$$s_2 = \sqrt{s^\star\cdot q_{\max}^{-1/2}}.$$

Now we define $\hat{s} = \min(s, s_2)$. This is the function we will actually consider. Lemma 3.9 tells us that the probability that $\mathcal{F}^N$ generates an instance with a monotone property $P$ is non-decreasing in $s$. For our original function $s$ and the smaller function $\hat{s}$ we just defined, this implies

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] \geq \Pr_{\Phi\sim\mathcal{F}^N(\hat{s})}[\,P(\Phi)=1\,].$$

Thus, any lower bound on the probability to generate instances with property $P$ at $\hat{s}$ carries over to $s$, while $\hat{s}$ also fulfills the requirement to be in $o(q_{\max}^{-1/2})$.

As before, we can use a Chernoff bound to show that $|\Phi| \geq s' = \hat{s} - \hat{s}^{2/3}$ with probability $1 - o(1)$. It now holds that

$$\Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \,] \geq \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \wedge |\Phi| \geq s' \,].$$

Again,

$$\begin{aligned}
&\Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \wedge |\Phi| \geq s' \,] \\
&= \sum_{j=s'}^{N} \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = j \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, |\Phi| = j \,] \\
&\geq \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \cdot \sum_{j=s'}^{N} \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, |\Phi| = j \,] \\
&= \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, |\Phi| \geq s' \,] \\
&\geq \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] - o(1),
\end{aligned}$$

since

$$\Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = j \,] \geq \Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = s' \,]$$

due to Lemma 6.2. Lemma 6.1 now yields

$$\Pr_{\Phi \sim \mathcal{F}^N(\hat{s})}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \geq \Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] - o(1),$$

since $\hat{s} \cdot m = \hat{s} \cdot s' \in o(q_{\max}^{-1})$. With the same condition it holds that

$$\Pr_{\Phi \sim \mathcal{D}^N(s')}[\, |\Phi| < s' \,] = o(1)$$

and thus

$$\begin{aligned}
&\Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \\
&\geq \Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \wedge |\Phi| = s' \,] \\
&= \Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \,] - \Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \wedge |\Phi| < s' \,] \\
&= \Pr_{\Phi \sim \mathcal{D}^N(s')}[\, P(\Phi) = 1 \,] - o(1),
\end{aligned}$$

which implies

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] \geq \Pr_{\Phi\sim\mathcal{D}^N(s')}[\,P(\Phi)=1\,] - o(1) = 1 - o(1)$$

as desired, since $s' \in \omega(m^\star)$ and $m^\star$ is an asymptotic threshold function. Both results together imply that $s^\star$ is an asymptotic threshold function for the property $P$ with respect to the parameter $s$. ■

Using essentially the same proof we can show that sharpness of a threshold for a monotone property in $\mathcal{F}^N$ carries over to $\mathcal{D}^N$. This is shown in the following lemma.

▶ **Lemma 6.5.** Let $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ be an ensemble of variable probability distributions on $n$ variables each and let $s^\star \in \omega(1)$ be a sharp threshold for $P$ on $\mathcal{F}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},s)$ with respect to $s$. If $s^\star \in o(q_{\max}^{-1/2})$, then $m^\star = s^\star$ is a sharp threshold for $P$ on $\mathcal{D}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},m)$ with respect to $m$. ◀

*Proof.* We want to show that $m^\star = s^\star$ is a sharp threshold function for $P$ on $\mathcal{D}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},m)$ with respect to $m$. For the sake of simplicity, we will use the shorthand notations $\mathcal{D}^N(m)$ and $\mathcal{F}^N(s)$ again.

First, let us consider any function $m \leq (1-\varepsilon_m)\cdot m^\star$ for a constant $\varepsilon_m \in (0,1)$. We have to show that

$$\Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\,] \in o(1).$$

As before, it holds that

$$\Pr_{\Phi\sim\mathcal{D}^N(m)}[\,|\Phi|<m\,] \in o(1)$$

if $m \in o(q_{\max}^{-1/2})$. Thus,

$$\begin{aligned}
&\Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\,]\\
&= \Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\wedge|\Phi|=m\,] + \Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\wedge|\Phi|<m\,]\\
&\leq \Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\wedge|\Phi|=m\,] + o(1)\\
&\leq \Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\mid|\Phi|=m\,] + o(1).
\end{aligned}$$

According to Lemma 6.1

$$\Pr_{\Phi\sim\mathcal{D}^N(m)}[\,P(\Phi)=1\mid|\Phi|=m\,] \leq \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\mid|\Phi|=m\,] + o(1)$$

for *any* $s$ with $s\cdot m \in o(q_{\max}^{-1})$. That means, we have to choose an appropriate $s$ below the sharp threshold, but well above $m$ so that the size of a formula

generated with $\mathcal{F}^N(s)$ is a. a. s. above $m$. Again, we can use Chernoff bounds to show that a. a. s. $|\Phi| \geq s - s^{2/3}$. Thus, we can choose some $s = m + o(m)$ so that $s' = s - s^{2/3} \geq m$ for sufficiently large $m$. One possibility to choose a suitable $s$ is $s = m + m^{5/6}$. This is still below the sharp threshold, since $m = (1 - \varepsilon_m) \cdot s^\star$ and thus $s = m + o(m) \leq (1 - \varepsilon_m + o(1)) \cdot s^\star \leq (1 - \varepsilon_m/2) \cdot s^\star$ for sufficiently large $n$. We can now see that

$$\begin{aligned}\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \,] &\geq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \wedge |\Phi| \geq s' \,] \\ &= \sum_{j=s'}^{N} \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = j \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| = j \,]\end{aligned}$$

and according to Lemma 6.2 this yields

$$\begin{aligned}&\geq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] \cdot \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, |\Phi| \geq s' \,] \\ &= \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = s' \,] - o(1) \\ &\geq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = m \,] - o(1)\end{aligned}$$

again due to $s' \geq m$ and to Lemma 6.2. This implies

$$\begin{aligned}\Pr_{\Phi \sim \mathcal{D}^N(m)}[\, P(\Phi) = 1 \,] &\leq \Pr_{\Phi \sim \mathcal{D}^N(m)}[\, P(\Phi) = 1 \mid |\Phi| = m \,] + o(1) \\ &\leq \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \mid |\Phi| = m \,] + o(1) \\ &\leq \left( \Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \,] + o(1) \right) \in o(1),\end{aligned}$$

since $s \leq (1 - \varepsilon_m/2) \cdot s^\star$ as shown before and thus $\Pr_{\Phi \sim \mathcal{F}^N(s)}[\, P(\Phi) = 1 \,] \in o(1)$.

Now we consider a function $m \geq (1 + \varepsilon_m) \cdot s^\star$ for some constant $\varepsilon_m > 0$. As in the proof of the former lemma, we have to make sure that $m \in o(q_{\max}^{-1/2})$ is still satisfied. Thus, we consider $m' = (1 + \varepsilon_m) \cdot s^\star$ instead. Since the probability for monotone properties to hold in $\mathcal{D}^N(m)$ is non-decreasing in $m$ according to Lemma 3.8, it holds that

$$\Pr_{\Phi \sim \mathcal{D}^N(m)}[\, P(\Phi) = 1 \,] \geq \Pr_{\Phi \sim \mathcal{D}^N(m')}[\, P(\Phi) = 1 \,].$$

Thus, a lower bound on the probability at $m'$ also holds at $m$. It then holds that

$$\begin{aligned}\Pr_{\Phi \sim \mathcal{D}^N(m')}[\, P(\Phi) = 1 \,] &\geq \Pr_{\Phi \sim \mathcal{D}^N(m')}[\, P(\Phi) = 1 \wedge |\Phi| = m' \,] \\ &\geq \Pr_{\Phi \sim \mathcal{D}^N(m')}[\, P(\Phi) = 1 \mid |\Phi| = m' \,] - o(1),\end{aligned}$$

since $\Pr_{\Phi \sim \mathcal{D}^N(m)}[\, |\Phi| < m' \,] \in o(1)$ due to $m' \in o(q_{\max}^{-1})$. Again, we can choose

an $s$ such that a formula generated with $\mathcal{F}^N(s)$ a. a. s. consists of at most $s' = s+s^{2/3}$ clauses due to a Chernoff bound. We have to choose an $s \geq (1+\varepsilon'_m)\cdot s^\star$ such that $s' \leq m'$. A possible choice is $s = m' - m'^{5/6}$. This ensures $s' = s + s^{2/3} \leq m'$ as well as $s = m' - o(m') \geq (1 + \varepsilon_m/2) \cdot s^\star$ for sufficiently large $n$. Thus,

$$\begin{aligned}\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] &= \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \wedge |\Phi| \leq s'\,] + o(1)\\ &= o(1) + \sum_{j=0}^{s'} \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \mid |\Phi|=j\,] \cdot \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,|\Phi|=j\,]\end{aligned}$$

and according to Lemma 6.2 we get

$$\begin{aligned}&\leq \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \mid |\Phi|=s'\,] + o(1)\\ &\leq \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \mid |\Phi|=m'\,] + o(1).\end{aligned}$$

This implies

$$\Pr_{\Phi\sim\mathcal{D}^N(m')}[\,P(\Phi)=1\,] \geq \Pr_{\Phi\sim\mathcal{D}^N(m')}[\,P(\Phi)=1 \mid |\Phi|=m'\,] - o(1)$$

and according to Lemma 6.1

$$\begin{aligned}&\geq \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1 \mid |\Phi|=m'\,] - o(1)\\ &\geq \Pr_{\Phi\sim\mathcal{F}^N(s)}[\,P(\Phi)=1\,] - o(1) = 1 - o(1),\end{aligned}$$

since $s \geq (1+\varepsilon'_m)\cdot s^\star$ and $s^\star$ is a sharp threshold for $P$. This proves the result. ■

Now we know that the threshold behavior of the two models is equivalent if the asymptotic threshold function is in $o({q_{\max}}^{-1/2})$. This holds for all monotone properties, but it especially holds for unsatisfiability and thus for the behavior of the satisfiability threshold. We will proceed to show under which requirements the satisfiability threshold is sharp in $\mathcal{F}^N$.

## 6.2 Coarse Thresholds

At this point, we want to take a some time to talk about our definitions of sharp and coarse thresholds again. First, both definitions only apply to properties with an asymptotic threshold function. Then, we say that a threshold for a monotone property $P$ is sharp with respect to parameter $p$ of a random model $\mathcal{M}$ iff there is a function $p^\star$ so that for every constant $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] = \begin{cases}0, & \text{if } p' = (1-\varepsilon)\cdot p^\star\\ 1, & \text{if } p' = (1+\varepsilon)\cdot p^\star\end{cases}$$

Otherwise we call the threshold coarse. Note that for satisfiability we defined the thresholds the other way around, i. e. approaching one below the threshold and zero above it. However, the definitions are interchangable, since we can always just consider the property $\overline{P}$. For example, for unsatisfiability the threshold behaves as described above.

The notion of coarseness used by Friedgut, Bourgain and O'Donnell [Fri99; ODo14] is slightly different from the one we use. In their definition, they fix a constant $\varepsilon \in (0, 1/2)$. Now they consider the parameter values $p_0, p_1$ such that

$$\Pr_{\Phi \sim \mathcal{M}(p_0)} [\, P(\Phi) = 1 \,] = \varepsilon$$

and

$$\Pr_{\Phi \sim \mathcal{M}(p_1)} [\, P(\Phi) = 1 \,] = 1 - \varepsilon.$$

The parameter value $p^\star$ where the probability is $1/2$ is somewhere in the interval $[p_0, p_1]$ due to the probability function being non-decreasing. Between the parameter values $p_0$ and $p_1$ the probability is bounded away both from zero and one by the constant $\varepsilon$. The authors now compare the size of this interval with the size of the value $p^\star$ in the limit. If $\lim \frac{p_1 - p_0}{p^\star}$ approaches zero, the threshold is sharp. This agrees with our intuition of a sharp threshold: The size of the interval around $p^\star$, where the probability does not approach zero or one, grows slower than the threshold value. Or in mathematical terms, the size of this interval is $o(\hat{p})$, where $\hat{p}$ is our asymptotic threshold function. One can show that this definition of a sharp threshold is equivalent to the one we use (c. f. Lemma 6.6). However, the definition of coarse thresholds by Friedgut, Bourgain and O'Donnell is slightly different. They speak of a coarse threshold only if $\lim \frac{p_1 - p_0}{p^\star}$ is bounded away from zero. This is not a dichotomy as we consider it, i. e. if there is an asymptotic threshold function, we do not automatically have either a sharp or a coarse threshold, we could also have neither. That would be the case if the limit of $\frac{p_1 - p_0}{p^\star}$ was not defined.

Is this a problem if we want to apply the sharp threshold theorem? As it turns out, it is not! Due to our definition, coarseness means that for every $p^\star$ there is a constant $\varepsilon > 0$ so that either

$$\lim_{n \to \infty} \Pr_{\Phi \sim \mathcal{M}(p', n)} [\, P(\Phi) = 1 \,] \neq 0$$

for $p' = (1 - \varepsilon) \cdot p^\star$ or

$$\lim_{n \to \infty} \Pr_{\Phi \sim \mathcal{M}(p', n)} [\, P(\Phi) = 1 \,] \neq 1$$

for $p' = (1 + \varepsilon) \cdot p^\star$. Now consider $p^\star$ to be the critical value, i. e. the probability

at $p^\star$ is $1/2$. W. l. o. g. let us assume we have a $p' = (1-\varepsilon)\cdot p^\star$ with

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] \neq 0.$$

Recalling our definition of limits, this means that there is some constant $\varepsilon' > 0$ so that for infinitely many values of $n \in \mathbb{N}$

$$\Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] \geq \varepsilon'.$$

Due to the non-decreasing nature of the probability function, for those selected values of $n$ everything between $p'$ and $p^\star$ is bounded away from zero and one. Furthermore, we know that there is an asymptotic threshold function $\hat{p}$. It has to hold that $p', p^\star \in \Theta(\hat{p})$. We can now concentrate on the values of $n$ that satisfy this property and we know that there are infinitely many of those. This is not exactly the definition of a coarse threshold due to Friedgut, Bourgain and O'Donnell, but it certifies that there are parameter values $p', p^\star \in \Theta(\hat{p})$ between which the probability function is bounded away from zero and one by constants. Furthermore the size of the interval $[p', p^\star]$ is asymptotically the same as the value $p^\star$, i. e. $\lim \frac{p^\star - p'}{p^\star}$ is bounded away from zero and one for the values of $n$ we consider. As it turns out, this is already enough to apply the Sharp Threshold Theorem towards a contradiction.

For the sake of completeness, we finish this section by showing the equivalence of our sharpness definition and the one by Friedgut [Fri99].

▶ **Lemma 6.6.** Let $\hat{p}$ be an asymptotic threshold function for a monotone property $P$ with respect to parameter $p$ of a random model $\mathcal{M}$. Then it holds that there is a function $p^\star$ so that for every $\varepsilon_p > 0$

$$\lim_{n\to\infty} \Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] = \begin{cases} 0, & \text{if } p' = (1-\varepsilon_p)\cdot p^\star \\ 1, & \text{if } p' = (1+\varepsilon_p)\cdot p^\star \end{cases}$$

if and only if for every constant $\varepsilon \in (0, 1/2)$ it holds that $\lim_{n\to\infty} \frac{p_1 - p_0}{p_{1/2}} = 0$, where

$$\Pr_{\Phi\sim\mathcal{M}(p_0,n)}[\,P(\Phi)=1\,] = \varepsilon,$$
$$\Pr_{\Phi\sim\mathcal{M}(p_1,n)}[\,P(\Phi)=1\,] = 1-\varepsilon, \text{ and}$$
$$\Pr_{\Phi\sim\mathcal{M}(p_{1/2},n)}[\,P(\Phi)=1\,] = 1/2.$$

◀

*Proof.* First, we are going to show that the first statement implies the second. Thus, we fix some $\varepsilon$ for the second statement. We want to show that for every constant $\varepsilon_l \in (0,1)$ there is some $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ it holds that

$\frac{p_1-p_0}{p_{1/2}} < \varepsilon_l$. From our premise, we know that for all sufficiently large $n$

$$\Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] \leq \varepsilon$$

if $p' = (1-\varepsilon_l/4)\cdot p^\star$ and

$$\Pr_{\Phi\sim\mathcal{M}(p',n)}[\,P(\Phi)=1\,] \geq 1-\varepsilon$$

if $p' = (1+\varepsilon_l/4)\cdot p^\star$. For those large enough values of $n$ it holds that $p_0 \geq (1-\varepsilon_l/4)\cdot p^\star$ and $p_1 \leq (1+\varepsilon_l/4)\cdot p^\star$, since the probability function is non-decreasing. With the same argument, it holds that $p_{1/2} > 1/2\cdot p^\star$ for large enough values of $n$. Thus, for those large enough values, we get

$$\frac{p_1-p_0}{p_{1/2}} \leq \frac{(1+\varepsilon_l/4)\cdot p^\star - (1-\varepsilon_l/4)\cdot p^\star}{1/2\cdot p^\star} = \varepsilon_l.$$

Since for each constant $\varepsilon_l \in (0,1)$ there is some $n_0 \in \mathbb{N}$ so that this holds for all $n \geq n_0$, we get $\lim_{n\to\infty}\frac{p_1-p_0}{p_{1/2}} = 0$.

Now we are going to show that the second statement also implies the first one. We use $p^\star = p_{1/2}$ and fix an $\varepsilon_p \in (0,1)$. We want to show that

$$\lim_{n\to\infty}\Pr_{\Phi\sim\mathcal{M}((1-\varepsilon_p)\cdot p^\star,n)}[\,P(\Phi)=1\,] = 0.$$

In other words, we want that for any $\varepsilon_l \in (0,1)$, there is some $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ we get

$$\Pr_{\Phi\sim\mathcal{M}((1-\varepsilon_p)\cdot p^\star,n)}[\,P(\Phi)=1\,] \leq \varepsilon_l.$$

Let us now assume that $p_0$ is the parameter value at which the probability is $\varepsilon_l$ and $p_1$ is the parameter value at which the probability is $1-\varepsilon_l$. Then, for any sufficiently large $n$, we have $\frac{p_1-p_0}{p_{1/2}} \leq \varepsilon_p$. Furthermore, $p_0 \leq p_{1/2} \leq p_1$. Thus,

$$(1-\varepsilon_p)\cdot p_{1/2} = p_{1/2} - \varepsilon_p\cdot p_{1/2} \leq p_1 - (p_1-p_0) = p_0.$$

Since the probability function is non-decreasing, the probability at $(1-\varepsilon_p)\cdot p_{1/2}$ must be smaller than the one at $p_0$ and thus smaller than $\varepsilon_l$ as desired. Again, for any choice of $\varepsilon_l$ we can find an $n_0 \in \mathbb{N}$ so that this holds. This implies

$$\lim_{n\to\infty}\Pr_{\Phi\sim\mathcal{M}((1-\varepsilon_p)\cdot p^\star,n)}[\,P(\Phi)=1\,] = 0.$$

With a similar argument, we can show that for any constant $\varepsilon_p > 0$ the parameter function $p' = (1+\varepsilon_p)\cdot p^\star$ satisfies

$$\lim_{n\to\infty}\Pr_{\Phi\sim\mathcal{M}((1+\varepsilon_p)\cdot p^\star,n)}[\,P(\Phi)=1\,] = 1.$$

This proves the equivalency of both statements. ■

## 6.3 The Sharp Threshold Theorem

In this section we present Bourgain's Sharp Threshold Theorem and the concepts it relies on. Colloquially speaking the theorem states that "All monotone graph properties with a coarse threshold may be approximated by a local property." In the context of Boolean satisfiability, this means that a coarse threshold implies the existence of a family of only a few short clauses, which certify unsatisfiability. These clauses have a high probability of appearing around the threshold and their existence increases the probability of a formula to be unsatisfiable.

