The decimation process in random k-SAT

Amin Coja-Oghlan and Angelica Y. Pachon-Pinzon*

University of Warwick, Mathematics and Computer Science, Coventry CV4 7AL, UK
{a.coja-oghlan,a.y.pachon-pinzon}@warwick.ac.uk

Abstract. Let $\boldsymbol{\Phi}$ be a uniformly distributed random k-SAT formula with n variables and m clauses. Non-rigorous statistical mechanics ideas have inspired a message passing algorithm called *Belief propagation guided decimation* for finding satisfying assignments of $\boldsymbol{\Phi}$. This algorithm can be viewed as an attempt at implementing a certain thought experiment that we call the *decimation process*. In this paper we identify a variety of phase transitions in the decimation process and link these phase transitions to the performance of the algorithm.
Key words: random structures, phase transitions, k-SAT, Belief Propagation.
Submission to track A.

1 Introduction

Let $k \geq 3$ and $n > 1$ be integers, let $r > 0$ be a real, and set $m = \lceil rn \rceil$. Let $\boldsymbol{\Phi} = \boldsymbol{\Phi}_k(n, m)$ be a propositional formula obtained by choosing a set of m clauses of length k over the variables $V = \{x_1, \ldots, x_n\}$ uniformly at random. For k, r fixed we say that $\boldsymbol{\Phi}$ has some property $\mathcal{P}$ *with high probability* ('w.h.p.') if $\lim_{n\to\infty} \mathrm{P}[\boldsymbol{\Phi} \in \mathcal{P}] = 1$. The interest in random k-SAT originates from the experimental observation that for certain densities r the random formula $\boldsymbol{\Phi}$ is satisfiable w.h.p. while a large class of algorithms, including and particularly the workhorses of practical SAT solving such as sophisticated DPLL-based solvers, fail to find a satisfying assignment efficiently [18]. Over the past decade, a fundamentally new class of algorithms has been proposed on the basis of ideas from statistical physics [6, 17]. Experiments performed for $k = 3, 4, 5$ indicate that these new 'message passing algorithms', namely *Belief Propagation guided decimation* and *Survey Propagation guided decimation* ('BP/SP decimation'), excel on random k-SAT instances [14]. Indeed, the experiments indicate that BP/SP decimation find satisfying assignments for r close to the threshold where $\boldsymbol{\Phi}$ becomes unsatisfiable w.h.p.

These algorithms are specifically inspired by a generic technique from statistical mechanics called the *cavity method*, which was used in [6, 15] to study the structure of the set $\mathcal{S}(\boldsymbol{\Phi})$ of satisfying assignments (or, more accurately, properties of the Gibbs measure) of the random formula $\boldsymbol{\Phi}$. The results obtained in these (non-rigorous) works identify phase transitions solely in terms of the formula density r. Furthermore, in base to them, it was hypothesized that (certain versions of) BP decimation should find satisfying assignments up to $r \sim \frac{2^k \ln k}{k}$ or even up to $r \sim 2^k \ln 2$ [15]. The argument given for the latter scenario in [15] is that the key obstacle for BP is the *condensation* phase, in which, the set of solutions is dominated by a few large cluster (with strongly fluctuating sizes). In terms of the parameter r, the condensation threshold was (non-rigorously) estimated to occur at $r = 2^k \ln 2 - \frac{3}{2} \ln 2$.

Since random k-SAT instances have widely been deemed extremely challenging benchmarks, the stellar experimental performance of the physicists' message passing algorithms has stirred considerable excitement. However, the statistical mechanics ideas that BP/SP decimation are based on are highly non-rigorous, and thus a rigorous analysis of these message passing algorithms is an important but challenging open problem.

A first step was made in [8], where it was shown that BP decimation does not outperform far simpler combinatorial algorithms for sufficiently large clause lengths k. More precisely, the main result of [8] is that there is a constant $\rho_0 > 0$ (independent of k) such that the 'vanilla' version of BP decimation fails to find satisfying assignments w.h.p. if $r > \rho_0 2^k/k$. The explanation for this discrepancy is that [6, 15] neglect the time parameter $\theta = 1 - t/n$ of the *decimation process*, an idealized thought experiment that the BP decimation algorithm aims to implement.

The analysis performed in [8] is based on an intricate method for directly tracking the execution of BP decimation. Unfortunately this argument does little to illuminate the conceptual reasons for the algorithms' demise. In particular, [8] does not provide a link to the statistical mechanics ideas that inspired the algorithm.

The present paper aims to remedy these defects. Here we study the *decimation process* and show that this experiment undergoes a variety of phase transitions that explain the failure of BP decimation for densities $r > \rho_0 \cdot 2^k/k$. Our results identify phase transitions jointly in terms of the clause/variable density r and with respect to the time parameter of the decimation process θ. The latter dimension was ignored in the original statistical mechanics work on BP [6, 17] but turns out to have a crucial impact on the performance of the algorithm. As Theorem 7

* Supported by EPSRC grant EP/G039070/2.

shows, even for *fixed* $\rho \geq \rho_0$ (independent of k) condensation occurs as the decimation process proceeds to θ in the regime (3). This means that decimating variables has a similar effect on the geometry of the set of satisfying assignments as increasing the clause/variable density.
On a non-rigorous basis an analysis both in terms of the formula density r and the time parameter θ was carried out in [20]. Thus, our results can be viewed as a rigorous version of [20] (with proofs based on completely different techniques).
The results of this paper can also be seen as a generalization of the ones obtained in [1], where is rigorously proved a substantial part of the results hypothesized in [15] on shattering and rigidity in terms of the clause/variable density r; this improved prior work [2, 5, 9]. The new aspect here is that we identify not only a transition for shattering/rigidity, but also for condensation and forcing in terms of *both* the density r and the time parameter θ of the decimation process (Theorems 5–7). Technically, that is showed upon the methods developed in [1] and new arguments needed to accommodate the time parameter θ to prove the statements on the marginals of the variables and the condensation phenomenon (Theorem 7). In addition, Theorem 10 confirms rigorously that for r, θ in the condensation phase, BP does not yield the correct marginals.
The present results have no immediate bearing on the conceptually more sophisticated SP decimation algorithm. However, we conjecture that SP undergoes a similar sequence of phase transitions and that the algorithm will not find satisfying assignments for densities $\rho \geq \rho_0$, with ρ_0 a certain constant independent of k.

1.1 The decimation process

BP decimation is a polynomial-time algorithm that aims to (heuristically) implement the 'thought experiment' shown in Experiment 1 [19, 20], which we call the *decimation process*.[1]

Experiment 1 ('decimation process'). *Input:* A satisfiable k-CNF Φ.
Result: A satisfying assignment $\sigma : V \to \{0, 1\}$ (with $0/1$ representing 'false'/'true').

0. Let $\Phi_0 = \Phi$.
1. For $t = 1, \ldots, n$ do
2. Compute the fraction $M_{x_t}(\Phi_{t-1})$ of all satisfying assignments of Φ_{t-1} in which the variable x_t takes the value 1.
3. Assign $\sigma(x_t) = 1$ with probability $M_{x_t}(\Phi_{t-1})$, and let $\sigma(x_t) = 0$ otherwise.
4. Obtain the formula Φ_t from Φ_{t-1} by substituting the value $\sigma(x_t)$ for x_t and simplifying (i.e., delete all clauses that got satisfied by assigning x_t, and omit x_t from all other clauses).
5. Return the assignment σ.

A moment's reflection reveals that, given a satisfiable input formula Φ, the decimation process outputs a uniform sample from the set of all satisfying assignments of Φ. The obvious obstacle to actually implementing this experiment is the computation of the marginal probability $M_{x_t}(\Phi_{t-1})$ that x_t takes the value 'true' in a random satisfying assignment of Φ_{t-1}, a $\#P$-hard problem in the worst case. Yet the key hypothesis underlying BP decimation is that these marginals can be computed efficiently on *random* formulas by means of a message passing algorithm. We will return to the discussion of BP decimation and its connection to Experiment 1 below.
In this work we are going to study the decimation process when applied to a random formula $\boldsymbol{\Phi}$ for densities $r < 2^k \ln 2 - k$ [3, 4], i.e., in the regime where $\boldsymbol{\Phi}$ is satisfiable w.h.p. More precisely, conditioning on $\boldsymbol{\Phi}$ being satisfiable, we let $\boldsymbol{\Phi}_t$ be the (random) formula obtained after running the first t iterations of Experiment 1. The variable set of this formula is $V_t = \{x_{t+1}, \ldots, x_n\}$, and each clause of $\boldsymbol{\Phi}_t$ consists of *at most* k literals. Let $\mathcal{S}(\boldsymbol{\Phi}_t) \subset \{0, 1\}^{V_t}$ be the set of all satisfying assignments of $\boldsymbol{\Phi}_t$ and $\boldsymbol{\sigma}_t$ a 'random' satisfying assignment of $\boldsymbol{\Phi}_t$ obtained by the following experiment.

D1. Generate a random formula $\boldsymbol{\Phi}$, conditioned on $\boldsymbol{\Phi}$ being satisfiable.
D2. Run the decimation process for t steps to obtain $\boldsymbol{\Phi}_t$.
D3. Choose a satisfying assignment $\boldsymbol{\sigma}_t \in \mathcal{S}(\boldsymbol{\Phi}_t)$ uniformly at random.
D4. The result is the pair $(\boldsymbol{\Phi}_t, \boldsymbol{\sigma}_t)$.

[1] Several different versions of BP decimation have been suggested. In this paper we refer to the simplest but arguably most natural one, also considered in [8, 19, 20]. Other versions decimate the variables in a different order, allowing for slightly better experimental results [6, 14].

We will identify various phase transition that the formulas $\boldsymbol{\Phi}_t$ undergo as t grows from 1 to n. As it turns out, these can be characterized via two simple parameters. The first one is the clauses density $r \sim m/n$. Actually, it will be most convenient to work in terms of

$$\rho = kr/2^k,$$

so that $m/n \sim \rho \cdot 2^k/k$. We will be interested in the regime $\rho_0 \leq \rho \leq k \ln 2$, where ρ_0 is a constant (independent of k). The upper bound $k \ln 2$ marks the point where satisfying assignments cease to exist [4]. The second parameter is the fraction

$$\theta = 1 - t/n$$

of 'free' variables (i.e., variables not yet assigned by time t). We say that *almost all* $\sigma_t \in \mathcal{S}(\boldsymbol{\Phi}_t)$ have a certain property $\mathcal{A}$ if $|\mathcal{A} \cap \mathcal{S}(\boldsymbol{\Phi}_t)| = (1 - o(1))|\mathcal{S}(\boldsymbol{\Phi}_t)|$.

1.2 Notation and results

To formalize the concept of 'typical' satisfying assignment of $\boldsymbol{\Phi}_t$, we define the probability distribution $\mathcal{U}_t = \mathcal{U}_t(k, n, m)$ induced by the following experiment, which is equivalent to **D1**–**D4**:

U1. Generate a random formula $\boldsymbol{\Phi}$, conditioned on $\boldsymbol{\Phi}$ being satisfiable.
U2. Choose $\boldsymbol{\sigma} \in \mathcal{S}(\boldsymbol{\Phi})$ uniformly at random.
U3. Substitute $\boldsymbol{\sigma}(x_i)$ for x_i for $1 \leq i \leq t$ and simplify to obtain a formula $\boldsymbol{\Phi}_t$.
U4. The result is the pair $(\boldsymbol{\Phi}_t, \boldsymbol{\sigma}_t)$, where $\boldsymbol{\sigma}_t : V_t \to \{0,1\}\,,\ x \mapsto \boldsymbol{\sigma}(x)$.

Definition 2. *Let Φ be a k-CNF on V, Φ_t a the formula obtained after t steps of the decimation process, $1 \leq t < n$, and suppose that $\sigma \in \mathcal{S}(\Phi_t)$. Let $dist(\cdot,\cdot)$ denotes the Hamming distance.*

1. *We call $x \in V_t$* loose *in (Φ_t, σ) if there is $\tau \in \mathcal{S}(\Phi_t)$ such that*

$$\tau(x) \neq \sigma(x) \text{ and } dist(\sigma, \tau) \leq \ln n.$$

2. *Given an integer $\omega \geq 1$, we say that $x \in V_t$ is* ω-rigid *in (Φ_t, σ) if for any $\tau \in \mathcal{S}(\Phi_t)$ with $\tau(x) \neq \sigma(x)$ we have*

$$dist(\sigma, \tau) \geq \omega.$$

3. *Finally, $x \in V_t$ is* forced *in (Φ_t, σ) if x occurs in a unit clauses of Φ_t.*

Definition 3. *We say that a set $S \subset \{0,1\}^{V_t}$ is* (α, β)-shattered *if it admits a decomposition $S = \bigcup_{i=1}^{N} R_i$ into pairwise disjoint subsets such that the following two conditions are satisfied.*

SH1. *We have $|R_i| \leq \exp(-\alpha\theta n)|S|$ for all $1 \leq i \leq N$.*
SH2. *If $1 \leq i < j \leq N$ and $\sigma \in R_i$, $\tau \in R_j$, then $dist(\sigma, \tau) \geq \beta\theta n$.*

Definition 4. *Let $\alpha > 0$. We say that a set $S \subset \{0,1\}^{\theta n}$ is* α-condensed *if for any $\sigma, \tau \in S$ we have $dist(\sigma, \tau) \leq \alpha n$.*

Denote by $M_x(\Phi_t)$ the marginal probability that x takes the value 'true' in a random satisfying assignment of Φ_t, for any $x \in V_t$, i.e.,

$$M_x(\Phi_t) = \frac{|\{\sigma \in \mathcal{S}(\Phi_t) : \sigma(x) = 1\}|}{|\mathcal{S}(\Phi_t)|}.$$

The transition in terms of the density ρ is well described in [1, 2, 15]. Here we keep $\rho < k \ln 2$, the values for which a random formula has solution, and study the decimation process in terms of the time parameter θ. Figure 1 illustrate the changes of the set $\mathcal{S}(\boldsymbol{\Phi}_t)$ when ρ and θ grow. Let us explain them.

The symmetric phase. In the early stages of the decimation process, while θ is 'big' (1), the correlations amongst the variables are mostly local. Thus, if we 'flip' one variable in σ, then we can 'repair' the unsatisfied clauses that this may cause by simply flipping another $\ln n$ variables. Finally, as the average distance between satisfying assignments is large on average, the set $\mathcal{S}(\boldsymbol{\Phi}_t)$ is 'well spread' over the Hamming cube $\{0,1\}^{V_t}$.

Shattering and rigidity. For θ in the regime (2), in most satisfying $\sigma \in \mathcal{S}(\boldsymbol{\Phi}_t)$ the values assigned to 99% of the variables are linked via long-range correlations, i.e., to 'repair' the damage done by flipping a single rigid variable it is inevitable to reassign a *constant fraction* of all variables. The set $\mathcal{S}(\boldsymbol{\Phi}_t)$ decomposes into exponentially many exponentially tiny subsets, which are mutually separated by a linear Hamming distance $\Omega(n)$, and remains 'well spread' over the Hamming cube $\{0,1\}^{V_t}$.

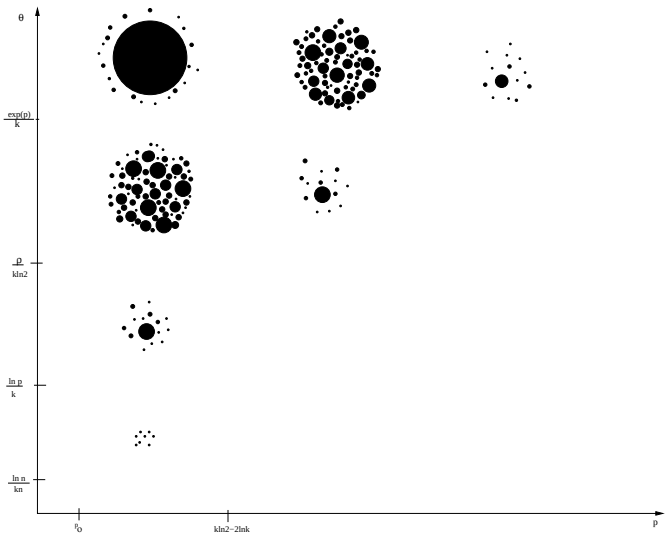

Fig. 1. ρ, θ-Phase transition.

The condensation phase. As the decimation process progresses to a point that θ satisfies (3), the set $\mathcal{S}(\boldsymbol{\Phi}_t)$ shrinks into a condensed subset of $\{0,1\}^{V_t}$ of tiny diameter, in contrast to a well-spread shattered set as in Theorem 6. Furthermore, most marginals $M_x(\boldsymbol{\Phi}_t)$ are either extremely close to 0 or extremely close to 1. In fact, there is a large set R of variables on which all satisfying assignments virtually agree (more precisely: any two can't disagree on more than $k2^{-k}n$ variables in R).

The forced phase. In the last stages of the decimation process, while θ is in (4), most of the variables are force, i.e., almost all clauses still unsatisfied are unit. Moreover, as the set $\mathcal{S}(\boldsymbol{\Phi}_t)$ is still condense, the geometry of the space of solutions is a disconnected subset of tiny diameter.

Now, we formulate these results in Theorems-5- 8.

Theorem 5. *There are constants $k_0, \rho_0 > 0$ such that for $k \geq k_0$, $\rho_0 \leq \rho \leq k\ln 2 - 2\ln k$, and*

$$k \cdot \theta > \exp\left[\rho\left(1 + \frac{\ln(\ln\rho + 10)}{\rho}\right)\right] \tag{1}$$

the random formula $\boldsymbol{\Phi}_t$ has the following properties w.h.p.

1. *In almost all satisfying assignments $\sigma \in \mathcal{S}(\boldsymbol{\Phi}_t)$ at least $0.99\theta n$ variables are loose.*
2. *At least $\theta n/3$ variables $x \in V_t$ satisfy $M_x(\boldsymbol{\Phi}_t) \in [0.01, 0.99]$.*
3. *W.h.p. almost all the satisfying assignments in $\mathcal{S}(\boldsymbol{\Phi}_t)$ are such that the distance between each two of them is at least $0.49\theta n$, i.e.,*

$$|\tau \in \mathcal{S}(\boldsymbol{\Phi}_t) : dist(\tau, \sigma) > 0.49\theta n| > (1 - exp(\Omega(n)))|\mathcal{S}(\boldsymbol{\Phi}_t)|.$$

Theorem 6. *There are constants $k_0, \rho_0 > 0$ such that for $k \geq k_0$, $\rho_0 \leq \rho \leq k\ln 2 - 2\ln k$, and*

$$\frac{\rho}{\ln 2}(1 + 2\rho^{-2}) \leq k\theta \leq \exp\left[\rho\left(1 - \frac{\ln\rho}{\rho} - \frac{2}{\rho}\right)\right] \tag{2}$$

the random formula $\boldsymbol{\Phi}_t$ has the following properties w.h.p.

1. *In almost all $\sigma \in \mathcal{S}(\boldsymbol{\Phi}_t)$ at least $0.99\theta n$ variables are $\Omega(n)$-rigid.*
2. *There exist $\alpha = \alpha(k,\rho) > 0, \beta = \beta(k,\rho) > 0$ such that $\mathcal{S}(\boldsymbol{\Phi}_t)$ is (α, β)-shattered.*
3. *At least $\theta n/3$ variables $x \in V_t$ satisfy $M_x(\boldsymbol{\Phi}_t) \in [0.01, 0.99]$.*
4. *W.h.p. almost all the satisfying assignments in $\mathcal{S}(\boldsymbol{\Phi}_t)$ are such that the distance between each two of them is at least $0.49\theta n$, i.e.,*

$$|\tau \in \mathcal{S}(\boldsymbol{\Phi}_t) : dist(\tau, \sigma) > 0.49\theta n| > (1 - exp(\Omega(n)))|\mathcal{S}(\boldsymbol{\Phi}_t)|.$$

Theorem 7. *There are constants $k_0, \rho_0 > 0$ such that for $k \geq k_0$, $\rho_0 \leq \rho \leq k \ln 2 - 2 \ln k$, and*

$$\ln \rho < k \cdot \theta < (1 - \rho^{-2}) \cdot \rho / (\ln 2) \tag{3}$$

the random formula $\boldsymbol{\Phi}_t$ has the following properties w.h.p.

1. *In almost all $\sigma \in \mathcal{S}(\boldsymbol{\Phi}_t)$ at least $0.99\theta n$ variables are $\Omega(n)$-rigid.*
2. *The set $\mathcal{S}(\boldsymbol{\Phi}_t)$ is $\exp(2-\rho)/k$-condensed.*
3. *At least $0.99\theta n$ variables $x \in V_t$ satisfy $M_x(\boldsymbol{\Phi}_t) \in \left[0, 2^{-k/2}\right] \cup \left[1 - 2^{-k/2}, 1\right]$.*
4. *There is a set $R \subset V_t$ of size $|R| \geq 0.99\theta n$ such that for any $\sigma, \tau \in \mathcal{S}(\boldsymbol{\Phi}_t)$ we have*

$$|\{x \in R : \sigma(x) \neq \tau(x)\}| \leq k 2^{-k} n.$$

Theorem 8. *There are constants $k_0, \rho_0 > 0$ such that for $k \geq k_0$, $\rho_0 \leq \rho \leq k \ln 2 - 2 \ln k$, and*

$$\ln(n)/n \ll k \cdot \theta < \ln(\rho)(1 - 10/\ln \rho) \tag{4}$$

the random formula $\boldsymbol{\Phi}_t$ has the following properties w.h.p.

1. *At least $0.99\theta n$ variables are forced.*
2. *The set $\mathcal{S}(\boldsymbol{\Phi}_t)$ is $\exp(2-\rho)/k$-condensed.*

Belief Propagation. As mentioned earlier, the BP decimation algorithm is an attempt at implementing the decimation process by means of an efficient algorithm. The key issue with this is the computation of the marginals $M_{x_t}(\Phi_{t-1})$ in step 2 of the decimation process. Indeed, the problem of computing these marginals is $\#P$-hard in the worst case. Thus, instead of working with the 'true' marginals, BP decimation uses certain numbers $\mu_{x_t}(\Phi_{t-1}, \omega)$ that can be computed efficiently, where $\omega \geq 1$ is an integer parameter. The precise definition of the $\mu_{x_t}(\Phi_{t-1}, \omega)$ can be found in [6]. Basically, they are the result of a 'local' dynamic programming algorithm ('Belief Propagation') that depends upon the assumption of a certain correlation decay property. For given k, ρ, the key hypothesis underpinning the BP decimation algorithm is

Hypothesis 9. *For any $\varepsilon > 0$ there is $\omega = \omega(\varepsilon, k, \rho, n) \geq 1$ such that w.h.p. for all $1 \leq t \leq n$ we have $|\mu_{x_t}(\boldsymbol{\Phi}_{t-1}, \omega) - M_{x_t}(\boldsymbol{\Phi}_{t-1})| < \varepsilon$.*

The Hypothesis 9 states that throughout the decimation process, the 'BP marginals' $\mu_{x_t}(\Phi_{t-1}, \omega)$ are a good approximation to the true marginals $M_{x_t}(\Phi_{t-1})$.
As we mentioned before, in [8] was proved that BP decimation fails to find satisfying assignments w.h.p. if $r > \rho_0 2^k/k$, for some $\rho_0 > 0$. However, this result does not show why that happen. The next result shows rigorously that in terms of ρ, θ, in the θ-condensation phase BP decimation does not yield the correct marginals. This result explain how the decimation process itself affect for the algorithm's demise.

Theorem 10. *There exist constants $c_0, k_0, \rho_0 > 0$ such that for all $k \geq k_0$, and $\rho_0 \leq \rho \leq k \ln 2 - 2 \ln k$ the following is true for any integer $\omega = \omega(k, \rho, n) \geq 1$. Suppose that*

$$c_0 \ln(\rho) < k \cdot \theta < \rho / \ln 2. \tag{5}$$

Then for at least $0.99\theta n$ variables $x \in V_t$ we have $\mu_x(\boldsymbol{\Phi}_t, \omega) \in [0.49, 0.51]$.

Comparing Theorem 7 with Theorem 10, we see that w.h.p. for θ satisfying (5) most of the 'true' marginals $M_x(\boldsymbol{\Phi}_t)$ are very close to either 0 or 1, whereas the 'BP marginals' lie in $[0.49, 0.51]$. Thus, in the regime described by (5) the BP marginals do *not* provide a good approximation to the actual marginals.

Corollary 11. *There exist constants $c_0, k_0, \rho_0 > 0$ such that for all $k \geq k_0$, $\rho_0 \leq \rho \leq k \ln 2 - 3 \ln k$ Hypothesis 9 is untrue.*

2 Summary and discussion.

Fix $k \geq k_0$ and $\rho \geq \rho_0$. Theorems 5–8 show how the space of satisfying assignments of $\boldsymbol{\Phi}_t$ evolves as the decimation process progresses. In the *symmetric phase* $k\theta \geq \exp((1+o_\rho(1))\rho)$ where there still is a large number of free variables, the correlations amongst the free variables are purely local ('loose variables'). As the number of free variables enters the regime $(1+o_\rho(1))\rho/\ln 2 \leq k\theta \leq \exp((1-o_\rho(1))\rho)$, the set $\mathcal{S}(\boldsymbol{\Phi}_t)$ of satisfying assignments shatters into exponentially many tiny 'clusters', each of which comprises only an exponentially small fraction of all satisfying assignments. Most satisfying assignments exhibit long-range correlations amongst the possible values that can be assigned to the individual variables ('rigid variables'). This phenomenon goes by the name of *dynamic replica symmetry breaking* in statistical mechanics [15].

While in the previous phases the set of satisfying assignments is scattered all over the Hamming cube (as witnessed by the average Hamming distance of two satisfying assignments), in the *condensation phase* $(1-o_\rho(1))\ln\rho \leq k\theta \leq (1-o_\rho(1))\rho/\ln 2$ the set of satisfying assignments has a tiny diameter. This is mirrored by the fact that the marginals of most variables are extremely close to either 0 or 1. Furthermore, in (most of) this phase the estimates of the marginals resulting from Belief Propagation are off (Theorem 10). As part 4 of Theorem 7 shows, the mistaken estimates of the Belief Propagation computation would make it impossible for BP decimation to penetrate the condensation phase. More precisely, even if BP decimation would emulate the decimation process perfectly up until the condensation phase commences, with probability $1-\exp(-\Omega(n))$ BP decimation would then assign at least $k2^{-k}n$ variables in the set R from part 4 of Theorem 7 'wrongly' (i.e., differently than they are assigned in any satisfying assignment). In effect, BP decimation would fail to find a satisfying assignment, regardless of its subsequent decisions. Finally, in the forced phase $k\theta \leq (1-o_\rho(1))\ln\rho$ there is an abundance of unit clauses that make it easy to read off the values of most variables. However, getting stuck in the condensation phase, BP decimation won't reach this regime.

These results suggest that the reason for the failure of BP decimation is the existence of the condensation phase. Intuitively, in the condensation phase the marginals are governed by genuinely global phenomena (essentially expansion properties) that elude the inherently local BP computation. By contrast, it is conceivable that BP does indeed yield the correct marginals in the previous phases. Verifying or falsifying this remains an important open problem.

3 Preliminares

3.1 Analyzing the decimation process

We perform some groundwork to facilitate a rigorous analysis of the decimation process. An analysis of **U1–U4** seems difficult because of **U2**: it is unclear how to analyze (or implement) this step directly. Following [1], we will surmount this problem by considering yet another experiment.

P1. Choose an assignment $\boldsymbol{\sigma}' \in \{0,1\}^V$ uniformly at random.
P2. Choose a formula $\boldsymbol{\Phi}'$ with m clauses that is satisfied by $\boldsymbol{\sigma}'$ uniformly at random.
P3. Substitute $\boldsymbol{\sigma}'(x_i)$ for x_i for $1 \leq i \leq t$ and simplify to obtain a formula $\boldsymbol{\Phi}'_t$.
P4. The result is the pair $(\boldsymbol{\Phi}'_t, \boldsymbol{\sigma}'_t)$, where $\boldsymbol{\sigma}'_t : V_t \to \{0,1\},\ x \mapsto \boldsymbol{\sigma}'(x)$.

The experiment **P1–P4** is easy to implement and, in effect, also amenable to a rigorous analysis. For given the assignment $\boldsymbol{\sigma}'$, there are $(2^k-1)\binom{n}{k}$ clauses in total that evaluate to 'true' under $\boldsymbol{\sigma}'$, and to generate $\boldsymbol{\Phi}'$ we merely choose m out of these uniformly and independently. Unfortunately, it is *not* true that the experiment **P1–P4** is equivalent to **U1–U4**. However, we will employ a result from [1] that establishes a connection between these two experiments that is strong enough to extend many results from **P1–P4** to **U1–U4**.

To state this result, observe that **P1–P4** and **U1–U4** essentially only differ in their first two steps. Thus, let $\Lambda_k(n,m)$ denote the set of all pairs (Φ,σ), where Φ is a k-CNF on $V = \{x_1,\ldots,x_n\}$ with m clauses, and $\sigma \in \mathcal{S}(\Phi)$. Let $\mathcal{U}_k(n,m)$ denote the probability distribution induced on $\Lambda_k(n,m)$ by **U1–U2**, and let $\mathcal{P}_k(n,m)$ signify the probability distribution induced by **P1–P2**; this distribution is sometimes called the *planted model*.

The key point to connect **P1–P2** with **U1–U2** is the following: if the number of satisfying assignment of a random formula, i.e., $|\mathcal{S}(\boldsymbol{\Phi})|$ is by an exponential factor smaller that the expectation, then, these estimates allows us to approximate the uniform model by the planted model. This result is obtained in [1](Lemma 22, Theorem 8) and given subsequently.