In the remainder of this chapter, we will use the notation from O'Donnell's book [ODo14]. This makes the application of the Sharp Threshold Theorem easier. The Sharp Threshold Theorem assumes a product probability space

$$(\Omega, \pi) = \Big(\{-1,1\}^N, \pi_1 \times \pi_2 \times \ldots \times \pi_N\Big).$$

We can now encode formulas in $k$-CNF as vectors $x \in \{-1,1\}^N$, where $N = \binom{n}{k} \cdot 2^k$ is the number of different $k$-clauses over $n$ variables. If a clause is chosen to be in the formula, we set its variable to $-1$, otherwise we set it to 1. With this encoding of formulas in $k$-CNF in mind, we can define a function $f\colon \{-1,1\}^N \to \{-1,1\}$, which returns $-1$ if the encoded formula is unsatisfiable and 1 otherwise. It is easy to see that $f$ is monotone in the sense that $f(x) \leq f(y)$ whenever $x \leq y$ coordinate-wise. This is the case, since setting a coordinate from $-1$ to 1 is equivalent to removing a clause from the encoded formula. By doing so, a satisfiable formula cannot be made unsatisfiable, i. e. the value of $f$ can only change from $-1$ to 1, but not the other way around. In a similar way, any monotone property can be described as a function $f\colon \{-1,1\}^N \to \{-1,1\}$.

We can now formally describe the product probability space of the non-uniform clause-flipping model $\mathcal{F}^N(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},s)$ with this notation. Given a variable probability distribution $\vec{p}^{(n)} = ({p_i}^{(n)})_{i=1,\ldots,n}$, the derived clause probability distribution $\vec{q}^{(n)} = ({q_i}^{(n)})_{i=1,\ldots,N}$, and the scaling factor $s$, we define our product probability space to be $(\Omega,\pi) = (\{-1,1\}^N, \pi_1 \times \pi_2 \times \ldots \times \pi_N)$ with $\pi_i(-1) = {q_i}^{(n)}(s)$ and $\pi_i(1) = 1 - {q_i}^{(n)}(s)$ for $i = 1,2,\ldots,N$. Here, ${q_i}^{(n)}(s) = \min(s \cdot {q_i}^{(n)}, 1)$ as described in Definition 3.7. We let $\mu_{\vec{p}^{(n)},s}$ denote the product probability measure, i.e. for $x \in \Omega$

$$\mu_{\vec{p}^{(n)},s}(x) = \prod_{i=1}^{N} \pi_i(x_i) = \prod_{i\in[N]\colon x_i=-1} q_i^{(n)}(s) \prod_{i\in[N]\colon x_i=1} \Big(1 - q_i^{(n)}(s)\Big).$$

For $S \subseteq \Omega$ we define $\mu_{\vec{p}^{(n)},s}(S) = \sum_{x\in S} \mu_{\vec{p}^{(n)},s}(x)$. We will use the shorthand notation $\mu$ instead of $\mu_{\vec{p}^{(n)},s}$ if the probability measure is clear from context. For a property $P$ we will also write $\mu(P) = \{x \in \Omega \mid P(x)\} = \{x \in \Omega \mid f(x) = -1\}$ if $f$

is the characteristic function of $P$, i. e. $f(x) = -1$ iff $P(x)$ holds. Furthermore, for an $N$-element vector $x = (x_1, x_2, \ldots, x_N)$ and a subset $T \subseteq [N]$ let $x_T = (x_i)_{i \in T}$ denote the *restriction of* $x$ *to* $T$. Last, for $x \in \{-1, 1\}^N$ we let $|x|_{-1}$ and $|x|_1$ denote the number of elements with value $-1$ and 1, respectively.

The following statement shows the relation between coarseness of a property's threshold and the derivative of its probability function. It says that, if a threshold is coarse, then the derivative of the probability function must be small at some point around the threshold. This is intuitively true, since a coarse threshold means that the probability function increases slowly around the threshold. It is easy to see that this derivative exists, because the probability $\mu_s(x)$ for each $x \in \Omega$ is a polynomial in $s$ and so is the probability $\mu_s(P)$ that property $P$ holds. The uniform equivalent of the following statement holds due to Friedgut [Fri99], but a simple argument shows that it also holds in the non-uniform case.

▶ **Lemma 6.7.** If a threshold for a property $P$ is coarse, then there are constants $K > 0$ and $\varepsilon \in (0, 1)$ such that for infinitely many $n \in \mathbb{N}$ there is a point $s^\star$ such that $\mu_{s^\star}(P) \in (\varepsilon, 1 - \varepsilon)$ and $s^\star \cdot \frac{d\mu_s(P)}{ds}|_{s=s^\star} \leq K$. ◀

*Proof.* The proof of the statement is a simple application of the mean value theorem. We know that the existence of a coarse threshold implies that there are constants $\varepsilon \in (0, 1/2)$ and $\varepsilon_s > 0$ and parameter functions $s_0$ and $s_1 = (1 + \varepsilon_s) \cdot s_0$ such that for infinitely many values of $n$ the probabilities between $s_0$ and $s_1$ are between $\varepsilon$ and $1 - \varepsilon$ (c. f. Lemma 6.6). For each such $n$ the mean value theorem now implies that there is a point $s^\star$ such that

$$\left.\frac{d\mu_s(P)}{ds}\right|_{s=s^\star} = \frac{\mu_{s_1}(P) - \mu_{s_0}(P)}{s_1 - s_0} \leq \frac{1 - 2 \cdot \varepsilon}{\varepsilon_s \cdot s_0}.$$

Since $s^\star \leq s_1 = (1 + \varepsilon_s) \cdot s_0$, this yields

$$s^\star \cdot \left.\frac{d\mu_s(P)}{ds}\right|_{s=s^\star} \leq \frac{(1 - 2 \cdot \varepsilon) \cdot (1 + \varepsilon_s)}{\varepsilon_s} = K$$

as desired. ■

Note that the point $s^\star$ usually depends on the value of $n \in \mathbb{N}$, i. e. $s^\star$ is actually a partial function on $\mathbb{N}$. Furthermore, the condition $\mu_s^\star(P) \in (\varepsilon, 1 - \varepsilon)$ ensures that $s^\star$ is in the same range as the asymptotic threshold function, i. e. "around the threshold". This will be crucial for showing our result with the Sharp Threshold Theorem.

Bourgain's Sharp Threshold Theorem will make use of the total influence of a Boolean function $f$. Intuitively, the *influence* $\mathbf{Inf}_i[f]$ of a function $f$ describes the probability that the value of the $i$-th coordinate influences the function value. The *total influence* $\mathbf{I}[f]$ of a function $f$ is the sum of the influence values for all coordinates. Both, $\mathbf{Inf}_i[f]$ and $\mathbf{I}[f]$ depend on the probability distribution $\pi$ and on the scaling parameter $s$, but we will omit this dependence if it is clear

from context. The following definition from [ODo14] formalizes our intuitive one.

▶ **Definition 6.8.** [Influence Function] Let $f\colon \{-1,1\}^N \to \{-1,1\}$ and let $(\Omega,\pi)$ be our product probability space with $\Omega = \{-1,1\}^N$ and $\pi = \pi_1 \times \ldots \times \pi_N$. The *influence* of the $i$-th coordinate is $\mathbf{Inf}_i[f] = \mathbb{E}_{x\sim\pi}[\, f(x)(L_i f)(x)\,]$,where $L_i f = f - E_i f$ and $E_i f(y) = \mathbb{E}_{\boldsymbol{y_i}\sim\pi_i}[\, f(y_1, y_2, \ldots, y_{i-1}, \boldsymbol{y_i}, y_{i+1} \ldots, y_{N-1}, y_N)\,]$. The *total influence* of $f$ is $\mathbf{I}[\, f\,] = \sum_{i=1}^{n} \mathbf{Inf}_i[f]$. ◀

The following lemma relates this notion of influence to the notion of coarseness due to Friedgut, more precisely to

$$\frac{d\mu_s(P)}{ds}s = \frac{d\mu_s(\{x \in \Omega \mid f(x) = -1\})}{ds}s,$$

where $\mu_s$ denotes the product probability measure of the clause-flipping non-uniform random $k$-SAT model $\mathcal{F}^N(n,k,(\vec{p}^{\,(n)})_{n\in\mathbb{N}},s)$.

▶ **Lemma 6.9.** Let $f\colon \{-1,1\}^N \to \{-1,1\}$ be monotone, and non-constant and let $(\Omega,\pi)$ be our product probability space with $\Omega = \{-1,1\}^N$ and $\pi = \pi_1 \times \ldots \times \pi_N$. Given clause probabilities $\vec{q}^{\,(n)} = (q_i{}^{(n)})_{i=1,\ldots,N}$, let $\pi_i(-1) = q_i{}^{(n)}(s)$ and $\pi_i(1) = 1 - q_i{}^{(n)}(s)$ for $i = 1,2,\ldots,N$, where $q_i{}^{(n)}(s) = \min\big(s \cdot q_i{}^{(n)}, 1\big)$. Further, let $P = \{x \in \Omega \mid f(x) = -1\}$. For $s < q_{\max}^{-1}$ it holds that

$$\mathbf{I}[\, f\,] \leq 4 \cdot \frac{d\mu_s(P)}{ds}s.$$

◀

*Proof.* Due to the requirement $s < q_{\max}^{-1}$ we can assume $q_i(s) = s \cdot q_i$ instead of $q_i(s) = \min(s \cdot q_i, 1)$. First, we are going to show that

$$\mathbf{I}[\, f\,] = 4 \sum_{x\in\Omega\colon f(x)=-1} \left( \mu_s(x) \cdot \sum_{i\in[N]\colon f(x\oplus i)=1} (1 - s \cdot q_i) \right).$$

Here, $x \oplus i$ denotes the encoding $x$ in which the $i$-th coordinate is flipped. Then, we will show

$$\frac{d\mu_s(P)}{ds}\Big|_{s=s'} = \frac{1}{s'} \sum_{x\in\Omega\colon f(x)=-1} \mu_{s'}(x) \cdot \sum_{i\in[N]\colon f(x\oplus i)=1} 1.$$

Together, this implies

$$\mathbf{I}[\, f\,] \leq 4 \cdot \frac{d\mu_s(P)}{ds}s$$

as desired.

We note that

$$\mathbf{Inf}_i[f] = \sum_{x\in\Omega} \mu_s(x) f(x)(f(x) - E_i f(x)).$$

As $E_i f$ only rerandomizes the $i$-th coordinate, it holds that $E_i f(x) = f(x)$ whenever $f(x) = f(x \oplus i)$. Remember that $x \oplus i$ denotes the encoding $x'$ where the $i$-th coordinate is flipped. The contribution of these $x \in \Omega$ is therefore zero and we can concentrate on the case that $f(x) \neq f(x \oplus i)$. As $f$ is monotone, it can only hold that $f(x) = -1$ if $y_i = -1$ and $f(x) = 1$ if $y_i = 1$. Thus,

$$E_i f(x) = (1 - s \cdot q_i) \cdot 1 + s \cdot q_i \cdot (-1) = 1 - 2s \cdot q_i.$$

So, for an $x$ with $f(x) = -1$ and $f(x \oplus i) = 1$, its contribution to $\mathbf{Inf}_i[f]$ is

$$\mu_s(x)(-1)(-1 - (1 - 2s \cdot q_i)) = \mu_s(x) \cdot 2(1 - s \cdot q_i) = 2s \cdot q_i \cdot \mu_s(x \oplus i).$$

The last inequality again holds since $f$ is monotone and $\mu(x)$ must therefore contain the factor $s \cdot q_i$ for $x_i = -1$. If we have an $x$ with $f(x) = 1$ and $f(x \oplus i) = -1$, we get

$$\mu_s(x)(1 - (1 - 2s \cdot q_i)) = 2s \cdot q_i \cdot \mu_s(x) = 2(1 - s \cdot q_i) \cdot \mu_s(x \oplus i).$$

So, if $f$ is not constant, the contribution of each $x \in \Omega$ with $f(x) = -1$ and $f(x \oplus i) = 1$ is counted exactly twice, once for $x$ and once for $x \oplus i$. Note that this only holds, since we consider a fixed $i \in N$. Thus,

$$\mathbf{Inf}_i[f] = 4 \cdot (1 - s \cdot q_i) \cdot \sum_{x \in \Omega \colon f(x) = -1,\ f(x \oplus i) = 1} \mu_s(x)$$

and

$$\mathbf{Inf}[f] = 4 \sum_{x \in \Omega \colon f(x) = -1} \left( \mu_s(x) \cdot \sum_{i \in [N] \colon f(x \oplus i) = 1} (1 - s \cdot q_i) \right).$$

Now we turn to the second statement. For a certain $x = (x_1, x_2 \ldots, x_N) \in \{-1, 1\}^N$ it holds that

$$\frac{d\mu_s(x)}{ds} = \frac{d\left( \left( \prod_{i \in [N] \colon x_i = -1} s \cdot q_i \right) \cdot \left( \prod_{i \in [N] \colon x_i = 1} (1 - s \cdot q_i) \right) \right)}{ds}.$$

We split this expression into two parts due to the product rule for derivatives:

$$\begin{aligned} \frac{d\mu_s(x)}{ds} = & \frac{d\left( \prod_{i \in [N] \colon x_i = -1} s \cdot q_i \right)}{ds} \cdot \left( \prod_{i \in [N] \colon x_i = 1} (1 - s \cdot q_i) \right) + \\ & + \left( \prod_{i \in [N] \colon x_i = -1} s \cdot q_i \right) \cdot \frac{d\left( \prod_{i \in [N] \colon x_i = 1} (1 - s \cdot q_i) \right)}{ds}. \end{aligned} \tag{6.8}$$

It holds that

$$\begin{aligned}\frac{d\left(\prod_{i\in[N]\colon x_i=-1} s\cdot q_i\right)}{ds} &= \frac{d\left(s^{|x|_{-1}}\prod_{i\in[N]\colon x_i=-1} q_i\right)}{ds}\\ &= \frac{|x|_{-1}\cdot s^{|x|_{-1}}}{s}\cdot\prod_{i\in[N]\colon x_i=-1} q_i\\ &= \frac{|x|_{-1}}{s}\cdot\prod_{i\in[N]\colon x_i=-1} s\cdot q_i,\end{aligned}$$

where $|x|_a$ denotes the number of appearances of $a$ in the vector $x$. It now holds that the first term of equation (6.8) is simply $(|x|_{-1}/s)\cdot\mu_s(x)$. Now let $i_1, i_2, \ldots, i_{|x|_1}$ be the indices of coordinates with value 1 in $x$. For the derivative in the second term, we can use the product rule again to obtain

$$\begin{aligned}\frac{d\left(\prod_{i\in[N]\colon x_i=1}(1-s\cdot q_i)\right)}{ds} = -\frac{1}{s}\cdot\frac{s\cdot q_{i_1}}{1-s\cdot q_{i_1}}\cdot\left(\prod_{i\in[N]\colon x_i=1}(1-s\cdot q_i)\right)+\\ +(1-s\cdot q_{i_1})\cdot\frac{d\left(\prod_{i\in[N]\setminus\{i_1\}\colon x_i=1}(1-s\cdot q_i)\right)}{ds}.\end{aligned}$$

By repeatedly using the product rule, we get

$$\frac{d\left(\prod_{i\in[N]\colon x_i=1}(1-s\cdot q_i)\right)}{ds} = -\frac{1}{s}\left(\sum_{i\in[N]\colon x_i=1}\frac{s\cdot q_i}{1-s\cdot q_i}\right)\cdot\left(\prod_{i\in[N]\colon x_i=1}(1-s\cdot q_i)\right),$$

i.e. the second term of equation (6.8) is

$$-\frac{1}{s}\left(\sum_{i\in[N]\colon x_i=1}\frac{s\cdot q_i}{1-s\cdot q_i}\right)\cdot\mu_s(x).$$

It now holds that

$$\frac{d\mu_s(x)}{ds} = \frac{1}{s}\left(|x|_{-1}-\sum_{i\in[N]\colon x_i=1}\frac{s\cdot q_i}{1-s\cdot q_i}\right)\cdot\mu_s(x). \tag{6.9}$$

Summing equation (6.9) over all $x \in \{-1,1\}^N$ with $f(x) = -1$ gives

$$\sum_{x\in\Omega\colon f(x)=-1} \frac{d\mu_s(x)}{ds} = \frac{1}{s} \sum_{x\in\Omega\colon f(x)=-1} \left( \left( |x|_{-1} - \sum_{i\in[N]\colon x_i=1} \frac{s\cdot q_i}{1-s\cdot q_i} \right) \cdot \mu_s(x) \right).$$

We know that $\frac{s\cdot q_i}{1-s\cdot q_i}\mu_s(x) = \mu_s(x')$ for a vector $x' \in \{-1,1\}^N$ which is the same as $x$ except for the $i$-th coordinate which is set from 1 to $-1$. As $f$ is monotone, it must hold that $f(x') = -1$. Now we can count how often each $\mu_s(x)$ appears in this sum. Each $\mu_s(x)$ is added once for each coordinate with $x_i = -1$ and subtracted once for each $\hat{x}$ with $f(\hat{x}) = -1$, where $\hat{x}$ is the same as $x$ except for one coordinate which is set from $-1$ in $x$ to 1 in $\hat{x}$. So the total number of times that a $\mu_s(x)$ with $f(x) = -1$ remains is $|\{i \in [1,N] |\, x_i = -1, f(x \oplus i) = 1\}| = |\{i \in [1,N] |\, f(x \oplus i) = 1\}|$ as $f$ is monotone. This yields

$$\left.\frac{d\mu_s(P)}{ds}\right|_{s=s'} = \frac{1}{s'} \sum_{x\in\Omega\colon f(x)=-1} \left( \mu_{s'}(x) \cdot \sum_{i\in[N]\,\colon f(x\oplus i)=1} 1 \right)$$

and gives us the result as desired. ■

To prove our main theorem, we will use the Sharp Threshold Theorem by Friedgut (and Bourgain) [Fri99] in O'Donnell's version [ODo14]. The theorem states that, if a monotone property $P$ has a coarse threshold, and therefore small influence, then there are local structures which approximate this property. These local structures are called $\tau$-boosters. Intuitively, a $\tau$-booster is a prescription for the existence and/or absence of certain clauses in a random formula such that conditioning on this prescription increases or decreases the probability for $P$ to hold by at least $\tau/2$ (or its expected value by $\tau$). The following is a formal definition of these structures.