Theorem 12 ([1]). *Let (Φ,σ) be a pair chosen from the experiment* **U1–U2**. *Suppose $k \geq 4$ and $0 < \rho < k\ln 2 - k^2/2^k$, then w.h.p.*

$$\frac{1}{n}\ln|\mathcal{S}(\Phi)| \geq \ln 2 + r\ln(1-2^{-k}) - \frac{0.99\rho n}{2^k}. \tag{6}$$

Theorem 13 ([1]). *Suppose $k \geq 4$, $0 < \rho < k\ln 2 - k^2/2^k$ and $|\mathcal{S}(\Phi)|$ satisfies (6). Let $\mathcal{E} \subset \Lambda_k(n,m)$. If $\mathrm{P}_{\mathcal{P}_k(n,m)}[\mathcal{E}] \geq 1-\exp(-\rho n/2^k)$ then $\mathrm{P}_{\mathcal{U}_k(n,m)}[\mathcal{E}] = 1-o(1)$.*

To establish Theorems 5-8 we need to connect **P1–P4** with **U1–U4**. This result is obtained just using the following consequence of Theorem 12 and following the proof of Theorem 13. The proof of Theorem 13 is given in [1].

Corollary 14. *Let (Φ_t, σ_t) be a pair chosen from the experiment* **U1–U4***, with $1 \leq t \leq n$, then w.h.p.*

$$\frac{1}{n}\ln|\mathcal{S}(\Phi_t)| \geq (1-t/n)\ln 2 + r\ln(1-2^{-k}) - \frac{\rho n}{2^k}. \tag{7}$$

Proof. Let Φ be a formula such that $\frac{1}{n}\ln|\mathcal{S}(\Phi)| \geq \ln 2 + r\ln(1-2^{-k}) - \rho n/2^k$. By Theorem 12 the random formula Φ has this property w.h.p. Thus, it suffices to show that for a random $\sigma \in \mathcal{S}(\Phi)$ the bound (7) holds w.h.p. To this end, let $\mathcal{I} = \{0,1\}^t$. Moreover, for each $\sigma \in \{0,1\}^n$ let $\sigma|_t$ be the vector $(\sigma(x_1),\ldots,\sigma(x_t)) \in \mathcal{I}$. For each $\sigma_* \in \mathcal{I}$ let $Z(\sigma_*)$ be the number of assignments $\sigma \in \mathcal{S}(\Phi)$ such that $\sigma|_t = \sigma_*$. If $\sigma \in \mathcal{S}(\Phi)$ is chosen uniformly at random, then for any $\sigma_* \in \mathcal{I}$ we have

$$\mathrm{P}[\sigma|_t = \sigma_*] = Z(\sigma_*)/Z, \text{ where } Z = \sum_{\tau\in\mathcal{I}} Z(\tau) = |\mathcal{S}(\Phi)|.$$

Let $\xi > 0$ be a sufficiently small number and let

$$q = \mathrm{P}[Z(\sigma|_t) < \exp(-t\ln 2 - \xi n)\cdot Z],$$

where $\sigma \in \mathcal{S}(\Phi)$ is chosen uniformly at random. Then

$$q \leq \sum_{\sigma_*\in\mathcal{I}: Z(\sigma_*)\leq \frac{Z}{\exp(\xi n + t\ln 2)}} Z(\sigma_*)/Z \leq \frac{2^t}{Z}\cdot\frac{Z}{\exp(\xi n + t\ln 2)} \leq \exp(-\xi n),$$

whence the assertion follows. □

We also need the following variant of the planted model.

P1'. Choose an assignment $\boldsymbol{\sigma}' \in \{0,1\}^V$ uniformly at random.
P2'. Choose a formula $\boldsymbol{\Phi}'$ by including each of the $(2^k-1)\binom{n}{k}$ possible clauses that are satisfied under $\boldsymbol{\sigma}'$ with probability $p = m/((2^k-1)\binom{n}{k})$ independently.
P3'. Substitute $\boldsymbol{\sigma}'(x_i)$ for x_i for $1 \leq i \leq t$ and simplify to obtain a formula $\boldsymbol{\Phi}'_t$.
P4'. The result is the pair $(\boldsymbol{\Phi}'_t, \boldsymbol{\sigma}'_t)$, where $\boldsymbol{\sigma}'_t : V_t \to \{0,1\},\ x \mapsto \boldsymbol{\sigma}'(x)$.

Steps **P1'–P2'** of this experiment induce a probability distribution $\mathcal{P}'_k(n,m)$ on formula/assignment pairs. The following corollary establishes a connection between this distribution and the distribution $\mathcal{U}_k(n,m)$.

Corollary 15 ([1]). *Suppose $k \geq 4$ and $0 < \rho < k\ln 2 - k^2/2^k$. Let $\mathcal{E}$ be any property of formula/assignment pairs. If $\mathrm{P}_{\mathcal{P}'_k(n,m)}[\mathcal{E}] \geq 1-\exp(-\rho n/2^k)$ then $\mathrm{P}_{\mathcal{U}_k(n,m)}[\mathcal{E}] = 1-o(1)$.*

Its proof only use that under $\mathcal{P}'_k(n,m)$, the total number of clauses equals m is $\Theta(m^{-1/2})$. Furthermore, the conditional distribution $\mathcal{P}'_k(n,m)$ given the total number of clauses is m, is identical to $\mathcal{P}_k(n,m)$. Therefore, the assertion follows from Theorem 12 and 13. The same result is obtained to connect **P1'–P4'** with **U1–U4**, the assertion follows from Theorem 13 and Corollary 14.
We will need the following elementary observation about the distribution $\mathcal{P}'_k(n,m)$.

Lemma 16. *Let (Φ,σ) be a pair chosen from the distribution $\mathcal{P}'_k(n,m)$.*

1. *For each literal l that is true under σ, the number of clauses supported by l, S_l, is binomially distributed $\mathrm{Bin}(\frac{k}{n}\cdot\binom{n}{k}, m/((2^k-1)\binom{n}{k}))$ with mean $\mu = \frac{kr}{2^k-1}$.*
2. *For any integer D, the number of true literals l under σ that support fewer than D clauses, NL^D, is also binomially distributed $\mathrm{Bin}(n,q)$, where $q = P(S_l < D)$.*

Proof. Without loss of generality we may condition on σ assigning the value true to all variables. Thus, for any true literal l under σ, the number of all possible clauses in which l is the only positive literal is equal to $\binom{n-1}{k-1} = \frac{k}{n} \cdot \binom{n}{k}$. As each of the clauses is included in Φ with probability $p = m/((2^k-1)\binom{n}{k})$ independently, then S_l has a binomial distribution $\mathrm{Bin}(\frac{k}{n} \cdot \binom{n}{k}, p)$. This establishes 1.
Now define the variable

$$Y_l = \begin{cases} 1 & \text{if the number of clauses supported by } l \text{ is fewer than } D \\ 0 & \text{otherwise,} \end{cases}$$

which has Bernoulli distribution, $Ber(q)$. As the number of true literals under σ are n, then $NL^D = \sum_{l=1}^{n} Y_l$ and is distributed $\mathrm{Bin}(n, q)$. □

3.2 Factor graph and tame variables

There is a natural way to associate a bipartite graph with a k-CNF Φ, known as the *factor graph*. Its vertices are the variables and the clauses of Φ, and each clause is adjacent to all the variables it contains. For a variable x we let $N_{2\omega}(x)$ be the subgraph of that is spanned by all vertices at distance at most 2ω from x. (where, of course, 'distance' just refers to the length of a shortest path in the factor graph). A variable x is *tame* if $N_{2\omega}(x)$ is acyclic and contains no more than $\ln(n)$ variables. To prove the item concern to loose variables (Theorem 5, part 1), we need the following well-known fact about random k-CNFs.

Proposition 17. *Suppose that $k \geq 3$ and $0 < r \leq 2^k \ln 2 - k$. W.h.p. all but $o(n)$ variables are tame in Φ.*

Proof. Assume that $k \geq 3$, $0 < r \leq 2^k \ln(k)/k$ and $\omega = \lceil \ln \ln \ln n \rceil$. Proposition 17 follows from the next two lemmas directly. □

Lemma 18. *W.h.p. for all but $o(n)$ variables the subgraph $N_{2\omega}(x)$ of the factor graph contains at most $\ln(n)$ variables.*

Proof. Fix a variable x and let L_j be the set of variables that have distance $2j$ from x in the factor graph. We are going to prove that $\mathrm{P}[\sum_{j=0}^{\omega} |L_j| \geq \ln n] = o(1)$. Then the assertion follows from Markov's inequality. Let $\mathcal{F}_j$ be the coarsest σ-algebra in which all events $\{v \in L_i\}$ with $v \in V$ and $0 \leq i \leq j$ are measurable. For each $x \in L_{j+1}$ there is a clause C that contains both x and a variable from L_j but no variable from L_i for $i < j$. Let Y_{j+1} be the number of such clauses. We claim that conditional on $\mathcal{F}_j$ the variable Y_{j+1} is stochastically dominated by a binomial variable $\mathrm{Bin}(m, k|L_j|/(n - \sum_{i<j} |L_i|))$. For the probability that a random clause conditional on not containing a variable from $\bigcup_{i<j} L_i$ contains a variable from L_j is bounded by $k|L_j|/(n - \sum_{i<j} |L_i|)$ (as the probability that any of the k slots contains a variable from L_j is $|L_j|/(n - \sum_{i<j} |L_i|)$. Furthermore, $|L_{j+1}| \leq (k-1)Y_{j+1}$. Therefore, $|L_{j+1}|$ given $\mathcal{F}_j$ is stochastically dominated by a scalar multiple $Z_j = (k-1) \cdot \mathrm{Bin}(m, k|L_j|/(n - \sum_{i<j} |L_i|))$ of a binomial random variable. If $\sum_{j=0}^{\omega} |L_j| \geq \ln n$, then there is $0 \leq j < \omega$ such that $|L_{j+1}|/|L_j| \geq (\ln n)^{1/2\omega}$. Moreover, if j_0 is an index such that $\sum_{i<j_0} |L_i| \leq \ln n$, then by Markov's inequality

$$\begin{aligned} \mathrm{P}\left[|L_{j_0+1}|/|L_{j_0}| \geq (\ln n)^{1/\omega} | \mathcal{F}_{j_0}\right] &\leq \mathrm{P}\left[Z_{j_0+1} \geq |L_{j_0}| \cdot (\ln n)^{1/2\omega}\right] \\ &\leq \frac{\mathrm{E}Z_{j_0+1}}{|L_{j_0}| \cdot (\ln n)^{1/2\omega}} \leq \frac{k(k-1)m}{n - \sum_{i<j_0} |L_i|} \cdot (\ln n)^{-1/2\omega} \\ &\leq \frac{k(k-1)m}{(n - \ln n)} \cdot (\ln n)^{-1/2\omega} \leq 2k^2 r \cdot (\ln n)^{-1/2\omega}. \end{aligned}$$

Applying the union bound over indices j_0, we conclude that

$$\mathrm{P}\left[\sum_{j=0}^{\omega} |L_j| \geq \ln n\right] \leq 2k^2 r \cdot (\ln n)^{-1/2\omega} \cdot \omega = o(1),$$

thereby completing the proof. □

Lemma 19. *W.h.p. the subgraph $N_{2\omega}(x)$ of the factor graph of Φ is acyclic if the number of variables in $N_{2\omega}(x)$ is at most $\ln n$.*

Proof. Observe that any cycle of length $2l$ corresponds to a sequence $C_1, \dots, C_l$ of clauses such that for all index pairs j_1, j_2 such that either $1 \le j_1 < l$ and $j_2 = j_1 + 1$ or $j_1 = 1$ and $j_2 = l$ the clauses C_{j_1} and C_{j_2} share a variable.
For each $x_i \in N_{2\omega}(x)$, let X_l^i be the number of cycles of length $2l$ such that, the clauses C_1 and C_l share the variable x_i, then the number of cycles in $N_{2\omega}(x)$ is given by $Y = \sum_{1 \le l \le \omega} \sum_{i \in N_{2\omega}(x)} X_l^i$ and we have

$$\{Y \ge 1\} \subseteq \{ \sum_{1 \le l \le \omega} \sum_{i \in N_{2\omega}(x)} X_l^i \ge |N_{2\omega}(x)| \}.$$

On the other hand, the probability that two independent random clauses C, C' share a variable is bounded by k^2/n (as there are k ways to choose a variable in C and k slots where this variable can appear in C', and because the probability that a given variable appears in any particular slot of C' is $1/n$). Let X_l be the number of cycles of length $2l$, i.e., $\sum_{i \in N_{2\omega}(x)} X_l^i$, Therefore, $\mathrm{E}X_l \le \left(mk^2/n\right)^l = (k^2 r)^l$. In effect, the expected number of cycles of length at most 2ω is bounded by

$$\mathrm{E} \sum_{1 \le l \le \omega} X_l \le \sum_{1 \le l \le \omega} (k^2 r)^l = o(\ln n).$$

Thus, by Markov's inequality $P(Y = 0) \ge 1 - o(1)$ if $|N_{2\omega}(x)| < \ln n$. □

Finally, to complete some of the following proofs, we use *Chernoff bound* on the tails of a binomially distributed random variable X with mean λ (see [12, pages 26–28]): For any $t > 0$

$$\mathrm{P}(X \ge \lambda + t) \le \exp\left(-t \cdot \varphi(t/\lambda)\right) \text{ and } \mathrm{P}(X \le \lambda - t) \le \exp\left(-t \cdot \varphi(-t/\lambda)\right), \tag{8}$$

where

$$\varphi(x) = (1 + x)\ln(1 + x) - x. \tag{9}$$

4 Overview

We describe the main results of this paper, i.e., Theorems 5–8, according to the various phases that the decimation process passes through. But to prove this results, it is necessary to proceed in a different order. To facilitate this, we will state the main results in the order in which the proofs proceed. We begin with the statements on the loose/rigid and forced variables. Recall that $V_t = \{x_{t+1}, \dots, x_n\}$ and let $L_t = \{x_{t+1}, \bar{x}_{t+1}, \dots, x_n, \bar{x}_n\}$ be the set of literals.

Theorem 20. *There exist constants $k_0, \rho_0 > 0$ such that for all $k \ge k_0$ and $\rho_0 \le \rho \le k \ln 2 - 2 \ln k$ the following three statements hold for a random pair $(\boldsymbol{\Phi}_t, \sigma_t)$ chosen from the experiment* **U1–U4** *w.h.p.*

1. *If $k\theta > \exp\left[\rho\left(1 + \frac{\ln\ln\rho}{\rho} + \frac{10}{\rho}\right)\right]$, then at least $0.99\theta n$ variables $x \in V_t$ are loose w.h.p.*
2. *If $1 < k\theta < \exp\left[\rho\left(1 - \frac{3\ln\rho}{\rho}\right)\right]$, then at least $\rho^3 \exp(-\rho)\theta n$ variables $x \in V_t$ are $\Omega(n)$-rigid w.h.p.*
3. *If $\ln(n)/n < \theta < (\ln(\rho) - 10)/k$, then at least $0.99\theta n$ variables are forced w.h.p.*

The second type of statements concern the global structure of the set of satisfying assignments, summarize in the following theorem.

Theorem 21. *There exist constants $k_0, \rho_0 > 0$ such that for all $k \ge k_0$, and $\rho_0 \le \rho \le k \ln 2 - 2 \ln k$ the following three statements hold.*

1. *If*
$$\frac{\rho}{\ln 2}(1 + \rho^{-2} + 2^{2-k}) \le k\theta \le \exp\left[\rho\left(1 - \frac{\ln\rho}{\rho} - \frac{2}{\rho}\right)\right]$$
then $\mathcal{S}(\boldsymbol{\Phi}_t)$ is $(\exp(2 - \rho) - \varepsilon, \exp(2 - \rho) + \varepsilon)$-shattered w.h.p. for some $\varepsilon = \varepsilon(k, \rho) > 0$.
2. *If $\theta < (\rho - 1/\rho)/(k \ln 2)$, then $\mathcal{S}(\boldsymbol{\Phi}_t)$ is $\exp(2 - \rho)$-condensed w.h.p.*
3. *If $\theta > \rho(1 + 2/\rho^2)/(k \ln 2)$, then, for almost all the satisfying assignments in $\mathcal{S}(\boldsymbol{\Phi}_t)$, the distance between two random elements is at least $0.49\theta n$ w.h.p.*

The next theorem contains the statements about the marginals of the truth values of individual variables.

Theorem 22. *There exist constants $k_0, \rho_0 > 0$ such that for all $k \geq k_0$, and $\rho_0 \leq \rho \leq k\ln 2 - 2\ln k$ the following two statements hold.*

1. *If $\theta \geq \frac{\rho}{k\ln 2}(1 + 1/\rho^2 + k/2^{k-2})$, then w.h.p. for at least $\theta n/3$ variables $x \in V_t$ we have*
$$M_x(\boldsymbol{\Phi}_t) \in [0.01, 0.99]\,.$$
2. *If $\ln(n)/n < \theta < \rho(1 - 1/\rho^2)/(k\ln 2)$, then w.h.p. for all but $\exp(-\rho)\theta n$ variables $x \in V_t$ we have*
$$M_x(\boldsymbol{\Phi}_t) \in \left[0, 2^{-k/2}\right] \cup \left[1 - 2^{-k/2}, 1\right].$$

Theorems 5–8 follow directly from Theorems 20–22 by reordering the individual statements according to the phases they appear in.
The statements about loose, rigid and force variables are build upon the techniques developed in [1]. For the type of statement on the global structure of the set of satisfying assignments, shattering and condensation, we adapt arguments from [1, 2, 9] to the situation where we have the *two* parameters θ, ρ (rather than just ρ). Finally, the statements about the marginals of the truth values of individual variables will be proved using the bounds for the distance of two satisfying assignments, in combination with a double-counting argument.

5 Proof of Theorem 20

5.1 Loose variables

Let σ be a satisfying assignment of a k-CNF Φ. Remember that a literal l *supports* a clause C of Φ if l is the only literal in C that is true under σ. Moreover, we say that a literal l is 1-*loose* if it is true under σ and supports no clause. In addition, l is 2-*loose* if l is true under σ and each clause that l supports contains a 1-loose literal from L_t. Thus, any 1-loose literal is 2-loose as well. We also need to establish the following.

Proposition 23. *Suppose that $\theta \geq 3\exp(\rho)(\ln\rho + 10)/k$ and $r \leq 2^k \ln 2 - k$. Let (Φ_t, σ_t) be a random pair chosen from the experiment* **U1–U4**. *Then there are at least $0.999\theta n$ 2-loose literals in L_t w.h.p.*

To prove Proposition 23, we start by estimating the number of 1-loose variables.

Lemma 24. *Suppose that $\theta \geq \exp(\rho)/k$ and $\rho \leq k\ln 2$. Let (Φ_t, σ_t) be a random pair chosen from the experiment* **P1'–P4'**. *With probability at least $1 - \exp(-k2^{2-k}n)$ the number of 1-loose in L_t is at least $\theta n \cdot \exp(-\rho)/2$.*

Proof. By Lemma 16 the number X of 1-loose literals in L_t has a binomial distribution with mean

$$\begin{aligned}\mathrm{E}X &\leq \theta n \cdot \mathrm{P}\left[\mathrm{Bin}\left(\frac{k}{n}\cdot\binom{n}{k}, \frac{m}{(2^k-1)\binom{n}{k}}\right) = 0\right]\\ &= \theta n \cdot \left(1 - \frac{m}{(2^k-1)\binom{n}{k}}\right)^{\frac{k}{n}\cdot\binom{n}{k}} \sim \theta n \cdot \exp\left(-\frac{kr}{2^k-1}\right) = \theta n\exp(-\rho - \rho/(2^k-1)).\end{aligned}$$

As $\theta \geq \exp(\rho)/k$ and $\rho \leq k\ln 2$, the Chernoff bound (8) shows that for large enough k

$$\mathrm{P}\left[X < \theta n\exp(-\rho)/2\right] \leq \exp\left[-\frac{\theta n}{8\exp(\rho)}\right] \leq \exp(-k2^{2-k}n),$$

as desired. □

Lemma 25. *Suppose that $\theta \geq 3\exp(\rho)(\ln\rho + 10)/k$, $\rho \geq \rho_0$ with ρ_0 as in Lemma 24, and that k is sufficiently large. Let (Φ_t, σ_t) be a random pair chosen from the experiment* **P1'–P4'**. *Then with probability at least $1 - \exp(-k2^{1-k}n)$ the number of 2-loose literals in L_t is at least $0.999\theta n$.*

Proof. To simplify the notation, we are going to condition on σ being the all-true assignment; this is without loss of generality. For each variable $x \in V_t$ we let S_x be the number of clauses supported by x. Moreover, let $S = \sum_{x \in V_t} S_x$ and let X be the number of variables $x \in V_t$ such that $S_x = 0$. Thus, X equals the number of 1-loose variables. Let $\mathcal{E}$ be the event that $X \geq \theta n \exp(-\rho)/2$ and $S \leq 2\rho\theta n$. Since the number of possible clauses with precisely one positive literal in L_t is $\theta n\binom{n-1}{k-1}$, S has a binomial distribution $\mathrm{Bin}[\theta n\binom{n-1}{k-1}, m/((2^k-1)\binom{n}{k})]$. Therefore, Lemma 24 implies that

$$\begin{aligned}\mathrm{P}\left[\neg\mathcal{E}\right] &\leq \mathrm{P}\left[X < \theta n\exp(-\rho)/2\right] + \mathrm{P}\left[S > 2\rho\theta n\right] \\ &\leq \exp\left[-k2^{2-k}n\right] + \mathrm{P}\left[\mathrm{Bin}\left(\theta n\binom{n-1}{k-1}, \frac{m}{(2^k-1)\binom{n}{k}}\right) > 2\rho\theta n\right]. \end{aligned} \tag{10}$$

We have

$$\theta n\binom{n-1}{k-1} \cdot \frac{m}{(2^k-1)\binom{n}{k}} \leq \frac{2^k}{2^k-1} \cdot \rho\theta n.$$

Hence, combining (10) with the Chernoff bound (8), we obtain for sufficiently large k

$$\mathrm{P}\left[\neg\mathcal{E}\right] \leq \exp\left[-k2^{2-k}n\right] + \exp\left[-0.99\rho\theta n\right] \leq 2\exp\left[-k2^{2-k}n\right], \tag{11}$$

where in the last step we used the assumption that $\rho \geq \rho_0$ for a fixed constant $\rho_0 > 0$. Let us now condition on the event that $S = s$ for some number $s \leq 2\rho\theta n$, and on the event $\mathcal{E}$. In this conditional distribution for each of the s clauses supported by some variable in V_t the $k-1$ negative literals that the clause contains are independently uniformly distributed. Therefore, for each such clause the number of negative literals $\bar{y}$ whose underlying variable y is 1-loose is binomially distributed $\mathrm{Bin}(k-1, X/n)$. Consequently, the number T of clauses supported by some variable in V_t in which no 1-loose variable occurs negatively has a binomial distribution with mean $s \cdot \mathrm{P}\left[\mathrm{Bin}(k-1, X/n) = 0\right]$. Hence,

$$\begin{aligned}\mathrm{E}\left[T|\mathcal{E}\right] &\leq 2\rho\theta n \cdot \mathrm{P}\left[\mathrm{Bin}(k-1, \theta\exp(-\rho)/2) = 0\right] \\ &= 2\rho\theta n \cdot (1-\theta\exp(-\rho)/2)^{k-1} \leq 2\rho\theta n\exp(-\theta\exp(-\rho)k/3) \leq 2\exp(-10)\theta n.\end{aligned}$$

Thus, the Chernoff bound (8) implies that for $k \geq k_0$ large enough

$$\mathrm{P}\left[T > 0.001\theta n|\mathcal{E}\right] \leq \exp(-0.001\theta n) \leq \exp\left[-k2^{2-k}n\right]. \tag{12}$$

Finally, the assertion follows from (11) and (12). □

Proof (Proposition 23). Let $\mathcal{E}$ be the event that a pair (Φ_t, σ_t) chosen from the experiment **U1**–**U4** has at least $0.999\theta n$ 2-loose literals. Lemma 25 shows that

$$\mathrm{P}_{\mathcal{P}'_k(n,m)}\left[\mathcal{E}\right] \geq 1 - \exp(-k2^{1-k}n) \geq 1 - \exp(-krn/4^k). \tag{13}$$

Moreover, Corollary 15 and (13) imply that $\mathrm{P}_{\mathcal{U}_k(n,m)}\left[\mathcal{E}\right] = 1 - o(1)$ as desired. □

Proof (Theorem 20, part 1). Let (Φ_t, σ_t) be a pair chosen from the experiment **U1**–**U4**. Without loss of generality we may condition on σ being the all-true assignment. Let $\mathcal{L}$ be the set of all tame variables that are 2-loose. By Propositions 17 and 23 we have $\mathcal{L} \geq (0.999 - o(1))\theta n \geq 0.99\theta n$ w.h.p. Assuming that this is the case, we are going to show that if $x \in \mathcal{L}$, then there is a satisfying assignment τ such that $\tau(x) \neq \sigma(x)$ and $\mathrm{dist}(\tau, \sigma) \leq \ln(n)$. Thus, fix a variable $x \in \mathcal{L}$. If x is 1-loose, then we can just set $\tau(x) = 1 - \sigma(x) = 0$ and $\tau(y) = \sigma(y) = 1$ for all $y \neq x$ to obtain a satisfying assignment with $\mathrm{dist}(\tau, \sigma) = 1$, because x does not support any clauses. Hence, assume that x is 2-loose but not 1-loose. Let $\mathcal{C}$ be the set of all clauses supported by x in (Φ_t, σ_t). By definition of 2-loose any clause $C \in \mathcal{C}$ contains a negative occurrence of a 1-loose variable $x_C \in V_t$. Define $\tau(x) = 0$, $\tau(x_C) = 0$ for all $C \in \mathcal{C}$, and $\tau(y) = \sigma(y) = 1$ for all other variables y.

We claim that τ is a satisfying assignment. To see this, assume for contradiction that there is a clause U that is unsatisfied under τ. Then U contains a variable from $\{x\} \cup \{x_C : C \in \mathcal{C}\}$ positively, while none of these variables occurs negatively in U. Hence, $U \notin \mathcal{C}$. Moreover, since the variables x_C, $C \in \mathcal{C}$, do not support any clauses, U indeed contains two variables from the set $\{x\} \cup \{x_C : C \in \mathcal{C}\}$ positively. There are two possible cases.

Case 1: x occurs in U. Let $C \in \mathcal{C}$ such that x_C occurs in U as well. Then the factor graph contains the cycle x, C, x_C, U, x, in contradiction to our assumption that x is tame.

Case 2: x does not occur in U. There exist $C_1, C_2 \in \mathcal{C}$ such that x_{C_1}, x_{C_2} occur in C. Hence, the factor graph contains the cycle $x, C_1, x_{C_1}, C, x_{C_2}, C_2, x$, once more in contradiction to the assumption that x is tame.

Hence, there is no clause U that is unsatisfied under τ. Finally, since all the variable x_C with $C \in \mathcal{C}$ have distance two from x in the factor graph, and as x is tame, we have $\mathrm{dist}(\sigma, \tau) \leq \ln n$. □

5.2 Rigid variables

Assume that k, ρ, θ satisfy the assumptions of Theorem 6. Let (Φ_t, σ_t) be the (random) outcome of **U1–U4**. Our goal is to show that w.h.p. most variables $x \in V_t$ are rigid. What is the basic obstacle that makes it difficult to 'flip' the value of x? Observe that we can simply assign x the opposite value $1-\sigma_t(x)$, unless Φ_t has a clause $\mathcal{C}$ in which either x or $\bar{x}$ is the *only* literal that is true under σ_t. If there is such a clause, we say that x *supports* $\mathcal{C}$. But even if x supports a clause $\mathcal{C}$ it might be easy to flip. For instance, if $\mathcal{C}$ features some variable $y \neq x$ that does not support a clause, then we could just flip both x, y simultaneously. Thus, to establish the existence of $\Omega(n)$-rigid variables we need to analyze the distribution of the number of clauses that a variable supports, the probability that these clauses only consists of variables that support further clauses, the probability that the same is true of those clauses, etc, i.e., we will need to establish the existence of a special set of literals defined as follow.
In a random pair (Φ_t, σ_t) chosen from the experiment **U1–U4**, we say that a set $\mathcal{S} \subset L_t$ of true literals under σ is a *t-self-contained* if each literal $l \in \mathcal{S}$ supports at least two clauses that contain literals from $\bar{\mathcal{S}}$ only, where $\bar{\mathcal{S}}$ is the set of all negations of literals in $\mathcal{S}$. These set muss be build with special attention by a recursive procedure and using the following result.

Proposition 26. *Suppose that $k \geq 6$ and $0 < r \leq 2^k \ln 2 - k$. Let $\mu = \rho \cdot 2^k/(2^k-1)$ and $\zeta = (1+\mu+\mu^2/2)/\exp(\mu)$, and assume that $2^k\theta\zeta\varphi(1) > \rho$. Then w.h.p. in a random pair (Φ_t, σ_t) chosen from the experiment* **U1–U4** *no more than $2\zeta\theta n$ literals in L_t support fewer than three clauses.*

Proof. We are going to work under the experiment **P1'–P4'**, thus here $\mathcal{P}'_k(n,m)$ will make reference to the distribution under such experiment. Let S be the number of literals $l \in L_t$ that support fewer than three clauses. We just need to show that

$$\mathrm{P}_{\mathcal{P}'_k(n,m)}\left[S > 2\zeta\theta n\right] \leq \exp(-krn/4^k), \tag{14}$$

then Corollary 15 implies the assertion.
Using the second part of Lemma 16, under **P1'–P4'** we can see that the random variable S is dominated by a binomial with mean $(1+o(1))\theta\zeta n$. Hence, the Chernoff bound (8) shows that

$$\mathrm{P}_{\mathcal{P}'_k(n,m)}\left[S > 2\zeta\theta n\right] \leq \exp\left(-(1+o(1))\theta\zeta\varphi(1)n\right). \tag{15}$$

By the assumptions on μ and θ we have $\theta\zeta\varphi(1) > \rho/2^k$; hence, (14) follows from (15). □

Proposition 27. *Suppose that $k \geq 4$ and $0 < r \leq 2^k \ln 2 - k$, and that $0 \leq \theta \leq 1$. Set*

$$\mu = \frac{\rho 2^k}{2^k - 1},\ \zeta = \frac{1+\mu+\mu^2/2}{\exp(\mu)},$$

If $\zeta < 1/3$, then w.h.p.in a random pair (Φ_t, σ_t) chosen from the experiment **U1–U4***, there is a t-self-contained set of size at least $(1-3\zeta)\theta n$.*

This proof is by construction. Let us explain it in two big steps.