▶ **Definition 6.10.** [$\tau$-booster] Let $f\colon \{-1,1\}^N \to \{-1,1\}$ and let $(\Omega,\pi)$ be a product probability space with $\Omega = \{-1,1\}^N$ and $\pi = \pi_1 \times \ldots \times \pi_N$. For $T \subseteq [N]$, $y \in \Omega$, and $\tau > 0$, we say that the restriction $y_T$ is a *$\tau$-booster* if $\mathbb{E}_{x\sim\pi}[\, f \mid x_T = y_T \,] \geq \mathbb{E}[\, f \,] + \tau$. If $\tau < 0$, we say that $y_T$ is a $\tau$-booster if $\mathbb{E}_{x\sim\pi}[\, f \mid x_T = y_T \,] \leq \mathbb{E}[\, f \,] - |\tau|$. ◀

Note that it depends on the probabilities $\pi_i$ and thus on the scaling factor $s$ if a restriction $y_T$ is a $\tau$-booster. However, we will omit this dependence if it is clear from context.

The Sharp Threshold Theorem is stated as follows:

▶ **Theorem 6.11.** [Bourgain's Sharp Threshold Theorem] Let $f\colon \{-1,1\}^N \to \{-1,1\}$ and let $(\Omega,\pi)$ be our product probability space with $\Omega = \{-1,1\}^N$ and $\pi = \pi_1 \times \ldots \times \pi_N$. If $\mathbf{I}[\, f \,] \leq K$ for a constant $K$, then there is some $\tau$ (either

negative or positive) with $|\tau| \geq \mathrm{Var}[f] \cdot \exp(-O(\mathbf{I}[f]^2/\mathrm{Var}[f]^2))$ such that

$$\Pr_{x\sim\pi}\left[\exists\, T \subseteq [n],\ |T| \in O\left(\frac{\mathbf{I}[f]}{\mathrm{Var}[f]}\right) \text{ such that } x_T \text{ is a } \tau\text{-booster}\right] \geq |\tau|.$$

◀

This theorem is not specific to probability spaces with uniform probability distributions. Due to O'Donnell the Sharp Threshold Theorem also holds for arbitrary product probability spaces. Müller [Mül17] also showed that a version of Bourgain's original theorem still holds for arbitrary product probability spaces. Furthermore, by carefully checking the proof of the theorem, one can see that the constants hidden in the $O$-notation do not depend on the product probability space $(\Omega, \pi)$. O'Donnell also states this in the arXiv version of his book [ODo21]. In its original form, the theorem additionally requires that $\mathrm{Var}[f] \geq 0.01$. However, one can see from the proofs that any constant lower bound on $\mathrm{Var}[f]$ suffices. This is the case in the setting we consider, i. e. assuming a coarse threshold for unsatisfiability. $\mathrm{Var}[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2 = 1 - \mathbb{E}[f]^2$ is bounded away from zero and one by constants, since $\mathbb{E}[f] = -\mu_{s^\star}(P)+1-\mu_{s^\star}(P) = 1-2\cdot\mu_{s^\star}(P)$ and $\mu_{s^\star}(P)$ is bounded away from zero and one by constants due to Lemma 6.7. Together with the prerequisite that $\mathbf{I}[f]$ is upper-bounded by a constant $K$, the theorem essentially says that $|\tau|$ is at least some constant, while $|T|$ is at most some constant.

## 6.4 Proof of Sharpness

This section will be dedicated to proving the following theorem.

▶ **Theorem 6.12.** Let $k \geq 3$, let $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ be an ensemble of probability distributions on $n$ variables each and let $s^\star$ be an asymptotic satisfiability threshold for $\mathcal{F}(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s)$ with respect to $s$. If $p_{\max} \in o(s^{\star-(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(s^\star))$, then the threshold for satisfiability on $\mathcal{F}(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s)$ with respect to $s$ is sharp. ◀

The framework to prove this statement is inspired by the seminal work of Friedgut [Fri99]. It's high level idea and the structure of this section are as follows.

We assume toward a contradiction that the threshold is coarse. Then the Sharp Threshold Theorem tells us that there have to be $\tau$-boosters of constant size that appear with constant probability in the random formula. These boosters have the property that conditioning on their existence *boosts* the probability of the random formula to be unsatisfiable by at least an additive constant.

One kind of booster are unsatisfiable subformulas of constant size. Conditioning on them would boost the probability to generate an unsatisfiable formula to one. We rule these out by showing that they do not appear with constant probability.

Then, we consider subformulas, which give the second highest boost: maximally quasi-unsatisfiable subformulas. These are subformulas which have only *one* satisfying assignment for the variables appearing in them and adding any new clause over those variables makes them unsatisfiable. We want to show that these cannot boost the probability of a formula to be unsatisfiable by a constant.

Again toward a contradiction, we assume that conditioning on a maximally quasi-unsatisfiable subformula $T$ is enough to boost the unsatisfiability probability by a constant. First, we prove that conditioning on $T$ is equivalent to adding a number of clauses of size shorter than $k$ to the random formula over variables not appearing in $T$. Then, we use a version of Friedgut's coverability lemma to show that, if adding these clauses of size smaller than $k$ makes the random formula unsatisfiable with constant probability, then so does adding $o(\sqrt{s^\star})$ clauses of size $k$. We prove that this probability is dominated by the probability to make the original random formula unsatisfiable for a slightly bigger scaling factor. However, we can show that the probability to make the original random formula unsatisfiable cannot be increased by a constant with this slightly increased scaling factor. This contradicts our assumption that the probability is boosted by a constant in the first place. Therefore, quasi-unsatisfiable subformulas cannot be boosters.

After showing this, every less restrictive subformula cannot be a booster either. That means, the only possible boosters are unsatisfiable subformulas, which we ruled out already. Therefore, the implication of the Sharp Threshold Theorem does not hold, which contradicts the assumption of a coarse threshold.

**Application of the Sharp Threshold Theorem** First, note that any asymptotic threshold function $s^\star$ must satisfy $s^\star \in \Omega(1)$. This is simply due to the fact that our model is defined in such a way that the scaling factor $s$ is the expected number of clauses we flip and any unsatisfiable formula in $k$-CNF needs at least $2^k$ different clauses. Thus,

$$\Pr_{\Phi\sim\mathcal{F}^N(s)}[\,\Phi \text{ unsat}\,] \leq \Pr_{\Phi\sim\mathcal{F}^N(s)}\left[\,|\Phi| \geq 2^k\,\right] \leq \frac{\mathbb{E}[\,|\Phi|\,]}{2^k} = \frac{s}{2^k}$$

due to Markov's inequality. This means, for any constant $s < 2^k$, the probability to generate an unsatisfiable instance is bounded away from one by a constant, which implies $s^\star \in \Omega(1)$. That implies

$$C_k \leq \left(1 - \frac{k^2}{2}\sum_{i=1}^{n} p_i^2\right)^{-1} = 1 + O\left(s^{\star -2/k}\right)$$

due to equation (5.2) as well as $\sum_{i=1}^n p_i^2 \leq p_{\max} \in o\left(s^{\star -\frac{3k-1}{4k-2}}\right) \in o\left(s^{\star -2/k}\right)$.

We know that there is an asymptotic threshold function $s^\star$ and we assume toward a contradiction that the threshold is coarse. Due to our definition of coarse thresholds and their implications (see Section 6.2) this means that there are

infinitely many values $n \in \mathbb{N}$ for which the probability function behaves as we intuitively imagine it: We have a range of $s$ of size $\Theta(s^\star)$ where the probability to generate unsatisfiable instances is bounded away from zero and one by constants, i. e. the probability function slowly increases in that range. More formally, it means we can define incomplete functions $s_0$, $s_1$ such that there are constants $\varepsilon > 0$, $0 < \varepsilon_0 < \varepsilon_1 < 1$ and $n_0 \in \mathbb{N}$ so that for all those infinitely many values of $n \geq n_0$ it holds that $s_1 - s_0 \geq \varepsilon \cdot s^\star$ as well as $\Pr_{\Phi \sim \mathcal{F}^N(s_0)}[\,\Phi \text{ unsat}\,] \geq \varepsilon_0$ and $\Pr_{\Phi \sim \mathcal{F}^N(s_1)}[\,\Phi \text{ unsat}\,] \leq \varepsilon_1$. From now on we will concentrate on this subset of the natural numbers. If we use asymptotic expressions, they will also only hold for this subset if not stated otherwise. This will be enough to derive a contradiction for some sufficiently large value of $n$ for which this property holds.

Due to Lemma 6.7 a coarse threshold implies

$$\frac{d\mu_s(P)}{ds} s \leq K$$

for some constant $K$ and some $s$ in the threshold interval. Let us call this scaling factor $s_c$. Note that $s_c = \Theta(s^\star)$, since $s_c$ is in the threshold interval and $s^\star$ is an asymptotic threshold function. Due to Lemma 6.9 this means $\mathbf{I}[\,f\,] \leq 4 \cdot K$ for the indicator function $f$ with $P = \{x \in \Omega \mid f(x) = -1\}$. For the corollary to hold, we have to assure $s_c < q_{\max}^{-1}$. This follows due to our assumption

$$p_{\max} \in o\left(s^{\star -\frac{3k-1}{4k-2}}\right) \in o\left(s^{\star -1/k}\right),$$

which implies

$$q_{\max}(s_c) = s_c \cdot q_{\max} = s_c \cdot O\left(p_{\max}^k\right) \in o(1). \tag{6.10}$$

Remember that $\mu_{s_c}(f)$ is constant and so are $\mathbb{E}[\,f\,]$ and $\mathrm{Var}[f]$ at this scaling factor.

Now we can use Theorem 6.12 to see that, at least with constant probability $\tau$, our formulas have a subformula (or lack thereof) consisting of at most $O(K) \in O(1)$ clauses, so that conditioning on the existence (or non-existence) of these clauses increases (or decreases) the probability that our random formulas in k-CNF are unsatisfiable by at least $\tau/2$. The subformulas with these properties are the boosters. The theorem actually allows us to choose appropriate specific constants for $\tau$ and the upper bound on $|T|$ that are independent of $n$. That means, throughout the proof we can assume that $\tau$ and $|T|$ are given fixed constants.

Since the property of being unsatisfiable is monotone, it would not be beneficial to forbid some clauses and demand others. We can therefore concentrate on the two cases of either only forbidding or only enforcing clauses in our boosters. The following lemma shows that it suffices to concentrate on enforcing boosters. The idea is that every constant-sized subset of clauses a. a. s. does not exist in the formula, since clause probabilities are $o(1)$. Therefore, conditioning on the

non-existence of such a subformula does not change the overall probability by too much.

▶ **Lemma 6.13.** Every constant-sized booster which assumes the non-existence of clauses only boosts the probability to be satisfiable or unsatisfiable by $o(1)$. ◀

*Proof.* Suppose we have an $x \in \{-1,1\}^N$ drawn from $(\Omega, \pi)$ and a set $T \subseteq [N]$ of clause indices, where $|T| = v$ is constant. It now holds that

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\pi}\left[ f \mid x_T = 1^{|T|} \right] &= \Pr\left[f(x) = 1 \mid x_T = 1^{|T|}\right] - \Pr\left[f(x) = -1 \mid x_T = 1^{|T|}\right] \\ &= 2 \cdot \Pr\left[f(x) = 1 \mid x_T = 1^{|T|}\right] - 1. \end{aligned}$$

Furthermore

$$\begin{aligned} \Pr\left[f(x) = 1 \mid x_T = 1^{|T|}\right] &= \frac{\Pr\left[f(x) = 1 \wedge x_T = 1^{|T|}\right]}{\Pr\left[x_T = 1^{|T|}\right]} \\ &\leq \frac{\Pr[f(x) = 1]}{\Pr\left[x_T = 1^{|T|}\right]} \\ &= \frac{1 - \mu_{s_c}(P)}{\Pr\left[x_T = 1^{|T|}\right]}, \end{aligned}$$

where the last equality holds because we are at the critical scaling factor $s_c$ for the property $P$ of having an unsatisfiable formula. Remembering that $\Pr\left[x_T = 1^{|T|}\right]$ is the probability that none of the clauses with indices from $T$ appear, we get

$$\begin{aligned} 1 - \Pr\left[x_T = 1^{|T|}\right] &= \Pr[\exists\, i \in T :\ x_i = -1] \\ &\leq \sum_{i \in T} q_i(s_c) \\ &\leq |T| \cdot q_{\max}(s_c) \\ &\in o(1). \end{aligned}$$

Here, the last line follows due to equation (6.10). If we plug this into our first equation, we get

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\pi}\left[ f \mid x_T = 1^{|T|} \right] &\leq \frac{2 - 2 \cdot \mu_{s_c}(P)}{1 - o(1)} - 1 \\ &= 1 - 2 \cdot \mu_{s_c}(P) + o(1) \\ &= \mathop{\mathbb{E}}_{x\sim\pi}[\, f \,] + o(1). \end{aligned}$$

Equivalently, we can show that

$$\mathop{\mathbb{E}}_{x\sim\pi}\left[ f \mid x_T = 1^{|T|} \right] = 1 - 2 \cdot \Pr\left[f(x) = -1 \mid x_T = 1^{|T|}\right]$$

$$\geq 1 - 2 \cdot \frac{\Pr[f(x) = -1]}{\Pr\left[x_T = 1^{|T|}\right]}$$
$$\geq 1 - 2 \cdot \frac{\mu_{s_c}(P)}{1 - o(1)}$$
$$= \mathop{\mathbb{E}}_{x \sim \pi}[f] - o(1).$$

This means, the set $T$ cannot be a $\tau$-booster for any constant $\tau$. ■

We can now concentrate on conditioning on the *existence* of clauses. Our goal is to show that no constant-sized $\tau$-boosters exist with constant probability.

**Unsatisfiable subformulas are too improbable** A sure way to boost the probability of being unsatisfiable to one is to condition on the existence of an unsatisfiable subformula. To rule this case out, the next lemma shows that the probability that our formulas have an unsatisfiable subformula of constant size is smaller than any constant $\tau$ for sufficiently large $n$. The proof essentially shows that any minimally unsatisfiable subformula of constant size cannot exist with constant probability. This can be seen from the fact that such subformulas contain each variable in them at least twice and the probability for this can be bounded using $\sum_{i=1}^{n} p_i^2$ and $p_{\max}$.

▶ **Lemma 6.14.** Let $a, k \in \mathbb{N}$ be constants and let $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ be an ensemble of variable probability distributions. If $p_{\max} \in o(s^{\star -1/k})$ and $\sum_{i=1}^{n} p_i^2 \in O(s^{\star -2/k})$, then a random formula from $\mathcal{F}(n, k, (\vec{p}_n)_{n\in\mathbb{N}}, s)$ with $s \in O(s^\star)$ has an unsatisfiable subformula of length at most $a$ with probability $o(1)$. ◀

*Proof.* Before we can state this result, we have to make some observations. First, if a formula is unsatisfiable, it also contains a minimal unsatisfiable subformula, i.e. an unsatisfiable subformula such that removing any clause from it would make it satisfiable. Second, in a minimal unsatisfiable formula each variable has to appear at least twice. Otherwise there would be a pure literal and the clause with this literal could be satisfied and eliminated from the formula, independently of all other variables. The formula would therefore not be *minimally* unsatisfiable. Third, a result by Aharoni and Linial [AL86] states that each unsatisfiable formula over $v$ variables consists of at least $v + 1$ clauses. Fourth, each subformula $T$ of constant length $a$ in $k$-CNF consists of $a \cdot k$ literals, and, hence, also of at most $a \cdot k$ variables.

For a constant $v$ let $\mathcal{T}^{(v)}$ be the set of all formulas over $v$ variables with at least $v + 1$ and at most $a$ clauses in which each variable appears at least twice and let $\mathcal{T} = \bigcup_{v<a} \mathcal{T}^{(v)}$. Now let $\Phi$ be a random formula drawn from $\mathcal{F}\left(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s\right)$. We use the notation $A \subseteq B$ to denote that $A$ is a subformula of $B$. Using a union bound we get

$$\Pr[\exists\, T \subseteq \Phi \colon T \text{ unsat} \wedge |T| \leq a] \leq \Pr[\exists\, T \in \mathcal{T} \colon T \subseteq \Phi]$$

$$\leq \sum_{v<a} \Pr\left[\exists\, T \in \mathcal{T}^{(v)} : T \subseteq \Phi\right].$$

Now we can concentrate on bounding $\Pr\left[\exists\, T \in \mathcal{T}^{(v)} : T \subseteq \Phi\right]$. For $v+1 \leq l \leq a$ let $\mathcal{T}_l^{(v)}$ be the subset of $\mathcal{T}^{(v)}$ containing only formulas of length $l$. It holds that

$$\Pr\left[\exists\, T \in \mathcal{T}^{(v)} : T \subseteq \Phi\right] \leq \sum_{l=v+1}^{a} \Pr\left[\exists\, T \in \mathcal{T}_l^{(v)} : T \subseteq \Phi\right].$$

We can see that

$$\Pr\left[\exists\, T \in \mathcal{T}_l^{(v)} : T \subseteq \Phi\right] = \sum_{S \in \mathcal{P}_v([n])} \sum_{F \in C_l(S)} \prod_{c \in F} q_c(s),$$

where $C_l(S)$ is the collection of all sets of clauses of size $l$ over the variables with indices in $S$ such that each variable appears at least twice. Let us take a look at a certain $S = \{j_1, j_2, \ldots, j_v\}$ and $F \in C_l(S)$. Due to equation (3.3) we have

$$\prod_{c \in F} q_c(s) = \left(s \cdot C_k \cdot \frac{k!}{2^k}\right)^l \prod_{i=1}^{v} (p_{j_i})^{m_i},$$

where $C_k = \left(1 + O\left(s^{\star -2/k}\right)\right)$ and $m_i$ is the number of appearances of the variable $X_{j_i}$ in the set of clauses $F$. Due to the definition of $C_l(S)$ each $F \in C_l(S)$ defines multiplicities $(m_1, m_2, \ldots, m_v)$ for the $v$ variables such that $m_i \geq 2$ for all $i \in [v]$ and $\sum_{i=1}^{v} m_i = k \cdot l$. This means, it holds that

$$\left(s \cdot C_k \cdot \frac{k!}{2^k}\right)^l \prod_{i=1}^{v} (p_{j_i})^{m_i} \leq \left(s \cdot C_k \cdot \frac{k!}{2^k}\right)^l \prod_{j \in S} (p_j)^2 \cdot p_{\max}^{k \cdot l - 2 \cdot v}.$$