1. Start by taking out all the true literals in L_t which support fewer than three clauses. Let Z be this set. By Proposition 26 we know that w.h.p.$|Z| \leq 2\zeta\theta n$.
2. Now, take out the set of true literal for which, the number of clauses each of these literals support without literals from $\bar{Z}$ is less than two. It muss be made by middle of a recursive procedure as follow.
   (a) Let $Z_1 = Z$.
   (b) Take any literal $l \notin Z_1$, then l support three or more clauses. Let T_l be the number of clauses supported by l in which a literal from $\bar{Z}_1$ occurs. If $S_l - T_l < 2$, add it to Z_1.
   (c) Repeat this procedure $\theta n - |Z|$ times. (We will show that after $\theta n - |Z|$ times, w.h.p.$|Z_1 - Z| \geq \zeta n$).

Thus, we will prove w.h.p.the existence of a t-self-contained set of size at least $(1-3\zeta)\theta n$.

Proof (Proposition27). Let (Φ_t, σ_t) be chosen from the experiment **P1'–P4'**. We may condition on σ being the all-true assignment, and on the event that at most $2\zeta\theta n$ true literals under σ, support fewer than three clauses, i.e., $|Z| \leq 2\zeta\theta n$.
Note that in a random clause supported by l, the variables underlying the $k-1$ negative literals in that clause are distributed uniformly over V_t. One more time, let $Z_1 = Z$ and think on the recursive procedure described above,

then, the probability that a random clause supported by $l \notin Z_1$ with at least one literal from $\bar{Z}_1$ is $1-(1-|Z_1|/\theta n)^{k-1}$. Thus, T_l is dominated by a Binomial, $\mathrm{Bin}(S_l,\lambda)$, with $\lambda = 1-(1-3\zeta)^{k-1}$ and

$$\gamma = P(S_l - T_l < 2 \mid S_l = j, j \geq 3) = P(T_l \geq S_l - 1 \mid S_l = j, j \geq 3) \tag{16}$$

$$= \sum_{j=3} P[\mathrm{Bin}(j,\lambda) \geq j-1] P(S_l = j \mid j \geq 3) \tag{17}$$

$$< (1+o(1)) \sum_{j=3} \frac{j\lambda^{j-1}\mu^j}{j!\exp(\mu)(1-\zeta)} \tag{18}$$

$$= (1+o(1)) \frac{\mu(\exp(\lambda\mu) - \lambda\mu - 1)}{(1-\zeta)\exp(\mu)}. \tag{19}$$

Now we need to calculate how many literals are in $Z_1 - Z$ after $\theta n - |Z|$ times.
For any true literal $l \notin Z_1$ that support more than two clauses i.e., $(S_l \geq 3)$, define the variable,

$$Y_l = \begin{cases} 1 \text{ if } T_l \geq S_l - 1 \\ 0 \text{ otherwise.} \end{cases}$$

Observe that $\sum_{i=1}^{(1-2\zeta)\theta n} Y_l$ is distributed $\mathrm{Bin}((1-2\zeta)\theta n, \gamma)$. Assuming that $\zeta < 1/3$ and $\gamma \geq \zeta/(1-2\zeta)$ we can use Chernoff bound (8) to get

$$P(\sum_{i=1}^{(1-2\zeta)\theta n} Y_l \leq \zeta\theta n) \leq \exp\Big\{ -\theta n[(1-2\zeta)\gamma - \zeta]\varphi\Big(\frac{\zeta}{(1-2\zeta)\gamma} - 1\Big)\Big\} \tag{20}$$

Finally, to pass from the experiment **P1'–P4'** to **U1–U4** we need

$$\theta[(1-2\zeta)\gamma - \zeta]\varphi\Big(\frac{\zeta}{(1-2\zeta)\gamma} - 1\Big) > \frac{\rho}{2^k},$$

??????????? Thus we have gotten that w.h.p., under the experiment **U1–U4**, $|Z_1 - Z| \geq \zeta\theta n$.

□

Proposition 28. *For any $k \geq 3$ there is a number $\chi = \chi(k) > 0$ such that for any $0 < r \leq 2^k \ln 2 - k$ the following is true. Let (Φ_t, σ_t) be a random pair chosen from the from the experiment* **U1–U4**. *Then w.h.p. for any t-self-contained set $\mathcal{S}$ all variables $x \in \mathcal{S} \cup \bar{\mathcal{S}}$ are χn-rigid.*

The proof of Proposition 28 uses an elementary 'expansion property' of the random formula Φ given by the following Lemma.

Lemma 29. *There is a number $\chi = \chi(k) > 0$ such that for all $0 < r \leq 2^k$ the random formula Φ has the following property w.h.p.*

$$\begin{aligned} &\textit{There is no set } Q \textit{ of } 1 \leq |Q| \leq \chi n \textit{ variables such that the number of} \\ &\textit{clauses containing at least two variables from } Q \textit{ is at least } 2|Q|. \end{aligned} \tag{21}$$

Proof. We use a first moment argument. Let $1 \leq q \leq \chi n$ and let $Q_0 = \{x_1, \ldots, x_q\}$ be a fixed set of size q. For any set Q we let $Y(Q)$ be the number of clauses containing at least two variables from Q. Moreover, let X_q be the number of sets Q of size q such that $Y(Q) \geq 2q$. Since the distribution $F_k(n,m)$ is symmetric with respect to permutations of the variables, we have

$$\mathrm{E}X_q \leq \binom{n}{q} \cdot \mathrm{P}\left[Y(Q_0) \geq 2q\right] \leq \exp\left[q(1+\ln(n/q))\right] \cdot \mathrm{P}\left[Y(Q_0) \geq 2q\right]. \tag{22}$$

Furthermore, the probability that a random k-clause contains two variables from Q_0 is at most $\binom{k}{2}(q/n)^2$ (because for each of the $\binom{k}{2}$ pairs of 'slots' in the clauses the probability that both of them are occupied by variables from

Q_0 is at most $(q/n)^2$). As Φ consists of m independent k-clauses, $Y(Q_0)$ is stochastically dominated by a binomial random variable $\mathrm{Bin}(m, \binom{k}{2}(q/n)^2)$. Consequently, assuming that $q/n \leq \chi$ is sufficiently small, we get

$$\begin{aligned}\mathrm{P}\left[Y(Q_0) \geq 2q\right] &\leq \mathrm{P}\left[\mathrm{Bin}\left(m, \binom{k}{2}q\right) \geq 2q\right] \\ &\leq \exp\left[-1.9q \cdot \left[\ln\left(\frac{2q}{\binom{k}{2}(q/n)^2 m}\right) - 1\right]\right] \text{ [by Chernoff bound (8)]} \\ &\leq \exp\left[-1.9q \cdot \ln\left(\frac{4}{\mathrm{e}k^2 r} \cdot \frac{n}{q}\right)\right] \leq \exp\left[-1.9q \cdot \ln\left(\frac{4}{\mathrm{e}k^2 2^k} \cdot \frac{n}{q}\right)\right]. \end{aligned} \tag{23}$$

Choosing $\chi = \chi(k)$ sufficiently small, we can ensure that $(q/n)^{1/4} \leq \chi^{1/4} \leq 4/(\mathrm{e}k^2 2^k)$. Plugging this bound into (23), we get

$$\mathrm{P}\left[Y(Q_0) \geq 2q\right] \leq \exp\left[-1.1q \cdot \ln\left(n/q\right)\right]. \tag{24}$$

Combining (22) and (24), we get $\mathrm{E}X_q \leq \exp\left[-0.1q \ln\left(n/q\right)\right]$.
In effect, $\mathrm{E}\sum_{1 \leq q \leq \chi n} X_q = O(n^{-0.1})$. Hence, Markov's inequality implies that w.h.p. $\sum_{1 \leq q \leq \chi n} X_q = 0$, in which case (21) holds. □

Proof (Proposition28). Let (Φ_t, σ_t) be a random pair chosen from the experiment **U1**–**U4**. By Lemma 29 we know that w.h.p., there is a number $\chi = \chi(k) > 0$ such that (21) is satisfied. We are going to assume that this is the case.
Let $\mathcal{S}$ be a self-contained set. Suppose that τ is a satisfying assignment such that the set Q of all variables $x \in \mathcal{S} \cup \bar{\mathcal{S}}$ such that $\tau(x) \neq \sigma(x)$ is non-empty. For each variable $x \in Q$ there are two clauses $C_1(x), C_2(x)$ that are supported by x under σ and that consist of literals from $\bar{\mathcal{S}}$ only (because $\mathcal{S}$ is self-contained). Since τ is satisfying and $\tau(x) \neq \sigma(x)$, both $C_1(x)$ and $C_2(x)$ contain another variable from Q. Hence, there are at least $2|Q|$ clauses that contain at least two variables from Q. Thus, (21) implies that w.h.p.$|Q| > \chi n$, and consequently $\mathrm{dist}(\sigma, \tau) \geq |Q| > \chi n$.

□

Proof (Theorem 20, part 2). The assertion follows directly from Proposition 28 and Proposition 27. □

5.3 Forced variables

Let (Φ, σ) be a formula/assignment pair. A clause C *forces* a variable $x \in V_t$ if C contains $k-1$ literals from $\{x_1, \bar{x}_1, \ldots, x_t, \bar{x}_t\}$, none of which satisfies C under σ, and either the literal x or $\bar{x}$, which does.

Lemma 30. *Suppose that $\rho \geq \rho_0$, $k \geq k_0$, and $k\theta \sim \ln(\rho) - 10$. Then w.h.p. in a pair (Φ_t, σ_t) chosen from the experiment* **U1**–**U4** *at least $0.991\theta n$ variables in V_t are forced.*

Proof. Let $\mathcal{F}$ be the event that at least $0.991\theta n$ variables in V_t are forced. We are going to show that

$$\mathrm{P}_{\mathcal{P}'_k(n,m)}\left[\mathcal{F}\right] \geq 1 - \exp(-1.1\rho/2^k), \tag{25}$$

so that the assertion follows from Corollary 15. Thus, let (Φ_t, σ_t) be a pair chosen from the experiment **P1'**–**P4'**. We may assume without loss of generality that σ_t is the all-true assignment. For each variable $x \in V_t$ the number of clauses that x supports has a binomial distribution with mean $\mu = \rho \cdot 2^k/(2^k - 1)$. Furthermore, if C is a random clauses supported by x, then C contains $k-1$ random negative literals; the probability that all of these are in $V \setminus V_t$ equals $(1 - \theta + o(1))^{k-1}$. Hence, the number F_x of forcing clauses for x is binomially distributed with mean

$$\begin{aligned}\mathrm{E}\left[F_x\right] &= \mu(1 - \theta + o(1))^{k-1} \geq \rho(1-\theta)^{k-1} \\ &\geq \rho \exp\left[-(\theta + \theta^2)(k-1)\right] \geq \rho \exp\left[-\theta k - \theta^2 k\right] \geq \exp(5).\end{aligned}$$

Therefore, for any $x \in V_t$ we have $\mathrm{P}\left[F_x = 0\right] \leq \exp(-\exp(5))$, and the events $(\{F_x = 0\})_{x \in V_t}$ are mutually independent. Hence, the number Z of variables $x \in V_t$ with $F_x = 0$ is binomially distributed with mean $\exp(-\exp(-5))\theta n$, then by Chernoff bounds

$$\mathrm{P}\left[Z \geq 0.009\theta n\right] \leq \exp(-0.009\theta n) \leq \exp(-1.1\rho/2^k).$$

This proves (25). □

Proof (Theorem 20, part 1). We need to deal with general values $\ln n/n \ll \theta \leq \theta_0 = (\ln(\rho) - 10)/k$. Let $t = (1-\theta)n$ and $t_0 = (1-\theta_0)n$. To obtain a pair (Φ_t, σ_t) from the experiment **U1–U4**, one can proceed as follows. First, choose a pair $(\Phi_{t_0}, \sigma_{t_0})$ from the experiment **U1–U4** with t_0 variables decimated. Then, assign the variables in $x \in V_{t_0} \setminus V_t$ with the truth values $\sigma_{t_0}(x)$, simplify the formula, and let $\sigma_t(y) = \sigma_{t_0}(y)$ for all $y \in V_t$. We are going to use this experiment to analyze the number of forced variables in (Φ_t, σ_t).
The above experiment shows that any variable $x \in V_t$ that is forced in (Φ_{t_0}, σ_0) remains forced in (Φ_t, σ_t). Let $\mathcal{F}$ be the set of forced variables in (Φ_{t_0}, σ_0). Given that $|\mathcal{F}| = j$, the set $\mathcal{F}$ is a uniformly random subset of V_{t_0}. Hence, if we condition on the event that $|\mathcal{F}| \geq 0.991\theta_0 n$, then $|\mathcal{F} \cap V_t|$ has a hypergeometric distribution with mean at least $0.991\theta n$. Therefore, by Chebyshev's inequality, we have $|\mathcal{F} \cap V_t| \geq (0.991\theta - o(1))n \geq 0.99\theta n$ w.h.p. (here we use that $\theta n \gg 1$). Thus, the theorem follows from Lemma 30. □

6 Proof of Theorem 21

6.1 Main ideas of the proof

We work here with the planted model. Let (Φ, σ) a pair chosen from the distribution $\mathcal{P}_k(n, m)$ and Φ_t denote the formula obtained from Φ by substituting the values $\sigma(x_1), \ldots, \sigma(x_t)$ for the first t variables. Without loss of generality, we may assume that $\sigma = \mathbf{1}$ is the all-true assignment. Let Z_{σ_t} be the number of solutions compatible with the partial assignment of variables up to time t i.e., $Z_{\sigma_t} = |\{\tau \in \mathcal{S}(\Phi_t)\}|$, then

$$E(Z_{\sigma_t}) = \sum_{\tau \in \Sigma} P\big[\tau \in \mathcal{S}(\Phi) \mid \tau(x_1) = \sigma(x_1), \ldots, \tau(x_t) = \sigma(x_t), \sigma \in \mathcal{S}(\Phi)\big], \tag{26}$$

where Σ is the set of all 2^n assignments.
Let c_i be the i-th clause of Φ and define the next events.

$$\begin{aligned} A_i &:= \{\tau \text{ satisfies } c_i\} \\ B_i &:= \{\sigma \text{ satisfies } c_i\} \\ C &:= \tau(x_1) = \sigma(x_1), \ldots, \tau(x_t) = \sigma(x_t), \end{aligned}$$

then observe that

$$\begin{aligned} P\big[\tau \in \mathcal{S}(\Phi) \mid \tau(x_1) = \sigma(x_1), \ldots, \tau(x_t) = \sigma(x_t), \sigma \in \mathcal{S}(\Phi)\big] \\ = P\big[\cap_{i=1}^m (A_i \mid B_i \cap C)\big]. \end{aligned}$$