Since there are at most $\binom{\binom{v}{k}2^k}{l}$ sets in $C_l(S)$, it holds that

$$\begin{aligned}
\Pr\left[\exists\, T \in \mathcal{T}_l^{(v)} : T \subseteq \Phi\right] &\leq \sum_{S \in \mathcal{P}_v([n])} \left(\binom{\binom{v}{k}2^k}{l} \left(s \cdot C_k \cdot \frac{k!}{2^k}\right)^l \cdot \left(\prod_{j \in S} (p_j)^2\right) \cdot p_{\max}^{k \cdot l - 2 \cdot v}\right) \\
&\leq \binom{\binom{v}{k}2^k}{l} \left(s \cdot C_k \cdot \frac{k!}{2^k}\right)^l \cdot p_{\max}^{k \cdot l - 2 \cdot v} \left(\sum_{i=1}^{n} p_i^2\right)^v \\
&\in \exp(O(s^{\star -2/k})) \cdot o\left(\cdot s^{\star l - (k \cdot l - 2 \cdot v)/k - 2v/k}\right) \in o(1),
\end{aligned}$$

where the last line follows due to our requirements $p_{\max} \in o(s^{\star -1/k})$ and $\sum_{i=1}^{n} p_i^2 \in O\left(s^{\star -2/k}\right)$, and due to $C_k = \left(1 + O\left(s^{\star -2/k}\right)\right)$. We can now conclude

that

$$\Pr[\exists\, T \subseteq \Phi \colon T \text{ unsat} \wedge |T| \leq a] \leq \sum_{v=k}^{a-1} \sum_{l=v+1}^{a} \Pr\left[\exists\, T \in \mathcal{T}_l^{(v)} \colon T \subseteq \Phi\right] \in o(1).$$

This is exactly what we wanted to show. ■

**Maximally quasi-unsatisfiable subformulas provide the second-highest boost** Since we ruled out unsatisfiable subformulas as the boosters we are looking for, we now turn our attention to satisfiable subformulas. Let $\Phi_T$ be the formula encoded by $x_T = (-1)^{|T|}$ and let $V(T) \subseteq \{X_1, \ldots, X_n\}$ be the variables in $\Phi_T$. Note that $|V(T)|$ is constant since $|T|$ is constant and each clause contains $k$ variables. We call $\Phi_T$ *maximally quasi-unsatisfiable (mqu)* if it is satisfiable by only one of the $2^{|V(T)|}$ assignments over its variable set (quasi-unsatisfiable) and if adding any new clause with variables only from $V(T)$ makes it unsatisfiable (maximally satisfiable). The following lemma formalizes a statement by Friedgut [Fri99], that the biggest possible boost any satisfiable subformula can give is achieved by mqu subformulas. The proof of the statement uses the fact that every satisfiable subformula can be extended to a mqu subformula over the same variables. It also uses positive correlation of increasing events [FKG71] and the fact that we have a product probability space.

▶ **Lemma 6.15.** For every $T \subseteq [N]$ so that $\Phi_T$ is satisfiable, there is a $T' \supseteq T$ so that $\Phi_{T'}$ is maximally quasi-unsatisfiable and

$$\Pr_{x\sim\pi}\left[\, f(x) = -1 \mid x_{T'} = (-1)^{|T'|} \,\right] \geq \Pr_{x\sim\pi}\left[\, f(x) = -1 \mid x_T = (-1)^{|T|} \,\right].$$

◀

*Proof.* First of all, note that any satisfiable formula $\Phi_T$ can be extended to a maximally quasi-unsatisfiable formula $\Phi_{T'}$ by first adding enough clauses to make it quasi-unsatisfiable and then adding clauses which do not make the resulting formula unsatisfiable as long as such clauses still exist. We now define functions $g_S(x) \colon \{-1,1\}^N \to \{-1,1\}$ for $S \subseteq [N]$ such that $g_S(x) = -1$ if $x_S = (-1)^{|S|}$ and $g_S(x) = 1$ otherwise. It is easy to see, that $g_S(x)$ is increasing (monotone) for all $S \subseteq [N]$. We can derive

$$\begin{aligned}
&\Pr_{x\sim\pi}\left[\, f(x) = -1 \mid x_{T'} = (-1)^{|T'|} \,\right] \\
&= \Pr_{x\sim\pi}\left[\, f(x) = -1 \mid g_{T'}(x) = -1 \,\right] \\
&= \frac{\Pr_{x\sim\pi}\left[\, f(x) = -1 \wedge g_{T'}(x) = -1 \,\right]}{\Pr_{x\sim\pi}\left[\, g_{T'}(x) = -1 \,\right]} \\
&= \frac{\Pr_{x\sim\pi}\left[\, f(x) = -1 \wedge g_T(x) = -1 \wedge g_{T'\setminus T}(x) = -1 \,\right]}{\Pr_{x\sim\pi}\left[\, g_{T'}(x) = -1 \,\right]}.
\end{aligned}$$

It now holds that $\{f(x) = -1\}$, $\{g_T(x) = -1\}$, and $\{g_{T'\setminus T}(x) = -1\}$ are decreasing events (monotone functions), i. e. the event that $\{f(x) = -1\}$ holds cannot increase if we increase $x$. Since the intersection of decreasing events is also decreasing, the same holds for $\{f(x) = -1 \wedge g_T(x) = -1\}$. The FKG theorem [FKG71] tells us that decreasing events are positively associated, thus

$$\begin{aligned}
&\Pr_{x\sim\pi}\left[ f(x) = -1 \mid x_{T'} = (-1)^{|T'|} \right] \\
&= \frac{\Pr_{x\sim\pi}\left[ f(x) = -1 \wedge g_T(x) = -1 \wedge g_{T'\setminus T}(x) = -1 \right]}{\Pr_{x\sim\pi}\left[ g_{T'}(x) = -1 \right]} \\
&\geq \frac{\Pr_{x\sim\pi}\left[ f(x) = -1 \wedge g_T(x) = -1 \right] \cdot \Pr_{x\sim\pi}\left[ g_{T'\setminus T}(x) = -1 \right]}{\Pr_{x\sim\pi}\left[ g_{T'}(x) = -1 \right]} \\
&= \frac{\Pr_{x\sim\pi}\left[ f(x) = -1 \wedge g_T(x) = -1 \right]}{\Pr_{x\sim\pi}\left[ g_T(x) = -1 \right]} \\
&= \Pr_{x\sim\pi}\left[ f(x) = -1 \mid x_T = (-1)^{|T|} \right]
\end{aligned}$$

In the last line we used the fact that we have a product probability space, which implies

$$\Pr_{x\sim\pi}\left[ g_{A\setminus B}(x) = -1 \right] = \frac{\Pr_{x\sim\pi}\left[ g_A(x) = -1 \right]}{\Pr_{x\sim\pi}\left[ g_B(x) = -1 \right]}$$

for all $B \subseteq A$. ■

**The part of the formula containing only variables from the booster is still satisfiable** We now turn to analyzing the boost maximally quasi-unsatisfiable subformulas can give. In the end will will show that they cannot boost the unsatisfiability probability by a constant. Lemma 6.15 implies that the same holds for all satisfiable subformulas, thus giving us the desired contradiction.

Let $T \subseteq [N]$ with $\Phi_T$ mqu. In order to see how big the boost by such a $T$ can be, we split $x$ into two parts, the part $x_S$, so that each clause in $\Phi_S$ only contains variables from $V(T)$, and the part $x_{\overline{S}}$, in which each encoded clause contains at least one variable from $\overline{V(T)} = \{X_1, \ldots, X_n\} \setminus V(T)$. Let $f(x_S)$ be $-1$ if $\Phi_S$ is unsatisfiable and 1 otherwise. The following lemma asserts that $\Phi_S$ can only be unsatisfiable with probability in $o(1)$. This is the case, because it is very unlikely to flip one of the constant number of clauses that can make the maximally satisfiable booster unsatisfiable.

▶ **Lemma 6.16.** Let $T \subseteq [N]$ with $\Phi_T$ mqu and let $S \subseteq [N]$ be the indices of all clauses that only contain variables from $V(T)$. Then,

$$\Pr_{x\sim\pi}\left[ f(x_S) = -1 \mid x_T = (-1)^{|T|} \right] \in o(1).$$

◀

*Proof.* It holds that $T \subseteq S$ by definition. Due to Lemma 6.13 we can assume $x_T = (-1)^{|T|}$. Furthermore, because $\Phi_T$ is maximally satisfiable it holds that $f(x') = -1$ if $x'_T = x_T$ and if there is an $i \in S \setminus T$ with $x'_i = -1$, i. e. $\Phi_T$ becomes unsatisfiable if at least one other clause over $S$ is flipped. Since we already condition on $x_T = (-1)^{|T|}$ and since the clauses get flipped independently, it holds that

$$\Pr_{x\sim\pi}\left[ f(x_S) = -1 \mid x_T = (-1)^{|T|} \right] = \Pr_{x\sim\pi}[\, \exists\, i \in S \setminus T \colon x_i = -1 \,] \leq \binom{|V(T)|}{k} \cdot 2^k \cdot C_k \cdot \frac{k! \cdot s_c}{2^k} \cdot q_{\max} \in o(1),$$

where we overestimated $|S \setminus T| \leq |S| \leq \binom{|V(T)|}{k} \cdot 2^k$ and used equation (6.10). ■

**The booster adds shorter clauses to the other part of the formula** We can now concentrate on the case that $\Phi_S$ is satisfiable. Since $\Phi_T$ is maximally satisfiable, it holds that $\Phi_S = \Phi_T$, and since $\Phi_T$ is quasi-unsatisfiable, $\Phi_S$ also only has one satisfying assignment. We now want to create $x_{\overline{S}}$ under these conditions. To this end, we assume that the variables $V(T)$ take the one assignment that makes $\Phi_S$ satisfiable. For a clause containing both variables from $V(T)$ and variables from $\overline{V(T)}$ this means the clause is either satisfied or the variables from $V(T)$ can be eliminated as their literals are all set to false. Effectively, this means that this partial assignment can create clauses over $\overline{V(T)}$ of length $0 < l < k$. The following lemma gives an upper bound on the number $D_l$ of $l$-clauses we can create this way. However, with our requirement on $p_{\max}$ we can only get clauses of size $k-1$. The proof of the statement is a simple application of the Markov bound.

▶ **Lemma 6.17.** Let $p_{\max} \in o(s^{\star -(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(s^\star))$ and let $T \subseteq [N]$ so that $\Phi_T$ is maximally quasi-unsatisfiable. Assuming a partial assignment for the variables $V(T)$ that satisfies $\Phi_T$ a. a. s. creates at most

$$D_{k-1} \in O\left(s^\star \cdot p_{\max}\right) \in o\left(s^{\star 1 - \frac{3k-1}{4k-2}} \cdot \log^{-\frac{k-1}{2k-1}}(s^\star)\right)$$

clauses of length $k-1$ over $\overline{V(T)}$ and no shorter clauses. ◀

*Proof.* A clause $(\ell_1 \vee \ell_2 \vee \ldots \vee \ell_l)$ with $0 < l < k$ and $|\ell_1|, \ldots, |\ell_l| \in \overline{V(T)}$ is created if we flip at least one clause, which contains $(\ell_1 \vee \ell_2 \vee \ldots \vee \ell_l)$ and $l-k$ variables from $V(T)$ so that these are not satisfied by the partial assignment. Thus, the probability of creating a clause $(\ell_1 \vee \ell_2 \vee \ldots \vee \ell_l)$ is at most

$$C_k \frac{k! \cdot s_c}{2^k} \prod_{i=1}^{l} p(|\ell_i|) \cdot \sum_{J \in \mathcal{P}_{k-l}(V(T))} \prod_{X \in J} p(X) \leq C_k \frac{k! \cdot s_c}{2^k} \cdot \binom{|V(T)|}{k-l} \cdot p_{\max}^{k-l} \cdot \prod_{i=1}^{l} p(|\ell_i|), \tag{6.11}$$

since $\sum_{J\in\mathcal{P}_{k-l}(V(T))}\prod_{X\in J}p(X)\leq\binom{|V(T)|}{k-l}\cdot p_{\max}^{k-l}$. Summing over all the possibilities to choose $|\ell_1|,\dots,|\ell_l|\in\overline{V(T)}$ and their $2^l$ signs, the expected number of clauses of length $l$ added would be at most

$$E_l = C_k\frac{k!\cdot s_c}{2^{k-l}}\cdot\binom{|V(T)|}{k-l}\cdot p_{\max}^{k-l}\cdot\sum_{I\in\mathcal{P}_l\left(\overline{V(T)}\right)}\prod_{X\in I}p(X)\in O\left(s^\star\cdot p_{\max}^{k-l}\right),$$

since $C_k=1+o(1)$ and $s_c\in O(s^\star)$. With our requirement on $p_{\max}$ it holds that

$$E_l\in o\left(s^{\star 1-(k-l)\cdot\frac{3k-1}{4k-2}}\cdot\log^{-(k-l)\cdot\frac{k-1}{2k-1}}(s^\star)\right).$$

This expression is $o(1)$ for $l\leq k-2$. That means, due to a Markov bound, we do not create any clauses of size $l\leq k-1$ with probability $1-o(1)$. It remains to bound the number of $(k-1)$-clauses we create.

Let $h(s^\star)=s^{\star 1-(3k-1)/(4k-2)}\cdot\log^{-(k-1)/(2k-1)}(s^\star)$. If this expression We can use a Markov bound to prove that there are at least

$$D_{k-1}=\sqrt{h(s^\star)\cdot E_{k-1}}\in o\left(h(s^\star)\right)$$

clauses of length $k-1$ with probability at most

$$\frac{E_{k-1}}{\sqrt{h(s^\star)\cdot E_{k-1}}}\in\Theta\left(\frac{\sqrt{E_{k-1}}}{\sqrt{h(s^\star)}}\right)\in o(1).$$

■

We now want to create the resulting formula over variables from $\overline{V(T)}$ in two parts. First we create $k$-clauses over $\overline{V(T)}$ with the usual clause-flipping model, where the clause-probabilities are the same as in $\mathcal{F}\left(n,k,(\vec{p}^{(n)})_{n\in\mathbb{N}},s_c\right)$. Then, we add $D_{k-1}$ $(k-1)$-clauses over $\overline{V(T)}$ with a separate clause-drawing model. We let $\hat{\Phi}$ denote the random formula that this approach produces.

The probability $q_c$ to add a clause $c=(\ell_1\vee\ell_2\vee\ldots\vee\ell_{k-1})$ of size $k-1$ in our original clause-flipping model $\mathcal{F}^N(s_c)$ can be upper-bounded by

$$q_c\leq C_k\cdot\frac{k!\cdot s_c}{2^k}\cdot|V(T)|\cdot p_{\max}\cdot\prod_{i=1}^{k-1}p(|\ell_i|)$$

due to equation (6.11). However, if we want to use those probabilities when drawing clauses, they have to be normalized. This results in probabilities

$$q_c'=\frac{C_k\cdot\frac{k!\cdot s_c}{2^k}\cdot|V(T)|\cdot p_{\max}\cdot\prod_{i=1}^{k-1}p(|\ell_i|)}{\sum_{c=(\ell_1'\vee\ldots\vee\ell_{k-1}')}C_k\cdot\frac{k!\cdot s_c}{2^k}\cdot|V(T)|\cdot p_{\max}\cdot\prod_{i=1}^{k-1}p(|\ell_i'|)}$$

$$= C'_{k-1} \cdot \frac{(k-1)!}{2^{k-1}} \prod_{i=1}^{k-1} p(|\ell_i|), \tag{6.12}$$

for the clause drawing model, where $C'_{k-1} = 1 + O(p_{\max})$, since

$$\sum_{c=(\ell'_1 \vee \ell'_2 \vee \ldots \vee \ell'_{k-1})} \prod_{i=1}^{k-1} p(|\ell'_i|) = 2^{k-1} \sum_{S \subseteq \overline{V(T)} \colon |S|=k-1} \prod_{x \in S} p_x$$
$$\geq \frac{2^{k-1}}{(k-1)!} \cdot ((1 - (|V(T)| + k - 2) \cdot p_{\max})^{k-1})$$

due to Lemma 4.1. We can then apply Lemma 6.1 to relate the probability that $\hat{\Phi}$ is unsatisfiable to the probability that the original formula is unsatisfiable under its single satisfying assignment for $V(T)$.

▶ **Lemma 6.18.** It holds that

$$\Pr_{x \sim \pi}\left[ f(x) = -1 \wedge f(x_S) = 1 \mid x_T = (-1)^{|T|} \right] \leq \Pr\left[\hat{\Phi} \text{ unsat}\right] + o(1).$$

◀

*Proof.* First, we note that

$$\Pr_{x \sim \pi}\left[ f(x) = -1 \wedge f(x_S) = 1 \mid x_T = (-1)^{|T|} \right]$$
$$\leq \Pr_{x \sim \pi}\left[ f(x) = -1 \mid f(x_S) = 1 \wedge x_T = (-1)^{|T|} \right].$$

This means, we still get an upper bound on the desired probability by conditioning on both the event that $T$ is present and the event that no other clause with only variables from $V(T)$ is present. With these two conditions $x_S$ is fixed to $x_T$. Thus, we do not have to sample this part of the formula. We can assume to have a clause-flipping model over variables in $\overline{V(T)}$, where $k$-clauses and $(k-1)$-clauses are flipped. Therefore, we only have to consider the probability that our clause-flipping model generates unsatisfiable instances on $\overline{V(T)}$. The clause probabilities in this model are as follows. For $k$-clauses we use the original probabilities $q_c(s)$ and for $(k-1)$-clauses $c = (\ell_1, \ldots \ell_{k-1})$ we use the upper bound

$$q_c = C_k \cdot \frac{k! \cdot s_c}{2^k} \cdot |V(T)| \cdot p_{\max} \cdot \prod_{i=1}^{k-1} p(|\ell_i|). \tag{6.13}$$

Let $(\Omega', \pi')$ be the product space of this model and let $f' \colon \Omega' \to \{-1, 1\}$ be the characteristic function of unsatisfiability in this model. Since $f$ is monotone and equation (6.13) is an upper bound for the real probabilities of those clauses to appear, it holds that

$$\Pr_{x \sim \pi}\left[ f(x) = -1 \mid f(x_S) = 1 \wedge x_T = (-1)^{|T|} \right] \leq \Pr_{x' \sim \pi'}[\, f'(x') = -1 \,] \tag{6.14}$$

due to the observation in Lemma 3.9, that increasing any clause probability also increases the probability for a monotone property to hold.

The probability to have at most $D_{k-1}$ $(k-1)$-clauses in the clause-flipping model described above is $1-o(1)$ due to the same Markov bound as in Lemma 6.17. This holds since Lemma 6.17 uses exactly the same probabilities as upper bounds as the new clause-flipping model uses as clause probabilities for $(k-1)$-clauses. Therefore,

$$\Pr_{x'\sim\pi'}[\, f'(x') = -1 \,] \leq \Pr_{x'\sim\pi'}[\, f'(x') = -1 \wedge \ \leq D_{k-1}\ (k-1)\text{-clauses} \,] + o(1). \tag{6.15}$$

We can now use Lemma 6.3 and the monotonicity of $f'$ to derive

$$\begin{aligned}
&\Pr_{x'\sim\pi'}[\, f'(x') = -1 \wedge \leq D_{k-1}\ (k-1)\text{-clauses} \,] \\
&= \sum_{i=0}^{D_{k-1}} \left( \Pr_{x'\sim\pi'}[\, f'(x') = -1 \mid i\ (k-1)\text{-clauses} \,] \cdot \Pr_{x'\sim\pi'}[\, i\ (k-1)\text{-clauses} \,] \right) \\
&\leq \Pr_{x'\sim\pi'}[\, f'(x') = -1 \mid D_{k-1}\ (k-1)\text{-clauses} \,] \cdot \Pr_{x'\sim\pi'}[\, \leq D_{k-1}\ (k-1)\text{-clauses} \,] \\
&= \Pr_{x'\sim\pi'}[\, f'(x') = -1 \mid D_{k-1}\ (k-1)\text{-clauses} \,] + o(1).
\end{aligned} \tag{6.16}$$

This is possible, since we consider a monotone function on a product probability space and we condition on the number of clauses flipped in the restriction of $x'$, which encodes $(k-1)$-clauses. We now want to substitute flipping $(k-1)$-clauses with drawing $(k-1)$-clauses on $\overline{V(T)}$. The normalized probabilities of our models on $(k-1)$-clauses are

$$q'_c = C'_{k-1} \cdot \frac{(k-1)!}{2^{k-1}} \prod_{i=1}^{k-1} p(|\ell_i|)$$

with $C'_{k-1} = 1 + O(p_{\max})$ according to equation (6.12). Thus, for the flipping model we have a scaling factor of

$$s' = \frac{q_c}{q'_c} = \frac{C_k \cdot \frac{k!\cdot s_c}{2^k} \cdot |V(T)| \cdot p_{\max} \cdot \prod_{i=1}^{k-1} p(|\ell_i|)}{C'_{k-1} \cdot \frac{(k-1)!}{2^{k-1}} \prod_{i=1}^{k-1} p(|\ell_i|)} \in \Theta\left(s^\star \cdot p_{\max}\right)$$

and in the drawing model we draw $m' = D_{k-1} \in O(s^\star \cdot p_{\max})$ clauses. The maximum clause probability is $q'_{\max} \in \Theta(p_{\max}^{k-1})$. This implies $s' \cdot m' \cdot q_{\max} \in O(s^{\star 2} \cdot p_{\max}^{k+1}) \in o(1)$ due to our choice of $p_{\max}$. Thus, we can use Lemma 6.1 to derive

$$\begin{aligned}
&\Pr_{x'\sim\pi'}[\, f'(x') = -1 \mid \leq D_{k-1}\ (k-1)\text{-clauses} \,] \\
&\qquad = \Pr\left[\hat{\Phi} \text{ unsat} \mid \text{no } (k-1)\text{-clause drawn twice}\right] + o(1).
\end{aligned} \tag{6.17}$$

Let $C_{k-1}(V)$ be the set of all $(k-1)$-clauses over the variables $V \subseteq \{X_1, \ldots, X_n\}$. In the clause-drawing phase of creating $\hat{\Phi}$ the probability to draw a $(k-1)$-clause twice is at most

$$\binom{D_{k-1}}{2} \cdot \sum_{c \in C_{k-1}(\overline{V(T)})} {q'_c}^2 \leq D_{k-1}^2 \cdot q'_{\max} \in O(s^{\star 2} \cdot p_{\max}^{k+1}) \in o(1).$$

Thus, it holds that

$$\Pr\left[\hat{\Phi} \text{ unsat} \mid \text{no } (k-1)\text{-clause drawn twice}\right] \leq \Pr\left[\hat{\Phi} \text{ unsat}\right] + o(1). \quad (6.18)$$

Putting equation (6.14), equation (6.15), equation (6.17), equation (6.16), and equation (6.18) together yields the desired result of Lemma 6.18. ■

**Shorter clauses can be substituted with k-clauses** We now want to bound $\Pr\left[\hat{\Phi} \text{ unsat}\right]$. To this end, let $\tilde{\Phi}$ be the part of $\hat{\Phi}$ only consisting of $k$-clauses. Let us *assume* $\Pr\left[\hat{\Phi} \text{ unsat}\right] \geq \mu_{s_c}(f) + \delta$ for some constant $\delta > 0$. We know that $\tilde{\Phi}$ is unsatisfiable with probability at most $\mu_{s_c}(f)$, since it is drawn from $\mathcal{F}\left(n, k, (\vec{p}^{(n)})_{n \in \mathbb{N}}, s_c\right)$ with the difference that only clauses over $\overline{V(T)}$ are flipped. This implies $\Pr\left[\hat{\Phi} \text{ unsat} \wedge \tilde{\Phi} \text{ sat}\right] \geq \delta$. We now define a more general concept of coverability, analogously to Friedgut [Fri99]. This will allow us to substitute $(k-1)$-clauses with $k$-clauses while maintaining essentially the same probability to make $\hat{\Phi}$ unsatisfiable.

▶ **Definition 6.19.** Let $D_1, \ldots, D_a \in \mathbb{N}$ and $l_1, \ldots, l_a \in \mathbb{N}$ and let $\vec{q}_1, \ldots, \vec{q}_a$ be probability distributions. For $A \subseteq \{0,1\}^n$, we say that $A$ is $((d_1, l_1, \vec{q}_1), (d_2, l_2, \vec{q}_2), \ldots, (d_a, l_a, \vec{q}_a), \varepsilon)$-coverable, if the union of $d_i$ subcubes of co-dimension $l_i$ chosen according to probability distribution $\vec{q}_i$ for $1 \leq i \leq a$ has a probability of at least $\varepsilon$ to cover $A$. ◀

In contrast to Friedgut's definition, we permit subcubes of arbitrary co-dimension and with arbitrary probability distributions instead of only subcubes of co-dimension 1 with a uniform distribution. In the context of satisfiability we say that a specific formula (*not* a random formula) $F$ is $((d_1, l_1, \vec{q}_1), \ldots, (d_a, l_a, \vec{q}_a), \varepsilon)$-coverable if the probability to make it unsatisfiable by adding $d_i$ random clauses of size $l_i$ chosen according to distribution $\vec{q}_i$ for $i = 1, 2, \ldots a$ is at least $\varepsilon$ in total.

Now let $\vec{q}'_{k-1}$ be a vector of the clause drawing probabilities $q'_c$ for all clauses of size $k-1$ over $\overline{V(T)}$. It holds that with a sufficiently large constant probability $\tilde{\Phi}$ is $((D_{k-1}, k-1, \vec{q}'_{k-1}), \delta)$-coverable. The next lemma shows that formulas with this property are also $((g(n), k, \vec{q}'_k), \delta')$-coverable for some function $g(n) \in o(\sqrt{s^\star})$ and any constant $\delta' < \delta$. Here, $\vec{q}'_k$ is the vector of normalized clause probabilities for $k$-clauses on $\overline{V(T)}$, i. e. for a clause $c = (\ell_1, \ldots, \ell_k)$ with $|\ell_1|, \ldots, |\ell_k| \in \overline{V(T)}$

the clause probability is

$$q'_c = C'_k \cdot \frac{k!}{2^k} \prod_{i=1}^{k} p(|\ell_i|)$$

with $C'_k = 1 + O(p_{\max})$, equivalently to equation (6.12). The proof of the lemma is essentially a more precise version of Friedgut's original proof.

▶ **Lemma 6.20.** Let $\vec{q}_k$ be our original clause probability distribution, let $\vec{q}'_{k-1}$ be as described in equation (6.12), and let $D_{k-1}$ be as defined. If a concrete formula $F$ is $((D_{k-1}, k-1, \vec{q}'_{k-1}), \delta)$-coverable for some constant $\delta > 0$, it is also $((g(n), k, \vec{q}'_k), \delta')$-coverable for some function $g(n) \in o(\sqrt{s^\star})$ and for every constant $0 < \delta' < \delta$. ◀

*Proof.* Let $C_i$ denote the $i$-th random clause of length $k-1$ we add. We have to show that, if $F$ is $((D_{k-1}, k-1, \vec{q}'_{k-1}), \delta)$-coverable for some constant $\delta > 0$, it is also $((g(n), k, \vec{q}'_k), \delta')$-coverable for $g(n) \in o(\sqrt{s^\star})$ and some other constant $\delta' > 0$. For the sake of simplicity, let $\gamma_i$ denote the probability that the $i$-th $(k-1)$-clause makes $F$ unsatisfiable and that it was not made unsatisfiable by any formerly added $(k-1)$-clauses:

$$\gamma_i = \Pr\left[\left(F \bigwedge_{j=1}^{i} C_j\right) \text{unsat} \ \wedge \left(F \bigwedge_{j=1}^{i-1} C_j\right) \text{sat}\right].$$

Now we look at $\gamma_i$, starting from $i = D_{k-1}$. $\gamma_i$ represents the contribution of clause $C_i$ to the overall probability $\delta$ to cover $F$. If $\gamma_i < \delta/(2 \cdot D_{k-1})$, we simply delete that clause. Otherwise, we can substitute it with $\Theta(D_{k-1}^{k/(k-1)} \cdot \log D_{k-1})$ $k$-clauses, while losing at most $\delta/(4 \cdot D_{k-1})$ of the total probability $\delta$. This fact will be shown in the next step. We then reorder the clauses to add $k$-clauses first. If we repeat this step until all $D_{k-1}$ $(k-1)$-clauses are either deleted or replaced, the remaining probability will be at least $\delta' = \delta/2$.

If we want to substitute $(k-1)$-clauses with $k$-clause, it holds that we have a random formula $\Phi = F \bigwedge_{j=1}^{i-1} C_j$ so that $\Pr[(\Phi \wedge C) \text{ unsat} \wedge \Phi \text{ sat}] = \gamma_i \geq \delta/(2 \cdot D_{k-1})$ for some constant $\delta > 0$ and some $(k-1)$-clause $C$ drawn at random according to distribution $\vec{q}'_{k-1}$. Now we want to know what the probability is to have a concrete formula $\Phi'$ which is satisfiable and satisfies $\Pr[(\Phi' \wedge C) \text{ unsat}] \geq \delta/(4 \cdot D_{k-1})$. Let this probability be called $P_{\text{good}}$. It holds that

$$\begin{aligned}
\gamma_i &= \Pr[(\Phi \wedge C) \text{ unsat} \wedge \Phi \text{ sat}] \\
&= \sum_{\Phi' \text{ sat}} (\Pr[\Phi = \Phi'] \cdot \Pr[(\Phi' \wedge C) \text{ unsat}]) \\
&< P_{\text{good}} + (1 - P_{\text{good}}) \cdot \frac{\delta}{4 \cdot D_{k-1}},
\end{aligned}$$

since with probability $P_{\text{good}}$ we have a good formula with $\Pr[\Phi' \wedge C \text{ unsat}] \in$

$[\delta/(4 \cdot D_{k-1}), 1]$ and with probability $\Pr[\Phi \text{ sat}] - P_{\text{good}} < 1 - P_{\text{good}}$ we have a satisfiable formula with $\Pr[(\Phi' \wedge C) \text{ unsat}] < \delta/(4 \cdot D_{k-1})$. From this we can derive

$$P_{\text{good}} \geq \frac{\gamma_i - \frac{\delta}{4 \cdot D_{k-1}}}{1 - \frac{\delta}{4 \cdot D_{k-1}}}.$$

We will show subsequently that for exactly those formulas $\Phi'$, we can substitute the random $(k-1)$-clause with $D'_{k-1} \in \Theta(D_{k-1}^{k/(k-1)} \log D_{k-1})$ $k$-clauses so that after the substitution it holds that

$$\Pr\left[\Phi' \bigwedge_{j=1}^{D'_{k-1}} C_j^{(k)} \text{ unsat}\right] \geq \left(1 - \frac{\delta}{4 \cdot D_{k-1}}\right).$$

This implies

$$\Pr\left[\Phi \bigwedge_{j=1}^{D'_{k-1}} C_j^{(k)} \text{ unsat} \wedge \Phi \text{ sat}\right] \geq P_{\text{good}} \cdot \left(1 - \frac{\delta}{4 \cdot D_{k-1}}\right) \geq \gamma_i - \frac{\delta}{4 \cdot D_{k-1}}.$$

This means, we only lose $\delta/(4 \cdot D_{k-1})$ of the total probability $\delta$ as desired.

Now assume we had a concrete satisfiable formula $\Phi'$ with

$$\Pr\left[\Phi' \wedge C^{(k-1)} \text{ unsat}\right] = x \geq \frac{\delta}{4 \cdot D_{k-1}}$$

for a $(k-1)$-clause $C^{(k-1)}$ drawn at random according to distribution $\vec{q}'_{k-1}$. Now let us see how many $k$-clauses we need to substitute this $(k-1)$-clause. The fact that $\Phi'$ is coverable with a single $(k-1)$-clause with probability at least $\delta/(4 \cdot D_{k-1})$ means, that there is a subset of literals $L$ which appears in all satisfying assignments. Furthermore, the probability to draw a clause which forbids those literals is at least $\delta/(4 \cdot D_{k-1})$. This is the case if the clause contains literals from $L$, but with inverted signs. Let us denote by $\overline{L}$ the set of literals from $L$ with inverted signs. To cover $\Phi'$ with a $k$-clause, the $k$-clause has to contain only literals from $\overline{L}$. It holds that

$$\begin{aligned} x &\leq \Pr\left[C^{(k-1)} \subseteq \overline{L}\right] \\ &= C'_{k-1} \cdot \frac{(k-1)!}{2^{k-1}} \sum_{S \subseteq \overline{L}\colon |S|=k-1} \prod_{\ell \in S} p(|\ell|) \leq C'_{k-1} \cdot \left(\frac{1}{2} \sum_{\ell \in \overline{L}} p(|\ell|)\right)^{k-1} \end{aligned}$$

due to equation (6.12). The last inequality gives us

$$\frac{1}{2} \sum_{\ell \in \overline{L}} p(|\ell|) \geq \left(\frac{x}{C'_{k-1}}\right)^{1/(k-1)}.$$

Note that $x^{1/(k-1)} \in \Omega(D_{k-1}^{-1/(k-1)}) \in \omega(p_{\max})$ and thus, $|L| \geq k$. The probability to cover $\Phi'$ with a $k$-clause is now

$$\begin{aligned}
\Pr\left[C^{(k)} \subseteq \overline{L}\right] &= C'_k \cdot \frac{k!}{2^k} \sum_{S \subseteq \overline{L}\colon\, |S|=k} \prod_{\ell \in S} p(|\ell|) \\
&\geq C'_k \cdot \frac{k!}{2^k} \cdot \frac{1}{k!} \cdot \left( \sum_{\ell \in \overline{L}} p(|\ell|) - k \max_{\ell' \in \overline{L}} (p(|\ell'|)) \right)^k \\
&\geq C'_k \cdot \left( \frac{1}{2} \sum_{\ell \in \overline{L}} p(|\ell|) - \frac{k}{2} \cdot p_{\max} \right)^k \\
&= \Theta\left( \left( \frac{x}{C'_{k-1}} \right)^{k/(k-1)} \right),
\end{aligned}$$

since $\frac{1}{2} \sum_{\ell \in \overline{L}} p(|\ell|) \geq (x/C'_{k-1})^{1/(k-1)} \in \omega(p_{\max})$, $C'_k = 1 + o(s^{\star -1/k})$, and $C'_{k-1} = 1 + o(s^{\star -1/k})$. It follows that the probability to cover $\Phi'$ with $g(n)$ $k$-clauses is at least

$$1 - \left(1 - \Theta\left(x^{k/(k-1)}\right)\right)^{g(n)} \geq 1 - e^{-\Theta\left(x^{k/(k-1)}\right) \cdot g(n)} \geq 1 - \frac{\delta}{4 \cdot D_{k-1}}$$

for $g(n) \in \Omega(x^{-k/(k-1)} \log(4 \cdot D_{k-1}/\delta))$, which is $\Omega(D_{k-1}^{k/(k-1)} \log D_{k-1})$ for $x \geq \delta/(4 \cdot D_{k-1})$ as desired.

It now remains to count how many $k$-clauses we needed. In each substitution step we used $\Theta(D_{k-1}^{k/(k-1)} \log D_{k-1})$ $k$-clauses, while there are at most $D_{k-1}$ $(k-1)$-clauses. Therefore, we need at most $O(D_{k-1}^{k/(k-1)+1} \cdot \log D_{k-1})$ $k$-clauses to substitute $(k-1)$-clauses. Furthermore, in our case

$$g(n) \in \Theta\left(D_{k-1}^{k/(k-1)+1} \cdot \log D_{k-1}\right) \in o\left(\sqrt{s^\star}\right),$$

since $p_{\max} \in o(s^{\star -(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(s^\star))$. Please note, that, instead of additive errors $\delta/(4 \cdot D_{k-1})$ per substitution and $\delta/(2 \cdot D_{k-1})$ per deletion, we could have chosen any arbitrarily small constant fraction of $\delta/D_{k-1}$. With this in mind, we can actually achieve a cover probability of $\delta'$ for any constant $\delta' < \delta$ with the same asymptotic number of $k$-clauses.

■

The former lemma states that if our random formula $\tilde{\Phi}$ is at least $((D_{k-1}, k-1, \vec{q}'_{k-1}), \delta)$-coverable, we can substitute the second step of getting $\hat{\Phi}$ by instead adding $g(n)$ $k$-clauses. Let $\Phi'$ denote the random formula we get this way, i. e. $\tilde{\Phi}$ and $g(n)$ additional $k$-clauses. What is the overall probability that $\Phi'$ is unsatisfiable? We choose a constant $\varepsilon > 0$ and call a formula $F$ *good* if it is satisfiable and $((D_{k-1}, k-1, \vec{q}'_{k-1}), \varepsilon_F)$-coverable for some constant $\varepsilon_F \geq \varepsilon$. Also,

we let $R$ denote the set of random clauses of sizes $k-1$ we add to $F$. It holds that

$$\Pr\left[\hat{\Phi} \text{ unsat} \wedge \tilde{\Phi} \text{ sat}\right] = \sum_{F \text{ sat}} \Pr\left[\tilde{\Phi} = F\right] \cdot \Pr[F \wedge R \text{ unsat}].$$

That means, if we substitute shorter clauses with $k$-clauses, we decrease $\varepsilon_F = \Pr[F \wedge R \text{ unsat}]$ by at most $\varepsilon$ if $F$ is good. If $F$ is bad, we cannot guarantee anything, so we might lose the contribution of those formulas completely. However, bad formulas $F$ satisfy $\Pr[F \wedge R \text{ unsat}] < \varepsilon$. That means, in total we lose at most

$$\sum_{F \text{ good}} \Pr\left[\tilde{\Phi} = F\right] \cdot \varepsilon + \left(\Pr\left[\tilde{\Phi} \text{ sat}\right] - \Pr\left[\tilde{\Phi} \text{ good}\right]\right) \cdot \varepsilon \leq \Pr\left[\tilde{\Phi} \text{ sat}\right] \cdot \varepsilon \leq \varepsilon.$$

We can choose $\varepsilon = \delta/2$ to guarantee $\Pr[\Phi' \text{ unsat}] \geq \mu_{s_c}(f) + \delta/2$.