Let z be the overlap between σ and τ, i.e, the number of variables to which σ and τ assign the same value. It is well known that the probability that a fix pair of truth assignments σ and τ satisfies the i-th random clause c_i depends only on the number z. That follows because if c_i is not satisfied by σ, the only possibility for c_i neither be satisfied by τ is because all variables in c_i lie in the overlap.
Let $z = \gamma n$, then the sequence $\{A_i \cap B_i \mid z\}_{i \geq 1}$ is a sequence of independent events and $P(A_i \cap B_i \mid z) = 1 - \frac{1}{2^{k-1}} + \frac{\gamma^k}{2^k}$ (see equation (3) in [4]).
Furthermore, observe that the event $\{A_i \mid B_i \cap C\}$ imply that $\{A_i \mid B_i \cap \{\gamma n = t + \beta(n-t)\}\}$, for some $0 \leq \beta \leq 1$. Taking $\alpha = 1 - \beta$ we get

$$\begin{aligned} P\big[\cap_{i=1}^m (A_i \mid B_i \cap C)\big] &\leq P\big[\cap_{i=1}^m (A_i \mid B_i \cap \{\gamma = (1-\alpha\theta)\})\big] \\ &= \frac{\Pi_{i=1}^m P\big[(A_i \cap B_i \mid \gamma = (1-\alpha\theta))\big]}{P\big[B_i \mid \gamma = (1-\alpha\theta)\big]} \\ &= \Big(\frac{1 - \frac{1}{2^{k-1}} + \frac{\gamma^k}{2^k}}{1 - \frac{1}{2^k}}\Big)^m \\ &= \Big(1 - \frac{1-\gamma^k}{2^k - 1}\Big)^m. \end{aligned}$$

Hence,

$$\begin{aligned} E(Z_{\sigma_t}) &\leq \sum_{\gamma = t/n}^{n} \sum_{overlap(\sigma,\tau) = \gamma n} \Big(1 - \frac{1-\gamma^k}{2^k - 1}\Big)^m \\ &= \sum_{\gamma = t/n}^{n} \binom{n}{\gamma n} \Big(1 - \frac{1-\gamma^k}{2^k - 1}\Big)^m. \end{aligned}$$

Let $X_\gamma = |\{\tau \in \mathcal{S}(\Phi_{t,\sigma}) : overlap(\sigma,\tau) = \gamma n\}|$, then we can write $E(Z_{\sigma_t}) = \sum_\gamma \mathrm{E}X_\gamma$, thus

$$\mathrm{E}X_\gamma \leq \binom{n}{\gamma n}\Big(1 - \frac{1-\gamma^k}{2^k-1}\Big)^m. \tag{27}$$

Let us make a change of parameter. Note that if $overlap(\sigma,\tau) = \gamma n$ then $\mathrm{dist}(\sigma,\tau) = \alpha\theta n$. Consider $X_\alpha = |\{\tau \in \mathcal{S}(\Phi_{t,\sigma}) : \mathrm{dist}(\sigma,\tau) = \alpha\theta n\}|$, thus from (27) we have

$$\mathrm{E}X_\alpha \leq \binom{\theta n}{\alpha\theta n}\left(1 - \frac{1-(1-\alpha\theta)^k}{2^k-1}\right)^m. \tag{28}$$

Taking logarithms and bounding the binomial coefficient via Stirling's formula we obtain

$$\frac{\ln \mathrm{E}X_\alpha}{n} \leq \psi(\alpha), \tag{29}$$

where $\psi(\alpha) = -\alpha\theta\ln\alpha - (1-\alpha)\theta\ln(1-\alpha) + r\ln\Big(1 - \frac{1-(1-\alpha\theta)^k}{2^k-1}\Big)$.
The crucial point here is that if for some fix θ, exits $\alpha > 0$ such that; $\Psi(\alpha) < 0$, then by Markov inequality follows that w.h.p Φ_t does not have any satisfying assignment τ, such that $\mathrm{dist}(\sigma,\tau) = \alpha\theta n$. To establish the 'shattering' part we are going to prove the following: Under the assumptions of Theorem 21 part 1, there exist $0 < a_1 < \alpha < a_2 < 1$ depending only on k, ρ such that w.h.p. we have

$$\Psi(\alpha) < 0, \text{ for } 0 < a_1 < \alpha < a_2 < 1, \tag{30}$$

then, following the idea in [2] to show the existence of clusters in the solution space, we get a partition of the set $\mathcal{S}(\Phi_{t,\sigma})$ into well-separated regions. The picture of $\psi(\alpha)$ is described by Figure 2 left side.
These regions are build as follow:

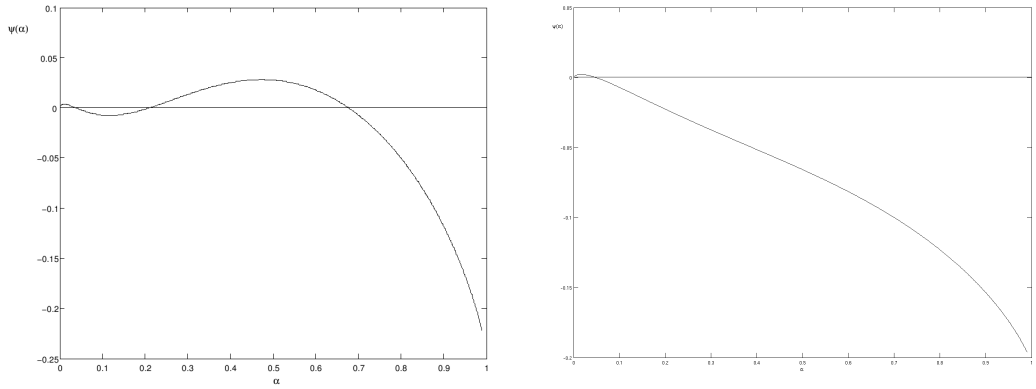

Fig. 2. Left and right sides represent $\psi(\alpha)$ for shattering and condensation phases respectively.

Choose any $\sigma_1 \in \mathcal{S}(\Phi_t)$ and let $\mathcal{C}_{\sigma_1} = \{\chi \in \mathcal{S}(\Phi_t) : \mathrm{dist}(\chi,\sigma_1) \leq a_1\theta n\}$. Then, choose $\sigma_2 \in \mathcal{S}(\Phi_t) \setminus \mathcal{C}_{\sigma_1}$ and let $\mathcal{C}_{\sigma_2} = \{\chi \in \mathcal{S}(\Phi_t) \setminus \mathcal{C}_{\sigma_1} : \mathrm{dist}(\chi,\sigma_2) \leq a_1\theta n\}$. Proceed inductively until all remaining satisfying assignments have been assigned to a region $\mathcal{C}_{\sigma_i}$ which diameter at most $a_1\theta n$. Suppose we find N of such regions $\mathcal{C}_{\sigma_i}$'s. Let $R_l = \mathcal{C}_{\sigma_l} \setminus \bigcup_{i=1}^{l-1}\mathcal{C}_{\sigma_j}$ and $R_0 = \mathcal{S}(\Phi_t) \setminus \bigcup_{i=1}^{N} R_i$. Note that $R_0, R_1, \ldots, R_N$ are disjoint sets such that; $\mathcal{S}(\Phi_{t,\sigma}) = \cup_0^N R_j$. Furthermore, since w.h.p. there are not any pair of truth assignment at distance $a_1\theta n \leq \mathrm{dist}(\chi,\tau) \leq a_2\theta n$, then $\mathrm{dist}(R_i, R_j) \geq a_2\theta n$ for any $i \neq j \in \{1,\ldots,N\}$. The decomposition $R_0,\ldots,R_N$ witnesses that $\mathcal{S}(\Phi_t)$ shatters.
To prove Condensation we also analyze the first moment of $X\alpha$. Here, we will prove that for some values of θ, i.e., when the fix variables overcome some value, w.h.p. it is not possible to find another solution τ such that; $\mathrm{dist}(\sigma,\tau) \geq \alpha$. Hence, if exist more solutions they should be close to each other with respect to the Hamming distance. The pictures suggested by this result comes in Figure 2 right side.
With respect to pairwise distances of satisfying assignments, under the situation described by Figure 2 left side,

we need to prove that $\sup_{0<\alpha\leq 0.49}\psi(\alpha) < \theta\ln 2 + 2^k\rho\ln(1-2^{-k})/k - \rho/2^{k-1}$. This implies that w.h.p. only an exponentially small fraction of all the satisfying assignments of Φ_t lies within distance $\leq 0.49\theta n$ between them. From this we derive the statement made in Theorem 21 on the pairwise distance.
In addition, the fact that the average pairwise distance of satisfying assignments is $\geq 0.49\theta n$ w.h.p. implies in combination with a double counting argument the claim about the marginals $M_x(\Phi_t)$ in Theorem 22.

6.2 Shattering

In this section we prove the first part of Theorem 21. The main step of the proof is summarized in the following proposition.

Proposition 31. *Let $k \geq 6$ and $r > 0$ be fixed. Moreover, let $0 < \theta \leq 1$ and let*

$$\psi : (0,1) \to \mathbf{R}, \quad \alpha \mapsto -\alpha\theta\ln\alpha - (1-\alpha)\theta\ln(1-\alpha) + r\ln\left(1 - \frac{1-(1-\alpha\theta)^k}{2^k-1}\right).$$

Suppose that there is a number $a \in (0,1)$ such that

$$\psi(a) + \rho/2^k < 0 \quad \text{and} \quad \sup_{0<\alpha<a}\psi(\alpha) < \theta\ln 2 + 2^k\rho\ln(1-2^{-k})/k - \rho/2^k. \tag{31}$$

Then there is $\varepsilon = \varepsilon(k,\rho)$ such that for Φ_t generated by the experiment **U1**–**U4***, the set $\mathcal{S}(\Phi_t)$ is $(a-\varepsilon, a+\varepsilon)$-shattered.*

To establish Proposition 31 and how it implies the first part of Theorem 21, first we need to prove the following.

Lemma 32. *Keep the assumptions from Proposition 31 and let*

$$b = \theta\ln 2 + 2^k\rho\ln(1-2^{-k})/k.$$

There exist numbers $\xi > 0$, $0 < a_1 < a_2 < 1$ such that a pair (Φ,σ) chosen from the distribution $\mathcal{U}_k(n,m)$ has the following two properties with probability at least $1-\exp(-\xi n)$.

1. *Φ_t does not have a satisfying assignment τ with $a_1 n < dist(\sigma,\tau) < a_2 n$.*
2. *$|\{\tau \in \mathcal{S}(\Phi_t) : dist(\sigma,\tau) < a_2 n\}| \leq \exp((b-\xi)n)$.*

Proof. Let (Φ,σ) a pair chosen from the distribution$\mathcal{P}_k(n,m)$. For $\alpha > 0$, let

$$X_\alpha = |\{\tau \in \mathcal{S}(\Phi_t) : \mathrm{dist}(\sigma,\tau) = \alpha\theta n\}|.$$

Assume ψ satisfies (31) and let $a \in (0,1)$ be such that $\psi(a) + \rho/2^k < 0$. As ψ is continuous there exist $0 < a_1 < \alpha < a_2 < 1$ and $\xi_1 > 0$ such that

$$\sup_{a_1\leq\alpha\leq a_2}\psi(\alpha) < -\rho/2^k - 2\xi_1. \tag{32}$$

Combining (29) and (32) we conclude that $\mathrm{E}X_\alpha \leq exp\left[-n(\rho/2^k + 2\xi_1)\right]$ for all $a_1 \leq \alpha \leq a_2$. Summing over integers $a_1 n \leq j \leq a_2 n$ we see that for large n

$$\sum_{a_1 n\leq j\leq a_2 n}\mathrm{E}X_{j/n} \leq n\exp\left[-n(\rho/2^k + 2\xi_1)\right] \leq \exp\left[-n(\rho/2^k + \xi_1)\right].$$

Hence, by Markov's inequality the probability that there is a satisfying assignment τ that coincides with σ on the first t variables such that $a_1 n \leq \mathrm{dist}(\sigma,\tau) \leq a_2 n$ is bounded by $\exp(-n(\rho/2^k + \xi_1))$. This proves the first assertion.
Since we are assuming that $\sup_{0<\alpha<a}\psi(\alpha) < b - \rho/2^k$, and as (32) shows that $\psi(\alpha) < -\rho/2^k - 2\xi_1 < b - \rho/2^k - 2\xi_1$ for all $a_1 \leq \alpha < a_2$, then, there is a number $\xi_2 > 0$ such that

$$\sup_{0<\alpha\leq a_2}\psi(\alpha) \leq b - \rho/2^k - 3\xi_2.$$

Hence, (29) implies that

$$\mathrm{E}X_\alpha \leq \exp(n\psi(\alpha)) \leq \exp(n(b - \rho/2^k - 3\xi_2)) \qquad \text{for all } 0 < \alpha \leq a_2.$$

Taking the sum over integers $0 \leq j \leq a_2 n$ we get for large enough n

$$\sum_{0 \leq j \leq a_2 n} \mathrm{E} X_{j/n} \leq n \exp(n(b - \rho/2^k - 3\xi_2)) \leq \exp(n(b - \rho/2^k - 2\xi_2)).$$

That is, the expected number of assignments $\tau \in \mathcal{S}(\Phi_t)$ such that $\mathrm{dist}(\sigma, \tau) \leq a_2 n$ is bounded by $\exp(n(b - \rho/2^k - 2\xi_2))$. Hence, Markov's inequality entails that with probability at least $1 - \exp(-n(\rho/2^k + \xi_2))$ there are at most $\exp(n(b - \xi_2))$ such satisfying assignments τ. This proves the second assertion.
Finally, the result on $\mathcal{U}_k(n, m)$ follows directly from Corollary 13. □

Proof (Proposition 31). Let ξ, a_1, a_2 be the numbers provided by Lemma 32 and let (Φ, σ) be a pair chosen from the distribution $\mathcal{U}_k(n, m)$. With each assignment $\tau \in \mathcal{S}(\Phi_t)$ we associate a set

$$\mathcal{C}(\tau) = \{\chi \in \mathcal{S}(\Phi_t) : \mathrm{dist}(\chi, \tau) \leq a_1 n\}.$$

Moreover, we call $\tau \in \mathcal{S}(\Phi_t)$ *good* if $|\mathcal{C}(\tau)| \leq \exp((b - \xi)n)$ and there is no $\chi \in \mathcal{S}(\Phi_t)$ such that $a_1 n \leq \mathrm{dist}(\chi, \tau) \leq a_2 n$. Let $\mathcal{S}_{good}$ be the set of all good $\tau \in \mathcal{S}(\Phi_t)$ and $\mathcal{S}_{bad} = \mathcal{S}(\Phi_t) \setminus \mathcal{S}_{good}$. By Corollary 14, lemma 32 and our choice of b ensure that Φ has the following two properties w.h.p.:

$$|\mathcal{S}(\Phi_t)| \geq \exp n(b - \rho/2^k), \tag{33}$$

$$|\mathcal{S}_{good}| \geq (1 - \exp(-\xi n)) \cdot |\mathcal{S}(\Phi_{t,\sigma})|. \tag{34}$$

Assuming that (33) and (34) hold and that n is sufficiently large, we are going to construct a decomposition of $\mathcal{S}(\Phi_t)$ into subsets as required by **SH1**–**SH2**. To this end, choose some $\sigma_1 \in \mathcal{S}_{good}$. Having defined $\sigma_1, \ldots, \sigma_l$, we choose an arbitrary $\sigma_{l+1} \in \mathcal{S}_{good} \setminus \bigcup_{j=1}^{l} \mathcal{C}(\sigma_j)$, unless this set is empty, in which case we stop. Let $\sigma_1, \ldots, \sigma_N$ be the resulting sequence and define

$$R_l = \mathcal{C}(\sigma_l) \setminus \bigcup_{j=1}^{l-1} \mathcal{C}(\sigma_j) \qquad \text{for } 1 \leq l \leq N, \quad \text{and } R_0 = \mathcal{S}(\Phi_{t,\sigma}) \setminus \bigcup_{l=1}^{N} R_l.$$

Then $\mathcal{S}(\Phi_t) = R_0 \cup \cdots \cup R_N$. (Observe that possibly $R_0 = \emptyset$ while $R_l \neq \emptyset$ for all $1 \leq l \leq N$ as $\sigma_l \in R_l$.) Furthermore, for each $1 \leq l \leq N$ we have $R_l \subset \mathcal{C}(\sigma_l)$ and thus

$$\begin{aligned} |R_l| &\leq |\mathcal{C}(\sigma_l)| \leq \exp((b - \xi)n) && \text{[because } \sigma_l \text{ is good]} \\ &\leq |\mathcal{S}(\Phi_t)| \cdot \exp(-\xi n) && \text{[by (33)].} \end{aligned} \tag{35}$$

Furthermore, as $R_0 \subset \mathcal{S}_{bad}$, (34) implies

$$|R_0| \leq |\mathcal{S}_{bad}| \leq \exp(-\xi n)) \cdot |\mathcal{S}(\Phi_{t,\sigma})|. \tag{36}$$

Combining (35) and (36) we see that the decomposition $R_0, \ldots, R_N$ satisfies **SH1**. Furthermore, **SH2** is satisfied by construction. □

Now, we just need to verify (31).

Lemma 33. *Assume that* $0 \leq \theta \leq \exp(\rho - 2)/(\rho k)$. *Let* $a = \exp(2 - \rho)$. *Then* $\psi(a) < -a\theta/2$.

Proof. We have

$$\begin{aligned} \psi(a) &\leq a\theta(1 - \ln a) - \frac{\rho}{k}\left(1 - (1 - a\theta)^k\right) \leq a\theta(1 - \ln a) - \frac{\rho}{k}\left(1 - \exp(-ak\theta)\right) \\ &\leq a\theta(1 - \ln a) - \frac{\rho}{k}\left(ak\theta - (ak\theta)^2/2\right) = a\theta\left[1 - \ln a - \rho(1 - ak\theta/2)\right], \end{aligned}$$

where we have used $\exp(-z) \leq 1 - z + z^2/2$ for $z \geq 0$. Since $k\theta\rho \leq \exp(\rho - 2)$ by assumption, our choice of a implies that $\psi(a) \leq a\theta\left[1 - \ln a - \rho + a\exp(\rho - 2)/2\right] = -a\theta/2$, as claimed. □

Lemma 34. *Assume that* $0 \leq \theta \leq \exp(\rho - 2)/(\rho k)$. *Let* $a = \exp(2 - \rho)$. *Then* $\sup_{\alpha < a} \psi(\alpha) \leq \frac{3}{2\mathrm{e}^2 k\rho}$.

Proof. Let $0 \leq \alpha < a$. We have

$$\psi(\alpha) \leq \theta(\alpha - \alpha \ln \alpha - \alpha\rho(1 - \alpha k\theta/2)).$$

Let $\psi_1(\alpha)$ be the expression on the r.h.s. Then

$$\frac{d}{d\alpha}\psi_1(\alpha) = \theta\left[-\ln\alpha - \rho + \alpha k\theta\right], \quad \frac{d^2}{d\alpha^2}\psi_1(\alpha) = \theta\left[k\theta - 1/\alpha\right].$$

Thus, our assumption on θ implies that $\frac{d^2}{d\alpha^2}\psi_1(\alpha) < 0$ for all $0 < \alpha < a$, and therefore ψ_1 has a unique local maximum in the interval $(0, \alpha)$. To pinpoint this maximum, note that for $\alpha_0 = \exp(-\rho)$ the first derivative $\frac{d}{d\alpha}\psi_1(\alpha_0)$ is positive. Moreover, at $\alpha_1 = \exp(1-\rho)$ we have $\frac{d}{d\alpha}\psi_1(\alpha_1) < 0$. Hence, the unique local maximum of ψ_1 lies in the interval (α_0, α_1). To study the maximum value, consider the function $\psi_2 : \alpha \mapsto \alpha - \alpha\ln\alpha - \alpha\rho$. Its derivative is $d/d\alpha\,\psi_2(\alpha) = \rho - \ln\alpha$, so that the maximum of this function occurs at α_0. Furthermore, the quadratic term $\alpha \mapsto \alpha^2 k/2$ is monotonically increasing in α. Therefore,

$$\sup_{0<\alpha<a}\psi(\alpha) \leq \sup_{0<\alpha<a}\psi_1(\alpha) = \sup_{\alpha_0<\alpha<\alpha_1}\psi_1(\alpha) \leq \theta(\psi_2(\alpha_0) + \alpha_1^2 k/2) = 3\theta\exp(-\rho)/2.$$

Finally, the assertion follows from the assumed bound on θ. □

Proof (Theorem 21, part 1). Assume that $\rho \leq k\ln 2 - \ln k$ and

$$\frac{\rho}{k\ln 2}(1 + \rho^{-2} + 2^{2-k}) \leq \theta \leq \exp(\rho - 2)/(\rho k).$$

Let $a = \exp(2-\rho)$. Lemma 33 shows that

$$\psi(a) + \rho/2^k \leq \rho/2^k - \exp(2-\rho)\theta/2 \leq \rho/2^k - \frac{\exp(2-\rho)\rho}{k\ln 2} = \frac{\rho}{2^k}\left(1 - \frac{2^k\exp(2-\rho)}{k\ln 2}\right).$$

Since $\rho \leq k\ln 2 - \ln k$, the r.h.s. is negative. By Lemma 34 we have

$$\begin{aligned}\theta\ln 2 + \frac{2^k\rho}{k}\ln(1-2^{-k}) - \rho/2^k &\geq \theta\ln 2 - \frac{\rho}{k} - \rho/2^{k-1} \\ &\geq \frac{1}{k\rho} + 2^{2-k}\rho\ln 2 - \rho/2^{k-1} \geq \frac{1}{k\rho} > \sup_{\alpha<a}\psi(\alpha). \end{aligned} \tag{37}$$

Thus, the assertion follows from Proposition 31. □

6.3 Condensation

Here we prove the second part of Theorem 21. The following proposition reduces that task to a problem in calculus.

Proposition 35. *Let $k \geq 3$ and $r > 0$ be fixed. Let $0 < \theta \leq 1$ and let*

$$\psi : (0,1) \to \mathbf{R}, \quad \alpha \mapsto -\alpha\theta\ln\alpha - (1-\alpha)\theta\ln(1-\alpha) + r\ln\left(1 - \frac{1-(1-\alpha\theta)^k}{2^k - 1}\right).$$

If there is a number $a \in (0,1)$ such that

$$\sup_{a<\alpha\leq 1}\psi(\alpha) + \rho/2^k < 0 \tag{38}$$

then Φ_t is $2a\theta$-condensed.

Proof. Let (Φ, σ) be a pair chosen from the planted distribution $\mathcal{P}_k(n,m)$. For $\alpha > 0$, and

$$X_\alpha = \left|\{\tau \in \mathcal{S}(\Phi_t) : \mathrm{dist}(\sigma,\tau) = \alpha\theta n\}\right|.$$

By (29) we know $\frac{1}{n}\ln \mathrm{E}X_\alpha \leq \psi(\alpha)$, then, assuming (38) $\frac{1}{n}\ln \mathrm{E}X_\alpha < -\rho/2^k$ for $\alpha > a$. By Markov's inequality we have

$$\mathrm{P}\left[\exists\tau \in \mathcal{S}_t(\Phi_t) : d(\sigma,\tau) \geq a\theta n\right] \leq \theta n \cdot \exp(-(\Omega(1) + \rho/2^k)n) < \exp(-\rho n/2^k).$$

Therefore, the assertion follows from Corollary 13. □

Lemma 36. *Suppose that $\rho \leq k\ln 2 - 2\ln k$ and $\theta = (1-1/\rho^2)\frac{\rho}{k\ln 2}$. Moreover, assume that $\rho \geq \rho_0$ and $k \geq k_0$ for certain constants ρ_0, k_0. Let $a = \exp(2-\rho)$. Then (38) is satisfied.*

Proof. Let $h(\cdot)$ be the entropy function. We have

$$\psi(\alpha) \leq \theta h(\alpha) - \frac{\rho}{k}(1-\exp(-\alpha k\theta)).$$

To bound the r.h.s., we are going to consider several cases.

Case 1: $\alpha \leq 1/(k\rho\theta)$**.** As $\alpha \geq a = \exp(2-\rho)$, we obtain

$$\psi(\alpha) \leq \alpha\theta\left[1 - \ln\alpha - \rho + \alpha k\rho\theta/2\right] \leq \alpha\theta\left[\frac{\alpha k\rho\theta}{2} - 1\right] \leq -\alpha\theta/2.$$

The assumption $\rho \leq k\ln 2 - 2\ln k$ ensures that the last term is smaller than $-\rho/2^k$.
Case 2: $1/(k\rho\theta) < \alpha < 1/(k\theta)$**.** We have

$$\begin{aligned}
\psi(\alpha) &\leq \alpha\theta\left[1 - \ln\alpha - \rho + \alpha k\rho\theta/2\right] \\
&\leq \alpha\theta\left[1 + \ln(k\rho\theta) - \rho + \frac{\alpha k\rho\theta}{2}\right] \\
&\leq \alpha\theta\left[1 + \ln(k\rho\theta) - \rho/2\right] && [\text{as } \alpha < 1/(k\theta)] \\
&\leq \alpha\theta\left[1 - \ln\ln 2 + 2\ln\rho - \rho/2\right] && [\text{as } \theta \leq \tfrac{\rho}{k\ln 2}] \\
&\leq -\alpha\theta\rho/4.
\end{aligned}$$

The assumption $\rho \leq k\ln 2 - 2\ln k$ ensures that the last term is smaller than $-\rho/2^k$.
Case 3: $1/(k\theta) < \alpha \leq \alpha_0 = 0.15$**.** We have

$$\begin{aligned}
\psi(\alpha) &\leq \theta h(\alpha) - \frac{\rho}{k}(1-\exp(-\alpha k\theta)) \leq \theta h(\alpha_0) - \frac{\rho}{k}(1-1/\mathrm{e}) \\
&\leq \frac{\rho}{k}\left[\frac{h(\alpha_0)}{\ln 2} - 1 + 1/\mathrm{e}\right].
\end{aligned}$$

The choice of α_0 ensures that the last term is smaller than $-\rho/2^k$.
Case 4: $\alpha_0 < \alpha$**.** As $k\theta = (1-1/\rho^2)\rho/\ln 2$, we get

$$\begin{aligned}
\psi(\alpha) &\leq \theta h(\alpha) - \frac{\rho}{k}(1-\exp(-\alpha k\theta)) \leq \theta\ln 2 - \frac{\rho}{k}(1-\exp(-\alpha_0(1-1/\rho^2)\rho/\ln 2)) \\
&\leq \frac{\rho}{k}\left[\exp(-\alpha_0\rho) - 1/\rho^2\right].
\end{aligned}$$

The last term is smaller than $-\rho/2^k$.

□

Proof (Theorem 21, part 2). Let $\theta_0 = (1-1/\rho^2)\rho/(k\ln 2)$ and $t_0 = (1-\theta_0)n$. Suppose that $\theta \geq \theta_0$. Then Φ_t is obtained from Φ_{t_0} by assigning some further variables. Therefore,

$$\max\left\{d(\sigma,\tau) : \sigma,\tau \in \mathcal{S}(\Phi_t)\right\} \leq \max\left\{d(\sigma,\tau) : \sigma,\tau \in \mathcal{S}(\Phi_{t_0})\right\}.$$

Hence, Proposition 35 and Lemma 36 imply that Φ_t is $\exp(2-\rho)$-condensed w.h.p. □

6.4 Pairwise distances

Recall that Φ_t denotes the formula obtained by substituting the values $\sigma(x_i)$ for x_i for $1 \leq i \leq t$. The next lemma tell us that w.h.p. almost all the satisfying assignments in $\mathcal{S}(\Phi_t)$ are such that, the distance between any two of them is at least $0.49\theta n$.