**Bounding the boost by bounding the slope of the probability function** We can now show that instead of adding $g(n)$ $k$-clauses, we can increase the scaling factor $s$ of our original clause-flipping model by a value $s' \in \Theta(g(n))$ to achieve the same probability. The proof uses Lemma 6.1. However, for the lemma to work, we have to ensure $s' \cdot g(n) \in o(q_{\max}^{-1})$. This condition is satisfied due to the requirement $p_{\max} \in o(s^{\star -(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(s^\star))$. It implies $g(n)^2 \in o(s^\star)$ and $s^\star \in o(q_{\max}^{-1})$ holds due to equation (6.10).

Also note that we assume $g(n) \in \omega(1)$ for the rest of the proof. Assume there was some constant that upper-bounded $s^\star$, i. e. $s^\star \in O(1)$. This means, due to $g(n) \in o(\sqrt{s^\star})$, we would not need to add any additional clauses to get a probability of at least $\mu_{s_c} + \delta/2$. However, in that case $\tilde{\Phi} = \Phi'$ and we know $\Pr\left[\tilde{\Phi} \text{ unsat}\right] \leq \mu_{s_c}$, a contradiction. That means, $s^\star \notin O(1)$. As with our definitions of coarse thresholds, that means for *every* constant $\varepsilon > 0$ there are infinitely many $n \in \mathbb{N}$ (among the ones we consider with the coarse threshold property) such that $s^\star(n) \geq \varepsilon$. From this we can derive that there is a series of values $n \in \mathbb{N}$ that satisfy $s^\star(n) \in \omega(1)$ by doing the following: Every time we encounter a value $s^\star(n)$, we restrict the partial function to values of at least $s^\star(n)$ from this point on. With $s^\star(n) \in \omega(1)$ we can now show, that $g(n) \in \omega(1)$ as well. We know that $g(n) \in o(\sqrt{s^\star})$. Since the actual value of $g(n)$ is not relevant, as long as $g(n) \in o(\sqrt{s^\star})$, we can choose $g'(n) = \max(g(n), s^{\star 1/3})$. This guarantees both $g'(n) \in o(\sqrt{s^\star})$ and $g'(n)^2 \cdot q_{\max} \in o(1)$. Also, increasing the number of clauses can only improve the cover probability. By this argumentation, we can assume $g(n) \in \omega(1)$ for the rest of the proof.

▶ **Lemma 6.21.** For $s'(n) = 4 \cdot g(n) \in o(\sqrt{s^\star})$ it holds that

$$\Pr[\Phi' \text{ unsat}] \leq \mu_{s_c(n)+s'(n)}(\{f(x) = -1\}) + o(1).$$

◀

*Proof.* Instead of adding $g(n) \in o(s^\star)$ $k$-clauses to get $\Phi'$, we add another phase of clause flipping. In this phase $k$-clauses with only variables from $\overline{V(T)}$ are

flipped with the normalized probabilities $\vec{q}'_k$ and scaling factor $s'$. We can relate the model drawing $g(n)$ $k$-clauses with the flipping model in the following way. Let $F$ be a satisfiable formula coverable by $k$-clauses, let $\mathcal{D}(g(n))$ be the drawing model, $\mathcal{F}(s')$ be the flipping model, and $S$ be the random set of clauses created by those models. Due to the requirement $g(n)^2 \in o(q_{\max}^{-1})$ it holds that the probability to draw one of the $k$-clauses twice is at most

$$\binom{g(n)}{2} \cdot \sum_{i \in N} q_i'^2 \leq g(n)^2 \cdot q'_{max} \in \Theta(g(n)^2 \cdot q_{max}) \in o(1),$$

since the probability vector $\vec{q}$ of the original clause probabilities and the vector $\vec{q}'_k$ of normalized probabilities on $\overline{V(T)}$ differ in a factor of at most $1 + O(p_{\max})$. Thus,

$$\begin{aligned}
&\Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \,] \\
&= \Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \wedge |S| = g(n) \,] + \Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \wedge |S| < g(n) \,] \\
&= \Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \wedge |S| = g(n) \,] + o(1) \\
&= \Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \mid |S| = g(n) \,] \cdot \Pr_{S \sim \mathcal{D}(g(n))}[\, |S| = g(n) \,] + o(1) \\
&\leq \Pr_{S \sim \mathcal{D}(g(n))}[\, S \text{ covers } F \mid |S| = g(n) \,] + o(1). \qquad (6.19)
\end{aligned}$$

In order to relate the two models, we have to ensure that the flipping model $\mathcal{F}(s')$ a. a. s. flips at least $g(n)$ clauses. The expected number of clauses flipped would be exactly $s'$ if clauses with variables from $V(T)$ were flipped as well, since clause probabilities are normalized in the original model. However, here we have to exclude their probabilities, which sum up to at most

$$C_k \cdot \frac{k!}{2^k} \cdot \left( \sum_{i=1}^{k} \binom{|V(T)|}{i} \cdot {p_{\max}}^i \right) \in O(p_{\max}).$$

Thus, it holds that $\mathbb{E}_{S \sim \mathcal{F}(s')}[\, |S| \,] = s' \cdot (1 - O(p_{\max}))$. Due to a Chernoff bound it would be sufficient to assume $s'(n) = 2 \cdot g(n)$ to get

$$\Pr_{S \sim \mathcal{F}(s')}[\, |S| < g(n) \,] < \exp\left( -\left( \frac{1 - O(p_{\max})}{2} \right)^2 \cdot g(n) \right) \in o(1).$$

With $g(n) \in \omega(1)$ this implies $\Pr_{S \sim \mathcal{F}(s')}[\, |S| \geq g(n) \,] = 1 - o(1)$ and thus

$$\begin{aligned}
&\Pr_{S \sim \mathcal{F}(s')}[\, S \text{ covers } F \,] \\
&\geq \Pr_{S \sim \mathcal{F}(s')}[\, S \text{ covers } F \wedge |S| \geq g(n) \,]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=g(n)}^{N} \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \wedge |S| = i \,] \\
&= \sum_{i=g(n)}^{N} \left( \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \mid |S| = i \,] \cdot \Pr_{S\sim\mathcal{F}(s')}[\, |S| = i \,] \right) \\
&\geq \sum_{i=g(n)}^{N} \left( \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \mid |S| = g(n) \,] \cdot \Pr_{S\sim\mathcal{F}(s')}[\, |S| = i \,] \right) \\
&= \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \mid |S| = g(n) \,] \cdot \Pr_{S\sim\mathcal{F}(s')}[\, |S| \geq g(n) \,] \\
&= \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \mid |S| = g(n) \,] - o(1), \qquad (6.20)
\end{aligned}
$$

where we used Lemma 6.2 in line 5. We can do this, since the property that a randomly flipped set of clauses $S$ covers a given formula $F$ is a monotone property. Equation (6.19) and equation (6.20) together with Lemma 6.1 now yield

$$
\begin{aligned}
\Pr_{S\sim\mathcal{D}(g(n))}[\, S \text{ covers } F \,] &\leq \Pr_{S\sim\mathcal{D}(g(n))}[\, S \text{ covers } F \mid |S| = g(n) \,] + o(1) \\
&\leq \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \mid |S| = g(n) \,] + o(1) \leq \Pr_{S\sim\mathcal{F}(s')}[\, S \text{ covers } F \,] + o(1).
\end{aligned}
$$

Note that we can use Lemma 6.1 due to $s' \cdot g(n) \in \Theta(g(n)^2) \in o(q_{\max}^{-1})$. We have now established that instead of drawing $g(n)$ $k$-clauses with variables only from $\overline{V(T)}$, we can flip those clauses with their normalized probabilities $\vec{q}'_k$ and scaling factor $s' = 2 \cdot g(n)$. Thus, for every clause $c = (\ell_1 \vee \ell_2 \vee \ldots \vee \ell_k)$ we independently flip a coin twice and take it into the formula if at least one of the flips is successful. This results in clause probabilities of at most

$$
\begin{aligned}
&1 - \left(1 - C_k \cdot s_c \cdot \frac{k!}{2^k} \cdot \prod_{i=1}^{k} p(|\ell_i|)\right) \cdot \left(1 - C'_k \cdot s' \cdot \frac{k!}{2^k} \cdot \prod_{i=1}^{k} p(|\ell_i|)\right) \\
&\leq \left(C_k \cdot s_c + C'_k \cdot s'\right) \cdot \frac{k!}{2^k} \cdot \prod_{i=1}^{k} p(|\ell_i|) \\
&\leq C_k \cdot (s_c + 2 \cdot s') \cdot \frac{k!}{2^k} \cdot \prod_{i=1}^{k} p(|\ell_i|),
\end{aligned}
$$

since $C_k \geq 1$, $C'_k = 1 + O(p_{\max}) = 1 + o(1)$, and thus $C'_k \leq 2 \cdot C_k$ for sufficiently large $n$. Thus, we can instead flip each clause with its original probability and a scaling factor of $s_c + 4 \cdot g(n)$. Since we consider a monotone property (making a formula unsatisfiable) this only increases the probability for the property to hold. ■

Under the assumption that $\Pr\left[\hat{\Phi} \text{ unsat}\right] \geq \mu_{s_c}(f) + \delta$ for a constant $\delta > 0$, it

follows that $\mu_{s_c+s'}(f) \geq \mu_{s_c}(f) + \varepsilon$ for $s' = 4 \cdot g(n)$ and some constant $\varepsilon > 0$. We show that this cannot be the case. The proof of this lemma requires $s' \in o(\sqrt{s_c})$, which is ensured by $p_{\max} \in o(s^{\star -(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(s^\star))$.

▶ **Lemma 6.22.** It holds that $\mu_{s_c+s'}(f) \leq \mu_{s_c}(f) + o(1)$ if $s' \in o(\sqrt{s^\star})$. ◀

*Proof.* Remember that $s_c \in \Theta(s^\star)$. We let $h(n) = {s_c}^{1/4}/{s'}^{1/2} \in \omega(1)$. Due to a Chernoff bound it holds that

$$\Pr_{\Phi \sim \mathcal{F}(s_c+s')}\left[\, |\Phi| > (s_c+s') + \sqrt{(s_c+s')} \cdot h(n) \,\right] < e^{-\frac{h(n)^2}{3}} \in o(1).$$

We will now compare $\mu_{s_c}(f)$ and $\mu_{s_c+s'}(f)$ directly. It holds that

$$\mu_{s_c}(f) = \sum_{x \in \{-1,1\}^N \colon f(x)=-1} \mu_{s_c}(x),$$

where

$$\mu_{s_c}(x) = \left(\prod_{i \in [N] \colon x_i=-1} s_c \cdot q_i\right) \cdot \left(\prod_{i \in [N] \colon x_i=1} (1 - s_c \cdot q_i)\right).$$

Due to the upper bound on the size of $\Phi$ it holds that

$$\mu_{s_c+s'}(f) = o(1) + \sum_{\substack{x \in \{-1,1\}^N \colon f(x)=-1, \\ |x|_{-1} \leq (s_c+s') + \sqrt{(s_c+s')} \cdot h(n)}} \mu_{s_c+s'}(x).$$

This allows us to compare the probabilities for a given $x \in \{-1,1\}^N$ as follows

$$\begin{aligned}
\mu_{s_c+s'}(x) &= \left(\frac{s_c+s'}{s_c}\right)^{|x|_{-1}} \cdot \left(\prod_{i \in [N] \colon x_i=1} \frac{1 - (s_c+s') \cdot q_i}{1 - s_c \cdot q_i}\right) \cdot \mu_{s_c}(x) \\
&= \left(1 + \frac{s'}{s_c}\right)^{|x|_{-1}} \cdot \left(\prod_{i \in [N] \colon x_i=1} 1 - \frac{s' \cdot q_i}{1 - s_c \cdot q_i}\right) \cdot \mu_{s_c}(x) \\
&\leq \left(1 + \frac{s'}{s_c}\right)^{(s_c+s') + \sqrt{(s_c+s')} \cdot h(n)} \cdot \left(\prod_{i \in [N] \colon x_i=1} 1 - s' \cdot q_i\right) \cdot \mu_{s_c}(x) \\
&\leq \exp\left(\frac{s'}{s_c} \cdot \left((s_c+s') + \sqrt{(s_c+s')} \cdot h(n)\right) - s' \cdot \sum_{i \in [N] \colon x_i=1} q_i\right) \cdot \mu_{s_c}(x).
\end{aligned}$$

We want to show that the exponent of the leading factor is $o(1)$ and thus

$$\mu_{s_c+s'}(f) \leq o(1) + e^{o(1)} \cdot \sum_{\substack{x \in \{-1,1\}^N \colon f(x)=-1, \\ |x|_{-1} \leq (s_c+s') + \sqrt{(s_c+s')} \cdot h(n)}} \mu_{s_c}(x) \leq \mu_{s_c}(f) + o(1).$$

With

$$\sum_{i\in[N]\colon x_i=1} q_i \geq 1 - ((s_c + s') + \sqrt{(s_c + s')} \cdot h(n)) \cdot q_{\max}$$

it holds that

$$\begin{aligned}
&\frac{s'}{s_c} \cdot ((s_c + s') + \sqrt{(s_c + s')} \cdot h(n)) - s' \cdot \sum_{i\in[N]\colon x_i=1} q_i \\
&\leq s' + \frac{s'^2}{s_c} + \frac{\sqrt{s_c + s'} \cdot h(n) \cdot s'}{s_c} - s' + s' \cdot ((s_c + s') + \sqrt{(s_c + s')} \cdot h(n)) \cdot q_{\max} \\
&\leq \frac{s'^2}{s_c} + 2\frac{s'^{1/2}}{s_c^{1/4}} + s' \cdot (s_c + s') \cdot q_{\max} + \sqrt{(s_c + s')} \cdot h(n) \cdot q_{\max}.
\end{aligned}$$

We chose $p_{\max}$ in such a way that $s'^2 \in \Theta(g(n)^2) \in o(s^\star)$. It also holds that $s' \cdot s_c \cdot q_{\max} \in o(s^{\star 3/2} \cdot q_{\max}) \in o(1)$. This yields an exponent of $o(1)$ and thus establishes the result as desired. ■

The last lemma contradicts our conclusion of $\mu_{s_c+4\cdot g(n)}(f) \geq \mu_{s_c}(f) + \varepsilon$ for some constant $\varepsilon > 0$. Therefore, our assumption $\Pr\left[\hat{\Phi} \text{ unsat}\right] \geq \mu_{s_c}(f) + \delta$ for $\delta > 0$ constant has to be false, i.e. for every constant $\varepsilon > 0$ it holds that $\Pr\left[\hat{\Phi} \text{ unsat}\right] \leq \mu_{s_c}(f) + \varepsilon$ for all sufficiently large values of $n$. Now we can put all error probabilities together to see

$$\Pr_{x\sim\pi}\left[ f(x) = -1 \mid x_T = (-1)^{|T|} \right] \leq \mu_{s_c}(f) + \varepsilon + o(1).$$

Especially, for any given $\tau$ this is smaller than $\mu_{s_c}(f) + \tau$ for all sufficiently large values of $n$. This means, for every constant $\tau$ the maximally quasi-unsatisfiable subformula $\Phi_T$ cannot be a $\tau$-booster. Due to Lemma 6.15 the boost by every satisfiable subformula is at most as big as the one by a mqu subformula. Thus, no $T$ which encodes a satisfiable subformula can be a $\tau$-booster. Since we already ruled out unsatisfiable subformulas, this means there are no $\tau$-boosters which appear with probability at least $\tau/2$. This contradicts the implication of the Sharp Threshold Theorem and therefore the assumption of a coarse threshold, thus proving Theorem 6.12. ■

As stated in the introduction of this chapter, our sharpness result for the clause flipping model $\mathcal{F}^N$ together with the results relating $\mathcal{F}^N$ and $\mathcal{D}^N$ yield the following corollary. It states that the sharpness result also holds for the clause drawing model $\mathcal{D}^N$ with the same parameters and with respect to the number of drawn clauses $m$.

▶ **Corollary 6.23.** Let $k \geq 3$, let $(\vec{p}^{(n)})_{n\in\mathbb{N}}$ be an ensemble of probability distributions on $n$ variables each and let $m^\star$ be an asymptotic satisfiability threshold for $\mathcal{D}(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ with respect to $m$. If $p_{\max} \in o(m^{\star -(3k-1)/(4k-2)}$ .

$\log^{-(k-1)/(2k-1)}(m^\star)$), then the satisfiability threshold on $\mathcal{D}(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m)$ with respect to $m$ is sharp. ◀

*Proof.* In order to use Lemma 6.4 and Lemma 6.5 to relate the clause flipping and clause drawing models, we have to ensure $m^\star \cdot s^\star \cdot q_{\max} \in o(1)$. This holds due to the prerequisite $p_{\max} \in o(m^{\star-(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(m^\star))$, which implies $p_{\max} \in o(m^{\star-2/k})$ and thus $q_{\max} \in \Theta(p_{\max}^k) \in o(m^{\star-2})$. We can now show the corollary as follows:

1. $\mathcal{D}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m\big)$ has asymptotic threshold function $m^\star$.
2. $\mathcal{F}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s\big)$ has the same asymptotic threshold function $s^\star = m^\star$. See Lemma 6.4.
3. $\mathcal{F}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s\big)$ with asymptotic threshold function $s^\star$ has a sharp threshold. See Theorem 6.12.
4. Sharpness of the threshold in $\mathcal{F}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, s\big)$ implies sharpness of the threshold in $\mathcal{D}\big(n, k, (\vec{p}^{(n)})_{n\in\mathbb{N}}, m\big)$. See Lemma 6.5.

■

## 6.5 Examples

We can now analyze the sharpness of satisfiability thresholds of non-uniform random $k$-SAT with given ensembles of probability distributions and known asymptotic threshold functions. As before, we consider the three models random $k$-SAT, power-law random $k$-SAT, and geometric random $k$-SAT. We already know the asymptotic threshold functions of those models from Section 5.3.

### 6.5.1 Random k-SAT

For random $k$-SAT the probability ensemble is

$$\forall n \in \mathbb{N} \colon \vec{p}^{(n)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

We know that the asymptotic threshold function is $m^\star \in \Theta(n)$. It holds that $p_{\max} = n^{-1} \in o(n^{-(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(n))$ for $k \geq 2$. Thus, the threshold of random $k$-SAT is also sharp for $k \geq 3$.

### 6.5.2 Power-Law Random k-SAT

Corollary 6.23 now implies the following corollary for power-law random $k$-SAT.

▶ **Corollary 6.24.** For power-law random $k$-SAT with $\beta > \frac{5k-3}{k-1}$ the satisfiability threshold is sharp. ◀

*Proof.* For power-law random $k$-SAT we assume some fixed $\beta > 2$. Then, for $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{(n/i)^{\frac{1}{\beta-1}}}{\sum_{j=1}^{n}(n/j)^{\frac{1}{\beta-1}}}.$$

It already holds that $p_1 \geq p_2 \geq \ldots \geq p_n$. Lemma 3.12 yields

$$p_1 = p_{\max} = (1 \pm o(1)) \cdot \left(\frac{\beta-2}{\beta-1}\right) \cdot n^{-\frac{\beta-2}{\beta-1}}.$$

We also know that the asymptotic threshold function is $m^\star \in \left(n^{k \cdot \frac{\beta-2}{\beta-1}}\right)$ for $\beta < \frac{2k-1}{k-1}$ and $m^\star \in \Theta(n)$ for $\beta > \frac{2k-1}{k-1}$. In the first case, the requirement $p_{\max} \in o(n^{-(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(n))$ is not fulfilled. In the second case, the requirement is fulfilled for $\beta > \frac{5k-3}{k-1}$. ■

### 6.5.3 Geometric Random k-SAT

Corollary 6.23 also implies the following corollary for power-law random $k$-SAT.

▶ **Corollary 6.25.** For geometric random k-SAT with base $b > 1$, the satisfiability threshold is sharp. ◀

*Proof.* For $n \in \mathbb{N}$ the distribution is $\vec{p}^{(n)} = \left(p_1^{(n)}, p_2^{(n)} \ldots, p_n^{(n)}\right)$ with

$$p_i^{(n)} = \frac{b \cdot (1 - b^{-1/n})}{b-1} \cdot b^{-(i-1)/n}.$$

Again, it already holds that $p_1 \geq p_2 \geq \ldots \geq p_n$. Lemma 3.13 states

$$p_1 = p_{\max} = \frac{b \cdot (1 - b^{-1/n})}{(b-1)} = (1 + o(1)) \cdot \frac{b \cdot \ln b}{(b-1)} \cdot n^{-1}$$

and the asymptotic threshold function is $m^\star \in \Theta(n)$. Thus, as for random $k$-SAT it holds that $p_{\max} \in \Theta(n^{-1}) \in o(n^{-(3k-1)/(4k-2)} \cdot \log^{-(k-1)/(2k-1)}(n))$ for $k \geq 2$. Therefore, the satisfiability threshold is sharp for geometric random $k$-SAT according to Corollary 6.23. ■

## 6.6 Remarks

We defined sharpness and coarseness of thresholds in such a way that we have a dichotomy as soon as a threshold exists. Thus, if we have an asymptotic threshold function, the threshold must be either sharp or coarse. However, the

result we provide is only fit to identify sharp thresholds. It only works if some specific conditions on the ensemble of probability distributions in relation to the asymptotic threshold position are fulfilled, but we do not know if these conditions correctly identify the dichotomy. We are simply missing some condition on the coarseness of thresholds.

There is some evidence that suggests that our sharpness result can be improved. For power-law random $k$-SAT we have an asymptotic threshold function of $m^\star = n^{k(\beta-2)/(\beta-1)}$ for $\beta < \frac{2k-1}{k-1}$ and an asymptotic threshold function of $m^\star = n$ for $\beta > \frac{2k-1}{k-1}$. We also know that in the former case, the threshold is coarse. However, in the latter case, we can only show that the threshold is sharp for $\beta > \frac{5k-3}{k-1}$. But what happens for $\beta \in (\frac{2k-1}{k-1}, \frac{5k-3}{k-1}]$? We conjecture that the threshold is sharp in that range of $\beta$ as well, but we might need more involved techniques to prove it.

# 7 Conclusions & Outlook

In this thesis we studied a generalization of the random $k$-SAT model, which we call non-uniform random $k$-SAT. The model incorporates expected frequencies for the Boolean variables of a random formula in $k$-CNF by means of a probability distribution $\vec{p}$ over these variables according to which they appear in random clauses. Given an ensemble of probability distributions $\left(\vec{p}^{(n)}\right)_{n\in\mathbb{N}}$ and a clause size $k$, we can analyze the limiting behavior of non-uniform random $k$-SAT instances as the number of variables $n$ increases. In the introduction of this thesis, we posed two questions regarding this model: First, how does the satisfiability threshold behave? Second, how hard is it to solve instances of the model?

Regarding the first question, we thoroughly analyzed the threshold behavior of non-uniform random $k$-SAT. We showed that the position and sharpness of the satisfiability threshold for non-uniform random 2-SAT depends on the two highest variable probabilities and their relations to the sum of squares of the remaining probabilities. If $p_{\max}^2 \in o\left(\sum_{i=1}^n p_i^2\right)$, the threshold is sharp at $m^\star = 1/\left(\sum_{i=1}^n p_i^2\right)$. Otherwise, the threshold is coarse at $m^\star = (1 - (\sum_{i=1}^n p_i^2))/((\sum_{i=2}^n p_i^2) + p_1 \cdot \left(\sum_{i=2}^n p_i^2\right)^{1/2})$. Depending on the relation of $p_2$ (the second-highest variable probability) to $\sum_{i=2}^n p_i^2$, the coarseness either stems from the emergence of an unsatisfiable subformula containing only the two most-frequent Boolean variables or an unsatisfiable subformula with four clauses over three different Boolean variables. This completely characterizes the threshold behavior of non-uniform random 2-SAT.

For $k \geq 3$ we were able to prove the existence and asymptotic position of the satisfiability threshold for some ensembles of probability distributions. In order to prove unsatisfiability of instances we used different first moment methods. To prove satisfiablity of instances we restricted formulas in $k$-CNF to formulas in 2-CNF and used our results on the threshold behavior of non-uniform random 2-SAT. We also derived some conditions on the sharpness of the threshold depending on the maximum variable probability in relation to the asymptotic threshold position. However, our results do not completely characterize the threshold behavior for non-uniform random $k$-SAT with $k \geq 3$. There are some ensembles of probability distributions for which we do not know if a satisfiability threshold exists and some for which we know the asymptotic threshold function, but we do not know if the threshold is sharp or coarse. Thus, some straightforward extensions of our work include improved bounds for the asymptotic threshold function and a full characterization of the sharp/coarse-dichotomy for $k \geq 3$. A very ambitious goal might also be to derive the exact threshold function if the satisfiability threshold is sharp. Even for the most well-researched special case random $k$-SAT finding the exact threshold function

up to leading factors is still a challenging open question for $k \geq 3$ up to some very large values.

Regarding the second question, there are results only for a few specific ensembles of probability distributions. For random $k$-SAT the resolution size of unsatisfiable instances sampled around the satisfiability threshold is exponential [BW01; CS88]. Thus, CDCL-based SAT solvers need exponential time to certify unsatisfiability for those instances. We also showed that power-law random $k$-SAT has exponential resolution size around the threshold for power law exponents $\beta > \min(\frac{2k-2}{k-2}, 3)$ [Blä+21]. These exponential lower bounds suggest that a power law distribution alone is not enough to explain the effectiveness of CDCL on industrial instances. However, that does not rule out the existence of a distribution which fits this role better. Therefore, showing general bounds on the resolution size of non-uniform random $k$-SAT depending on the ensemble of probability distributions is still an important future work.

Although we only show rigorous lower bounds in [Blä+21], for smaller power law exponents the resolution size seems to scale exponentially in $n^x$ with $x$ slowly increasing from zero to one for increasing $\beta > \frac{2k-1}{k-1}$. If this was indeed the case, power-law random $k$-SAT would be a good model to randomly generate formulas in $k$-CNF with a certain resolution size as benchmarks for SAT solvers. Instances with the same resolution size can be created with random $k$-SAT as well when the number of clauses is $n^{1+1/2-\varepsilon}$ for constants $\varepsilon \in (0, 1/2)$. However, in power-law random $k$-SAT only a linear number of clauses is necessary and the resolution size can be controlled with the power law exponent $\beta$, i. e. instances generated with power-law random $k$-SAT can be much smaller. Thus, it would be interesting to improve our results on the resolution size of power-law random $k$-SAT and to complement them with upper bounds.

Resolution size is used to measure the hardness of unsatisfiable instances, but what about satisfiable instances? On random $k$-SAT local search solvers usually perform pretty well on satisfiable instances [Bie+09, Chapter 6]. However, they only work for clause-variable ratios below [CHH17; Coj17] or well above the satisfiability threshold [BS15; KP92]. In [Fri+21] we consider satisfiable instances of non-uniform random $k$-SAT and show that a simple local search algorithm finds a satisfying assignment with high probability if the number of clauses is high enough and the probability distributions in the ensemble are not too non-uniform. We actually show this result for a planted equivalent of non-uniform random $k$-SAT, where clauses are drawn in such a way that a random satisfying assignment is guaranteed to exist. However, for the same high enough number of clauses the planted and the original model are so closely related that our results carry over. This work implies that local search is successful for satisfiable instances of power-law and geometric random $k$-SAT with $\Omega(n \log n)$ clauses and generalizes earlier results [BS15; KP92], which showed that the same holds for random $k$-SAT.

In [Blä+21] we also studied if another promising feature could explain the unreasonable effectiveness of state-of-the-art SAT solvers on industrial instances:

the existence of some underlying geometry. The idea is that the Boolean variables have positions in some space, for example Euclidean or hyperbolic space. The random clauses have positions in that space as well and contain $k$ Boolean variables with probabilities depending on their distance to them. This results in some sort of clustering, since Boolean variables that are closer to each other tend to appear together in clauses, a feature that Ansótegui et al. [AGL12] observed in some classes of industrial SAT instances. However, we showed that instances generated with such a model and linear number of clauses are almost always trivially unsatisfiable if the influence of distances on the connection probabilities is high. Since industrial instances are usually not trivially unsatisfiable, this suggests that either geometry alone is not a realistic feature for those instances or the influence of an underlying geometry is only small.

This thesis aimed at analyzing the influence of different frequency distributions for Boolean variables on the satisfiability threshold of $k$-SAT instances. Although our results are incomplete, they showcase some interesting connections between the probabilities of Boolean variables to appear in a random formula and the behavior of the satisfiability threshold. These connections might go unnoticed when only studying random $k$-SAT and its uniform probability distribution. However, the whole point of studying models with prescribed expected frequencies is to see if those frequencies can explain the running time of state-of-the-art solvers on real-world instances. Thus, the next step is to analyze the influence of those distributions on the hardness of solving instances. At least for power law distributions, our related work suggests that the distribution alone might not be sufficient, while the assumption of an underlying geometry might be too strong. Therefore, finding other promising properties of industrial instances which may make them easy for state-of-the-art SAT solvers and ingraining them into realistic models for those instances is still an important task for future work.

# Bibliography


[ABL09a] Carlos Ansótegui, Maria Luisa Bonet, and Jordi Levy. **On the Structure of Industrial SAT Instances**. In: *Proceedings of the 15th International Conference on Principles and Practice of Constraint Programming (CP'2009)*. Vol. 5732. Lecture Notes in Computer Science. Springer, 2009, 127–141. DOI: 10.1007/978-3-642-04244-7_13 (see pages 2, 4).

[ABL09b] Carlos Ansótegui, Maria Luisa Bonet, and Jordi Levy. **Towards Industrial-Like Random SAT Instances**. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'2009)*. 2009, 387–392. URL: http://ijcai.org/Proceedings/09/Papers/072.pdf (see pages 2, 4, 17, 24, 26).

[Ach+01] Dimitris Achlioptas, Lefteris M. Kirousis, Evangelos Kranakis, and Danny Krizanc. **Rigorous results for random (2+p)-SAT**. *Theor. Comput. Sci.* 265:1-2 (2001), 109–129. DOI: 10.1016/S0304-3975(01)00154-2 (see page 4).

[AGL12] Carlos Ansótegui, Jesús Giráldez-Cru, and Jordi Levy. **The Community Structure of SAT Formulas**. In: *Proceedings of the 15th International Conference on Theory and Applications of Satisfiability Testing (SAT'2012)*. Vol. 7317. Springer, 2012, 410–423. DOI: 10.1007/978-3-642-31612-8_31 (see pages 2, 4, 151).

[AL86] Ron Aharoni and Nathan Linial. **Minimal non-two-colorable hypergraphs and minimal unsatisfiable formulas**. *J. Comb. Theory, Ser. A* 43:2 (1986), 196–204. DOI: 10.1016/0097-3165(86)90060-9 (see page 128).

[Ans+15] Carlos Ansótegui, Maria Luisa Bonet, Jesús Giráldez-Cru, and Jordi Levy. **On the Classification of Industrial SAT Families**. In: *Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence (CCIA'2015)*. Vol. 277. Frontiers in Artificial Intelligence and Applications. IOS Press, 2015, 163–172. DOI: 10.3233/978-1-61499-578-4-163 (see pages 2, 4).

[AP04] Dimitris Achlioptas and Yuval Peres. **The threshold for random $k$-SAT is $2^k \log 2 - O(k)$**. *Journal of the American Mathematical Society* 17:4 (2004), 947–973. DOI: 10.1090/S0894-0347-04-00464-3 (see page 14).

[APT79] Bengt Aspvall, Michael F. Plass, and Robert Endre Tarjan. **A Linear-Time Algorithm for Testing the Truth of Certain Quantified Boolean Formulas**. *Information Processing Letters* 8:3 (1979), 121–123. DOI: 10.1016/0020-0190(79)90002-4 (see page 12).

[ASV15] Dan Alistarh, Thomas Sauerwald, and Milan Vojnović. **Lock-free algorithms under stochastic schedulers**. In: *34th ACM Symposium on Principles of Distributed Computing (PODC'2015)*. ACM, 2015, 251–260. DOI: 10.1145/2767386.2767430 (see page 86).

[BC16] Victor Bapst and Amin Coja-Oghlan. **The Condensation Phase Transition in the Regular k-SAT Model**. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016*. Vol. 60. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016, 22:1–22:18. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2016.22 (see page 4).

[Bee06] Peter van Beek. "Backtracking Search Algorithms." In: *Handbook of Constraint Programming*. Vol. 2. Foundations of Artificial Intelligence. Elsevier, 2006, 85–134. DOI: 10.1016/S1574-6526(06)80008-8 (see page 2).

[Bie+09] A. Biere, A. Biere, M. Heule, H. van Maaren, and T. Walsh. **Handbook of Satisfiability: Volume 185 Frontiers in Artificial Intelligence and Applications**. NLD: IOS Press, 2009. ISBN: 1586039296 (see pages 11, 150).

[Blä+21] Thomas Bläsius, Tobias Friedrich, Andreas Göbel, Jordi Levy, and Ralf Rothenberger. **The Impact of Heterogeneity and Geometry on the Proof Complexity of Random Satisfiability**. In: *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'2021)*. SIAM, 2021, 42–53. DOI: 10.1137/1.9781611976465.4 (see pages 24, 150).

[Bou+05] Yacine Boufkhad, Olivier Dubois, Yannet Interian, and Bart Selman. **Regular Random *k*-SAT: Properties of Balanced Formulas**. *J. Autom. Reason.* 35:1-3 (2005), 181–200. DOI: 10.1007/s10817-005-9012-z (see page 4).

[BP14] Milan Bradonjic and Will Perkins. **On Sharp Thresholds in Random Geometric Graphs**. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014*. Vol. 28. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2014, 500–514. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2014.500 (see page 4).

[BS14] Paul Beame and Ashish Sabharwal. **Non-Restarting SAT Solvers with Simple Preprocessing Can Efficiently Simulate Resolution**. In: *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'2014)*. 2014, 2608–2615. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8397 (see page 2).

[BS15] Andrei A. Bulatov and Evgeny S. Skvortsov. **Phase Transition for Local Search on Planted SAT**. In: *40th Intl. Symp. Math. Foundations of Computer Science (MFCS)*. Vol. 9235. Lecture Notes in Computer Science. Springer, 2015, 175–186. DOI: 10.1007/978-3-662-48054-0_15 (see page 150).

[BW01] Eli Ben-Sasson and Avi Wigderson. **Short proofs are narrow - resolution made simple**. *Journal of the ACM* 48:2 (2001), 149–169. DOI: 10.1145/375827.375835 (see page 150).

[CHH17] Amin Coja-Oghlan, Amir Haqshenas, and Samuel Hetterich. **Walksat Stalls Well Below Satisfiability**. *SIAM Journal on Discrete Mathematics* 31:2 (2017), 1160–1173. DOI: 10.1137/16M1084158 (see page 150).

[CIP09] Chris Calabro, Russell Impagliazzo, and Ramamohan Paturi. **The Complexity of Satisfiability of Small Depth Circuits**. In: *4th International Workshop on Parameterized and Exact Computation (IWPEC)*. Vol. 5917. Lecture Notes in Computer Science. Springer, 2009, 75–85. DOI: 10.1007/978-3-642-11269-0_6 (see page 1).

[Coj14] Amin Coja-Oghlan. **The Asymptotic $k$-SAT Threshold**. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC'2014)*. ACM, 2014, 804–813. DOI: 10.1145/2591796.2591822 (see page 5).

[Coj17] Amin Coja-Oghlan. **Belief Propagation Guided Decimation Fails on Random Formulas**. *Journal of the ACM* 63:6 (2017), 49:1–49:55. DOI: 10.1145/3005398 (see page 150).

[Coo71] Stephen A. Cook. **The Complexity of Theorem-Proving Procedures**. In: *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC'1971)*. ACM, 1971, 151–158. DOI: 10.1145/800157.805047 (see pages 1, 12).

[CP16] Amin Coja-Oghlan and Konstantinos Panagiotou. **The asymptotic $k$-SAT threshold**. *Advances in Mathematics* 288 (2016), 985–1068 (see page 5).

[CR92] Vasek Chvátal and Bruce A. Reed. **Mick Gets Some (the Odds Are on His Side)**. In: *Proceedings of the 33rd Annual Symposium on Foundations of Computer Science (FOCS'1992)*. IEEE Computer Society, 1992, 620–627. DOI: 10.1109/SFCS.1992.267789 (see pages 5, 6, 12, 17, 29, 31, 35, 40, 47, 62, 81, 86, 89, 90, 94).

[CS88] Vasek Chvátal and Endre Szemerédi. **Many Hard Examples for Resolution**. *Journal of the ACM* 35:4 (1988), 759–768. DOI: 10.1145/48014.48016 (see pages 2, 150).

[CW18] Amin Coja-Oghlan and Nick Wormald. **The Number of Satisfying Assignments of Random Regular k-SAT Formulas**. *Combinatorics, Probability & Computing* 27:4 (2018), 496–530. DOI: 10.1017/S0963548318000263 (see page 4).

[Día+09] 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 (2009), 2920–2934. DOI: 10.1016/j.tcs.2009.02.020 (see page 5).

[DP09] Devdatt P. Dubhashi and Alessandro Panconesi. **Concentration of Measure for the Analysis of Randomized Algorithms**. Cambridge University Press, 2009. ISBN: 978-0-521-88427-3. URL: http://www.cambridge.org/gb/knowledge/isbn/item2327542/ (see page 10).

[DP60] Martin Davis and Hilary Putnam. **A Computing Procedure for Quantification Theory**. *Journal of the ACM* 7:3 (1960), 201–215. DOI: 10.1145/321033.321034 (see page 2).

[DSS15] Jian Ding, Allan Sly, and Nike Sun. **Proof of the Satisfiability Conjecture for Large K**. In: *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC'2015)*. ACM, 2015, 59–68. DOI: 10.1145/2746539.2746619 (see pages 5, 17).

[FKG71] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. **Correlation inequalities on some partially ordered sets**. *Communications in Mathematical Physics* 22:2 (June 1971), 89–103. ISSN: 1432-0916. DOI: 10.1007/BF01651330 (see pages 130, 131).

[FR18] Tobias Friedrich and Ralf Rothenberger. **Sharpness of the Satisfiability Threshold for Non-uniform Random k-SAT**. In: *Proceedings of the 21st International Conference on Theory and Applications of Satisfiability Testing (SAT'2018)*. Vol. 10929. Lecture Notes in Computer Science. Springer-Verlag, 2018, 273–291. DOI: 10.1007/978-3-319-94144-8_17 (see pages 11, 99).

[FR19] Tobias Friedrich and Ralf Rothenberger. **The Satisfiability Threshold for Non-Uniform Random 2-SAT**. In: *Proceedings of the 46th International Colloquium on Automata, Languages and Programming (ICALP'2019)*. Vol. 132. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, 61:1–61:14. DOI: 10.4230/LIPIcs.ICALP.2019.61 (see pages 11, 29).

[Fri+17a] 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 the 25th Annual European Symposium on Algorithms (ESA'2017)*. Vol. 87. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, 37:1–37:15. DOI: 10.4230/LIPIcs.ESA.2017.37 (see pages 11, 85, 89).

[Fri+17b] Tobias Friedrich, Anton Krohmer, Ralf Rothenberger, and Andrew M. Sutton. **Phase Transitions for Scale-Free SAT Formulas**. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'2017)*. AAAI Press, 2017, 3893–3899. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14755 (see pages 11, 29).

[Fri+21] Tobias Friedrich, Frank Neumann, Ralf Rothenberger, and Andrew M. Sutton. **Solving Non-uniform Planted and Filtered Random SAT Formulas Greedily**. In: *Proceedings of the 24th International Conference on Theory and Applications of Satisfiability Testing (SAT'2021)*. Vol. 12831. Lecture Notes in Computer Science. Springer, 2021, 188–206. DOI: 10.1007/978-3-030-80223-3_13 (see page 150).

[Fri05] Ehud Friedgut. **Hunting for sharp thresholds**. *Random Structures & Algorithms* 26:1-2 (2005), 37–51. DOI: 10.1002/rsa.20042 (see page 12).

[Fri99] Ehud Friedgut. **Sharp thresholds of graph properties, and the *k*-SAT problem**. *Journal of the American Mathematical Society* 12:4 (1999), 1017–1054. DOI: 10.1090/S0894-0347-99-00305-7 (see pages 6, 12, 15, 99, 115, 116, 119, 123, 124, 130, 136).

[GJ79] M. R. Garey and David S. Johnson. **Computers and Intractability: A Guide to the Theory of NP-Completeness**. W. H. Freeman, 1979. ISBN: 0-7167-1044-7 (see page 1).

[GL15] Jesús Giráldez-Cru and Jordi Levy. **A Modularity-Based Random SAT Instances Generator**. In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'2015)*. AAAI Press, 2015, 1952–1958. URL: http://ijcai.org/Abstract/15/277 (see page 4).

[GL17] Jesús Giráldez-Cru and Jordi Levy. **Locality in Random SAT Instances**. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'2017)*. 2017, 638–644. DOI: 10.24963/ijcai.2017/89 (see page 4).

[Goe96] Andreas Goerdt. **A Threshold for Unsatisfiability**. *Journal of Computer and System Sciences* 53:3 (1996), 469–486. DOI: 10.1006/jcss.1996.0081 (see pages 5, 31).

[Han+19] Thomas Dueholm Hansen, Haim Kaplan, Or Zamir, and Uri Zwick. **Faster *k*-SAT algorithms using biased-PPSZ**. In: *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC'2019)*. ACM, 2019, 578–589. DOI: 10.1145/3313276.3316359 (see page 12).

[HS03] Mohammad Taghi Hajiaghayi and Gregory B. Sorkin. **The Satisfiability Threshold of Random 3-SAT is at Least 3.52**. Tech. rep. RC22942. IBM, Oct. 2003 (see page 5).

[IP99] Russell Impagliazzo and Ramamohan Paturi. **Complexity of k-SAT**. In: *14th Annual IEEE Conference on Computational Complexity*. IEEE Computer Society, 1999, 237–240. DOI: 10.1109/CCC.1999.766282 (see page 1).

[Jan96] Svante Janson. **The Second Moment Method, Conditioning and Approximation**. In: *Random Discrete Structures*. New York, NY: Springer New York, 1996, 175–183. ISBN: 978-1-4612-0719-1 (see page 10).

[JS97] Roberto J. Bayardo Jr. and Robert Schrag. **Using CSP Look-Back Techniques to Solve Real-World SAT Instances**. In: *Proceedings of the 14th International Conference on Theory and Applications of Satisfiability Testing (SAT'1997)*. AAAI Press / The MIT Press, 1997, 203–208. URL: http://www.aaai.org/Library/AAAI/1997/aaai97-032.php (see page 2).

[Kar72] Richard M. Karp. **Reducibility Among Combinatorial Problems**. In: *Proceedings of a symposium on the Complexity of Computer Computations*. The IBM Research Symposia Series. Plenum Press, New York, 1972, 85–103. DOI: 10.1007/978-1-4684-2001-2_9 (see pages 1, 12).

[Kir+98] Lefteris M Kirousis, Evangelos Kranakis, Danny Krizanc, and Yannis C Stamatiou. **Approximating the unsatisfiability threshold of random formulas**. *Random Structures & Algorithms* 12:3 (1998), 253–269. DOI: 10.1002/(SICI)1098-2418(199805)12:3\<253::AID-RSA3\>3.0.CO;2-U (see pages 14, 85, 87).

[KKL06] Alexis C. Kaporis, Lefteris M. Kirousis, and Efthimios G. Lalas. **The probabilistic analysis of a greedy satisfiability algorithm**. *Random Structures & Algorithms* 28:4 (2006), 444–480. DOI: 10.1002/rsa.20104 (see page 5).

[KP92] Elias Koutsoupias and Christos H. Papadimitriou. **On the Greedy Algorithm for Satisfiability**. *Information Processing Letters* 43:1 (1992), 53–55. DOI: 10.1016/0020-0190(92)90029-U (see page 150).

[Lev73] L. A. Levin. **Universal problems of full search.** Russian. *Probl. Peredachi Inf.* 9:3 (1973), 115–116. ISSN: 0555-2923 (see pages 1, 12).

[LMS11] Daniel Lokshtanov, Dániel Marx, and Saket Saurabh. **Lower bounds based on the Exponential Time Hypothesis**. *Bulletin of the EATCS* 105 (2011), 41–72. URL: http://eatcs.org/beatcs/index.php/beatcs/article/view/92 (see page 1).

[McD92] Colin McDiarmid. **On a correlation inequality of Farr**. *Comb. Probab. Comput.* 1 (1992), 157–160. DOI: 10.1017/S096354830000016X (see page 88).

[MFS16] Nathan Mull, Daniel J. Fremont, and Sanjit A. Seshia. **On the Hardness of SAT with Community Structure**. In: *Proceedings of the 19th International Conference on Theory and Applications of Satisfiability Testing (SAT'2016)*. Vol. 9710. Lecture Notes in Computer Science. Springer, 2016, 141–159. DOI: 10.1007/978-3-319-40970-2_10 (see page 4).

[Mon+96] Rémi Monasson, Riccardo Zecchina, Scott Kirkpatric, Bart Selman, and Lidror Troyansky. **Phase transition and search cost in the 2+p-sat problem**. *4th Workshop on Physics and Computation* (1996) (see page 4).

[Mon+99] Rémi Monasson, Riccardo Zecchina, Scott Kirkpatrick, Bart Selman, and Lidror Troyansky. **2+p-SAT: Relation of typical-case complexity to the nature of the phase transition**. *Random Structures & Algorithms* 15:3-4 (1999), 414–435. DOI: 10.1002/(SICI)1098-2418(199910/12)15:3/4\<414::AID-RSA10\>3.0.CO;2-G (see page 4).

[MPZ02] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. **Analytic and algorithmic solution of random satisfiability problems**. *Science* 297:5582 (2002), 812–815. DOI: 10.1126/science.1073287 (see page 5).

[MR99] Rajeev Motwani and Prabhakar Raghavan. **Randomized Algorithms**. Chapman & Hall/CRC Applied Algorithms and Data Structures series. CRC Press, 1999. DOI: 10.1201/9781420049503-c16 (see page 72).

[MSL92] David G. Mitchell, Bart Selman, and Hector J. Levesque. **Hard and Easy Distributions of SAT Problems**. In: *Proceedings of the 10 AAAI Conference on Artificial Intelligence (AAAI'1992)*. AAAI Press / The MIT Press, 1992, 459–465. URL: http://www.aaai.org/Library/AAAI/1992/aaai92-071.php (see page 2).

[MU05] Michael Mitzenmacher and Eli Upfal. **Probability and Computing: Randomized Algorithms and Probabilistic Analysis**. Cambridge University Press, 2005. ISBN: 978-0-521-83540-4. DOI: 10.1017/CBO9780511813603 (see pages 8–10).

[Mül17] Tobias Müller. **The critical probability for confetti percolation equals 1/2**. *Random Structures & Algorithms* 50:4 (2017), 679–697. DOI: 10.1002/rsa.20675 (see page 124).

[MZ97] Rémi Monasson and Riccardo Zecchina. **Statistical mechanics of the random *K*-satisfiability model**. *Phys. Rev. E* 56 (2 Aug. 1997), 1357–1370 (see page 4).

[ODo14] Ryan O'Donnell. **Analysis of Boolean Functions**. Cambridge University Press, 2014. ISBN: 978-1-10-703832-5. URL: http://www.cambridge.org/de/academic/subjects/computer-science/algorithmics-complexity-computer-algebra-and-computational-g/analysis-boolean-functions (see pages 99, 115, 118, 120, 123).

[ODo21] Ryan O'Donnell. **Analysis of Boolean Functions**. *CoRR* abs/2105.10386 (2021). arXiv: 2105.10386. URL: https://arxiv.org/abs/2105.10386 (see page 124).

[PD11] Knot Pipatsrisawat and Adnan Darwiche. **On the power of clause-learning SAT solvers as resolution engines**. *Artificial Intelligence* 175:2 (2011), 512–525. DOI: 10.1016/j.artint.2010.10.002 (see page 2).

[Rat+10] Vishwambhar Rathi, Erik Aurell, Lars K. Rasmussen, and Mikael Skoglund. **Bounds on Threshold of Regular Random *k*-SAT**. In: *Proceedings of the 13 International Conference on Theory and Applications of Satisfiability Testing (SAT'2010)*. Vol. 6175. Lecture Notes in Computer Science. Springer, 2010, 264–277. DOI: 10.1007/978-3-642-14186-7_22 (see page 4).

[SS96] João P. Marques Silva and Karem A. Sakallah. **GRASP - a new search algorithm for satisfiability**. In: *International Conference on Computer-Aided Design, ICCAD*. IEEE, 1996, 220–227. DOI: 10.1109/ICCAD.1996.569607 (see page 2).

[Sze03] Stefan Szeider. **On Fixed-Parameter Tractable Parameterizations of SAT**. In: *Proceedings of the 6th International Conference on Theory and Applications of Satisfiability Testing (SAT'2013)*. Vol. 2919. Lecture Notes in Computer Science. Springer, 2003, 188–202. DOI: 10.1007/978-3-540-24605-3_15 (see page 12).

[Tse83] G. S. Tseitin. **On the Complexity of Derivation in Propositional Calculus**. In: *Automation of Reasoning: 2: Classical Papers on Computational Logic 1967–1970*. Berlin, Heidelberg: Springer Berlin/Heidelberg, 1983, 466–483. ISBN: 978-3-642-81955-1. DOI: 10.1007/978-3-642-81955-1_28 (see page 12).

# List of Publications

## Articles in Refereed Journals

[1] **Greed is Good for Deterministic Scale-Free Networks**. *Algorithmica* 82:11 (2020), 3338–3389. DOI: 10.1007/s00453-020-00729-z. Joint work with Ankit Chauhan and Tobias Friedrich.

[2] **Routing for on-street parking search using probabilistic data**. *AI Communications* 32:2 (2019), 113–124. DOI: 10.3233/AIC-180574. Joint work with Tobias Friedrich, Martin S. Krejca, Tobias Arndt, Danijar Hafner, Thomas Kellermeier, Simon Krogmann, and Armin Razmjou.

## Articles in Refereed Conference Proceedings

[3] **Probabilistic Routing for On-Street Parking Search**. In: *Proceedings of the 24th Annual European Symposium on Algorithms (ESA'2016)*. Vol. 57. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016, 6:1–6:13. DOI: 10.4230/LIPIcs.ESA.2016.6. Joint work with Tobias Arndt, Danijar Hafner, Thomas Kellermeier, Simon Krogmann, Armin Razmjou, Martin S. Krejca, and Tobias Friedrich.

[4] **Memory-Restricted Routing with Tiled Map Data**. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC'2018)*. IEEE, 2018, 3347–3354. DOI: 10.1109/SMC.2018.00567. Joint work with Thomas Bläsius, Jan Eube, Thomas Feldtkeller, Tobias Friedrich, Martin S. Krejca, J. A. Gregor Lagodzinski, Julius Severin, Fabian Sommer, and Justin Trautmann.

[5] **The Impact of Heterogeneity and Geometry on the Proof Complexity of Random Satisfiability**. In: *Symposium on Discrete Algorithms (SODA)*. SIAM, 2021, 42–53. DOI: 10.1137/1.9781611976465.4. Joint work with Thomas Bläsius, Tobias Friedrich, Andreas Göbel, and Jordi Levy.

[6] **Ultra-Fast Load Balancing on Scale-Free Networks**. In: *Proceedings of the 42nd International Colloquium on Automata, Languages and Programming (ICALP'2015)*. Vol. 9135. Lecture Notes in Computer Science. Springer-Verlag, 2015, 516–527. DOI: 10.1007/978-3-662-47666-6_41. Joint work with Karl Bringmann, Tobias Friedrich, Martin Hoefer, and Thomas Sauerwald.

[7] **Greed is Good for Deterministic Scale-Free Networks**. In: *36th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS'2016)*. Vol. 65. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016, 33:1–33:15. DOI: 10.4230/LIPIcs.FSTTCS.2016.33. Joint work with Ankit Chauhan and Tobias Friedrich.

[8] **Phase Transitions for Scale-Free SAT Formulas**. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'2017)*. AAAI Press, 2017, 3893–3899. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14755. Joint work with Tobias Friedrich, Anton Krohmer, and Andrew M. Sutton.

[9] **Solving Non-uniform Planted and Filtered Random SAT Formulas Greedily**. In: *Proceedings of the 24th International Conference on Theory and Applications of Satisfiability Testing (SAT'2021)*. Ed. by Chu-Min Li and Felip Manyà. Vol. 12831. Lecture Notes in Computer Science. Springer, 2021, 188–206. DOI: 10.1007/978-3-030-80223-3_13. Joint work with Tobias Friedrich, Frank Neumann, and Andrew M. Sutton.

[10] **Bounds on the Satisfiability Threshold for Power Law Distributed Random SAT**. In: *Proceedings of the 25th Annual European Symposium on Algorithms (ESA'2017)*. Vol. 87. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, 37:1–37:15. DOI: 10.4230/LIPIcs.ESA.2017.37. Joint work with Tobias Friedrich, Anton Krohmer, Thomas Sauerwald, and Andrew M. Sutton.

[11] **Greedy Maximization of Functions with Bounded Curvature under Partition Matroid Constraints**. In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'2019)*. AAAI Press, 2019, 2272–2279. DOI: 10.1609/aaai.v33i01.33012272. Joint work with Tobias Friedrich, Andreas Göbel, Frank Neumann, and Francesco Quinzan.

[12] **Sharpness of the Satisfiability Threshold for Non-uniform Random k-SAT**. In: *Proceedings of the 21st International Conference on Theory and Applications of Satisfiability Testing (SAT'2018)*. Vol. 10929. Lecture Notes in Computer Science. **Best Paper Award**. Springer-Verlag, 2018, 273–291. DOI: 10.1007/978-3-319-94144-8_17. Joint work with Tobias Friedrich.

[13] **Sharpness of the Satisfiability Threshold for Non-Uniform Random k-SAT**. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'2019)*. International Joint Conferences on Artificial Intelligence Organization, 2019, 6151–6155. DOI: 10.24963/ijcai.2019/853. Joint work with Tobias Friedrich.

[14] **The Satisfiability Threshold for Non-Uniform Random 2-SAT**. In: *Proceedings of the 46th International Colloquium on Automata, Languages and Programming (ICALP'2019)*. Vol. 132. LIPIcs. Schloss Dagstuhl

- Leibniz-Zentrum für Informatik, 2019, 61:1–61:14. DOI: 10.4230/LIPIcs.ICALP.2019.61. Joint work with Tobias Friedrich.

[15] **Dominating an s-t-Cut in a Network**. In: *Proceedings of the 41st Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'2015)*. Vol. 8939. Lecture Notes in Computer Science. Springer-Verlag, 2015, 401–411. DOI: 10.1007/978-3-662-46078-8_33. Joint work with Sascha Grau and Michael Rossberg.

[16] **Mixed Integer Programming versus Evolutionary Computation for Optimizing a Hard Real-World Staff Assignment Problem**. In: *Proceedings of the29th International Conference on Automated Planning and Scheduling (ICAPS'2019)*. AAAI Press, 2019, 541–554. URL: https://aaai.org/ojs/index.php/ICAPS/article/view/3521. Joint work with Jannik Peters, Daniel Stephan, Isabel Amon, Hans Gawendowicz, Julius Lischeid, Lennart Salabarria, Jonas Umland, Felix Werner, Martin S. Krejca, Timo Kötzing, and Tobias Friedrich.