Lemma 37. *Suppose that $\theta \geq \frac{\rho}{k\ln 2}(1 + 1/\rho^2 + k/2^{k-2})$. Let (Φ,σ) be a pair chosen from the distribution $\mathcal{U}_k(n,m)$. W.h.p. we have*

$$\left|\{\tau \in \mathcal{S}(\Phi_t) : \mathit{dist}(\tau,\sigma) \leq 0.49\theta n\}\right| \leq \exp(-\Omega(n))\left|\mathcal{S}(\Phi_t)\right|.$$

Proof. Consider the pair (Φ, σ) chosen from the planted model $\mathcal{P}_k(n, m)$. For $\alpha > 0$, let

$$X_\alpha = |\{\tau \in \mathcal{S}(\Phi_t) : \text{dist}(\sigma, \tau) = \alpha\theta n\}|.$$

We know that $\frac{\ln EX_\alpha}{n} \leq \psi(\alpha)$ by (29), where $\psi(\alpha)$ is the function given in Proposition 31. Assume that

$$\sup_{\alpha \leq 0.49} \psi(\alpha) - \theta \ln 2 - r \ln\left(1 - 2^{-k}\right) < -\rho/2^{k-1}, \tag{39}$$

and let $b = \theta \ln 2 + r \ln\left(1 - 2^{-k}\right)$. As ψ is continuos there exists $\epsilon > 0$ such that $\sup_{\alpha \leq 0.49} \psi(\alpha) \leq b - \rho/2^{k-1} - \epsilon$, thus for $\alpha \leq 0.49$

$$E[X_\alpha] \leq \exp(n\psi(\alpha) \leq \exp[n(b - \rho/2^{k-1} - \epsilon)],$$

and

$$\begin{aligned}\sum_{0 \leq j \leq 0.49\theta n} E[X_{j/\theta n}] &\leq \theta n \exp[n(b - \rho/2^{k-1} - \epsilon)] \\ &\leq \exp[n(b - \rho/2^{k-1})],\end{aligned}$$

then the expected number of assignments $\tau \in \mathcal{S}(\Phi_t)$ such that $\text{dist}(\sigma, \tau) \leq 0.49\theta n$ is bounded by $\exp[n(b - \rho/2^{k-1})]$. Hence, Markov's inequality entails that

$$P\big[\,|\{\tau \in \mathcal{S}(\Phi_t) : \text{dist}(\tau, \sigma_t) \leq 0.49\theta n\}| \geq \exp[n(b - \rho/2^k - \epsilon)]\big] \leq \exp[n(-\rho/2^k + \epsilon)],$$

thus with probability $1 - \exp[n(-\rho/2^k + \epsilon)]$:

$$|\{\tau \in \mathcal{S}(\Phi_t) : \text{dist}(\tau, \sigma_t) \leq 0.49\theta n\}| < \exp[n(b - \rho/2^k - \epsilon)].$$

By Corollary 14 we know w.h.p. $|\mathcal{S}(\Phi_t)| \geq \exp[n(b - \rho/2^k)]$, then we get

$$|\{\tau \in \mathcal{S}(\Phi_t) : \text{dist}(\tau, \sigma_t) \leq 0.49\theta n\}| < \exp(-\Omega(n))\,|\mathcal{S}(\Phi_t)|.$$

Now, we are going to show

$$\sup_{\alpha \leq 0.49} \psi(\alpha) - \theta \ln 2 - r \ln\left(1 - 2^{-k}\right) < -\rho/2^{k-1}.$$

We have

$$\begin{aligned}\psi(\alpha) - \theta \ln 2 - \frac{2^k\rho}{k} \ln\left(1 - 2^{-k}\right) &= \theta(h(\alpha) - \ln 2) + \frac{2^k\rho}{k} \ln\left[1 + \frac{(1 - \alpha\theta)^k - 2^{1-k}(1 - (1 - \alpha\theta)^k)}{2^k - 1}\right] \\ &\leq \theta(h(\alpha) - \ln 2) + \frac{\rho}{k}(1 - \alpha\theta)^k + 2^{-k} \\ &\leq \theta(h(\alpha) - \ln 2) + \frac{\rho}{k}\exp(-\alpha k\theta) + 2^{-k}.\end{aligned}$$

The differential of the last expression with respect to θ is negative, and thus the function is monotonically decreasing in θ. Therefore, it suffices to consider the minimum value $\theta = \rho/(k \ln 2)$. Thus, we obtain

$$\psi(\alpha) - \theta \ln 2 - \frac{2^k\rho}{k} \ln\left(1 - 2^{-k}\right) \leq \frac{\rho}{k}\left(\frac{h(\alpha)}{\ln 2} - 1 + \exp(-\alpha\rho/\ln 2)\right) + 2^{-k}.$$

We consider a few different cases.

Case 0: $\alpha < \exp(2 - \rho)$**.** Lemma 34 shows that $\psi(\alpha) \leq 1/(k\rho)$ and (37) shows that

$$\theta \ln 2 + 2^k \frac{\rho}{k} \ln(1 - 2^{-k}) \geq \theta \ln 2 - \rho/k - \rho/2^k.$$

Hence,

$$\psi(\alpha) - \theta \ln 2 - \frac{2^k\rho}{k} \ln\left(1 - 2^{-k}\right) \leq \frac{1}{k\rho} - \theta \ln 2 + \frac{\rho}{k} + \rho/2^k.$$

Since we are assuming that $\theta \geq \frac{\rho}{k \ln 2}(1 + 1/\rho^2 + k/2^{k-2})$, the r.h.s. is smaller than $\rho/2^{k-1}$.

Case 1: $\exp(2-\rho) \leq \alpha \leq \exp(-\rho/2)$**.** Bounding the exponential by a quadratic function, we get

$$\begin{aligned}\psi(\alpha) - \theta\ln 2 - r\ln\left(1-2^{-k}\right) &\leq \frac{\alpha\rho}{k\ln 2}\left[1-\ln\alpha-\rho+\frac{\alpha\rho^2}{4\ln 2}\right]+2^{-k}\\ &\leq \frac{\alpha\rho}{k\ln 2}\left[-1+\frac{(\alpha\rho)^2}{2\ln 2}\right]+2^{-k} < -\rho/2^{k-1},\end{aligned}$$

provided that $\rho_0 \leq \rho \leq k\ln 2 - 2\ln k$.

Case 2: $\exp(-\rho/2) \leq \alpha \leq 1/(2\rho)$**.** Bounding the exponential by a quadratic function, we get

$$\psi(\alpha) - \theta\ln 2 - r\ln\left(1-2^{-k}\right) \leq \frac{\alpha\rho}{k\ln 2}\left[1-\ln\alpha-\rho+\frac{(\alpha\rho)^2}{2\ln 2}\right]+2^{-k} < -\rho/2^{k-1},$$

provided that $\rho_0 \leq \rho \leq k\ln 2 - 2\ln k$.

Case 3: $1/(2\rho) < \alpha \leq 10\ln(2)/\rho$**.** Suppose that $\alpha = x\ln(2)/\rho$ for some $1/2 \leq x \leq 10\ln 2$. Then

$$\psi(\alpha) - \theta\ln 2 - r\ln\left(1-2^{-k}\right) \leq \frac{\rho}{k}\left[\frac{x\ln 2}{\rho}(1-\ln x-\ln\ln 2+\ln\rho)-1+\exp(-x)\right]+2^{-k}.$$

As x remains bounded away from 0, the term $\exp(-x)-1$ is strictly negative. Thus, the entire expression is smaller than $-\rho/2^{k-1}$ for $\rho \geq \rho_0$ sufficiently large.

Case 4: $10\ln(2)/\rho < \alpha \leq 0.49$**.** We have

$$\psi(\alpha) - \theta\ln 2 - r\ln\left(1-2^{-k}\right) \leq \frac{\rho}{k}\left(\frac{h(0.1)}{\ln 2}-1+\exp(-10)\right)+2^{-k}.$$

The r.h.s. is clearly smaller than $-\rho/2^{k-1}$.

The assertion follows from Theorem 13. □

Proof (Theorem 21, part 3). Lemma 37 directly implies the third part of Theorem 21. □

7 Proof of Theorem 22

7.1 Bounding the marginals away from 0, 1

We may assume that $\theta \geq \rho/(k\ln 2)$. The goal is to show that the marginals of a substantial fraction of the variables $x_{t+1},\ldots,x_n$ are bounded away from $0,1$.

We set up an auxiliary graph $\mathcal{G}$ whose vertices are all pairs (x,τ) of variables $x \in V_t$ and assignments $\tau \in \mathcal{S}(\Phi_t)$. A pair (x,τ) is connected by an edge with another pair (x,τ') if $\tau(x) = \tau'(x)$. (Thus, the graph consists of components $(x,\cdot)$ with $x \in V_t$.) Lemma 37 implies the following.

Corollary 38. *Let Φ_t be the formula obtained through the experiment* **U1–U4**. *W.h.p. the number of edges of $\mathcal{G}$, $|E(\mathcal{G})|$ satisfies*

$$2|E(\mathcal{G})| \leq 0.511|\mathcal{S}(\Phi_t)|^2\theta n.$$

Proof. By Lemma 37 for almost all the satisfying assignments in $\mathcal{S}(\Phi_t)$, the 'overlap' between any two of them is less than $0.51\theta n$, i.e., W.h.p. we have

$$|\{\tau \in \mathcal{S}(\Phi_t) : overlap(\tau,\sigma) < 0.51\theta n\}| > [1-\exp(-\Omega(n))]\,|\mathcal{S}(\Phi_t)|\,.$$

Thus, for any pair of such satisfying assignments τ,τ', the number of edges incident between $\{(\tau,x) : x \in V_t\}$ and $\{(\tau',x) : x \in V_t\}$ is at most $(1+o(1))0.51\theta n$. Counting the number of edges from each assignment τ, we have

$$2|E(\mathcal{G})| \leq (1+o(1))0.51|\mathcal{S}(\Phi_t)|^2\theta n \tag{40}$$

□

Proof (Theorem 22, part 1). To bound the marginals away from $0,1$ assume (40) and note that any variable x whose marginal does not lie in $(0.01,0.99)$ is such that the set
$\{(\tau,x) : \tau \in \mathcal{S}(\Phi_t)\}$ induces at least $(1+o(1))0.99|\mathcal{S}(\Phi_t)|(|\mathcal{S}(\Phi_t)|-1)/2$ edges. Hence, if we let ν be the number of such variables, then:
$(1+o(1))0.99|\mathcal{S}(\Phi_t)|(|\mathcal{S}(\Phi_t)|-1)\nu \leq 2|E(\mathcal{G})| \leq (1+o(1))0.51|\mathcal{S}(\Phi_t)|(|\mathcal{S}(\Phi_t)|-1)\theta n$.
Hence, $\nu \leq \frac{0.51}{0.99}\theta n \leq \frac{2}{3}\theta n$.

□

7.2 Concentration of the marginals about 0, 1

To prove Theorem 22, part 2, we need the following lemma.

Lemma 39. *Suppose that $\theta \leq \rho/(k\ln 2)$. Let (Φ,σ) be a pair chosen from the distribution $\mathcal{U}_k(n,m)$. W.h.p. there is no set of variables $Z \subset V_t$ of size $2kn/2^k \leq |Z| \leq (\mathrm{e}\rho)^{-4}\theta n$ such that each variable in Z supports two clauses under σ, each of which contains an occurrence of a variable in Z that evaluates to 'false' under σ.*

Proof. We work with the planted model $\mathcal{P}'_k(n,m)$. Let p be such that the expected number of clauses is m, i.e., $(2^k-1)\binom{n}{k}p = m$, and $Z \subset V_t$ a set of size z. For any $x \in Z$ we say that it is 'bad' if it supports two clauses under σ, each one of which contains an occurrence of another variable in Z that evaluates to 'false' under σ.
Note that for a fix $x \in Z$, the number of clauses it supports is a random variable with distribution $Bin\Big((z-1)\binom{n}{k-2},p\Big)$. Denote it by N_Z. Thus,

$$P(x \text{ is bad}) = P(N_Z = 2) < \left(z\binom{n}{k-2}p\right)^2.$$

Furthermore, observe that for any x and x' in Z the events $\{x \text{ is bad}\}$ and $\{x' \text{ is bad}\}$ are independent. Hence,

$$P(\{x \text{ is bad}, \forall x \in Z\}) < \left(z\binom{n}{k-2}p\right)^{2z} \leq (\alpha k\theta\rho)^{2z}, \text{ with } \alpha = z/(\theta n),$$

and the probability that there is a bad set of size z is bounded by

$$\binom{\theta n}{z}(\alpha k\theta\rho)^{2z} \leq \left(\frac{\mathrm{e}\theta n}{z}\right)^z (\alpha k\theta\rho)^{2z} = \left[\mathrm{e}\alpha(k\theta\rho)^2\right]^z \leq (\mathrm{e}\alpha\rho^4/\ln^2 2)^z \leq \exp(-z).$$

The assumption on z ensures that this is sufficiently small to move from the planted model to $\mathcal{U}_k(n,m)$ via Corollary 15. □

Proof (Theorem 22, part 2). If $k\theta < \ln(\rho)/2$, then the existence of forced variables immediately implies part 3 of Theorem 7. Thus, let us assume that $\ln(\rho)/2 \leq k\theta \leq \rho/\ln 2$. Let (Φ,σ) be a pair chosen from the distribution $\mathcal{U}_k(n,m)$. Let S be the set of rigid variables. By Theorem 20 part 2, we have $|S| \geq \rho^3\exp(-\rho)\theta n$ w.h.p.Ḋefine an auxiliary bipartite graph $\mathcal{G}$ as follow. The vertices of the graph are the variables in S and the satisfying assignments in $\mathcal{S}(\Phi_t)$. Each variable $x \in S$ is connected with all $\tau \in \mathcal{S}(\Phi_t)$ such that $\tau(x) \neq \sigma(x)$. Since $\mathcal{S}(\Phi_t)$ is $\exp(2-\rho)/k$ condense (part 2 of Theorem 7), then for any two satisfying assignment σ,τ we have

$$dist(\sigma,\tau) \leq \frac{\exp(2-\rho)\theta n}{k}, \tag{41}$$

which implies that $|Z| < (e\rho)^{-4}$ and by the Lemma 39 $|Z| \leq \frac{2kn}{2^k}$. Thus, the number of edges of $\mathcal{G}$ satisfies

$$\frac{|E(\mathcal{G})|}{2} = \sum_x \sum_{\sigma,\tau} I_{\sigma(x)\neq\tau(x)} \leq \frac{2kn}{2^k}|\mathcal{S}(\Phi_t)|^2. \tag{42}$$

By other side, assume $M(x) \notin (0,\epsilon)\cup(1-\epsilon,1]$, then

$$\sum_{\sigma,\tau} I_{\sigma(x)\neq\tau(x)} \geq \epsilon(1-\epsilon)|\mathcal{S}(\Phi_t)|^2. \tag{43}$$

Let ν be the number of variables which satisfies (43), thus

$$\sum_x \sum_{\sigma,\tau} I_{\sigma(x)\neq\tau(x)} \geq \nu\epsilon(1-\epsilon)|\mathcal{S}(\Phi_t)|^2, \tag{44}$$

and (42) and (43) together imply

$$\nu \geq \frac{2kn}{\epsilon(1-\epsilon)2^k}$$

.

□

Proof (Theorem 7, part 4). This follows directly by applying Lemma 39 to the self-contained set obtained in subsection 5.2. □

8 Proof of Theorem 10

8.1 Quasi-randomness properties

The proof of Theorem 10 is based on results from [8]. These results show that, in order to obtain Theorem 10, we essentially have to verify that the outcome $\boldsymbol{\Phi}_t$ of the experiment **U1–U4** enjoys certain quasi-randomness properties. We begin by stating the necessary properties. To this end, we define

$$\delta_t = \exp(-c\theta k), \tag{45}$$

where $c > 0$ is a small absolute constant (independent of k, r, t, n).
Fix a k-CNF Φ and an assignment $\sigma \in \{0,1\}^V$. Let $\Phi_{t,\sigma}$ denote the CNF obtained from Φ by substituting $\sigma(x_1), \ldots, \sigma(x_k)$ for $x_1, \ldots, x_t$ and simplifying. Let $G = G(\Phi_{t,\mathrm{id},\mathbf{1}})$ denote the factor graph. For a variable $x \in V_t$ and a set $Q \subset V_t$ let

$$N_{\leq 1}(x, Q) = \{b \in N(x) : |N(x) \cap Q \setminus \{x\}| \leq 1 \wedge 0.1\theta k \leq |N(b)| \leq 10\theta k\}. \tag{46}$$

Thus, $N_{\leq 1}(x, Q)$ is the set of all clauses that contain x (which may or may not be in Q) and at most one other variable from Q. In addition, there is a condition on the *length* $|N(b)|$ of the clause b in the decimated formula $\Phi_{t,\sigma}$. Observe that having assigned the first t variables, we should 'expect' the average clause length to be θk. For a linear map $\Lambda : \mathbf{R}^{V_t} \to \mathbf{R}^{V_t}$ let $\|\Lambda\|_\square$ signify the norm

$$\|\Lambda\|_\square = \max_{\zeta \in \mathbf{R}^{V_t} \setminus \{0\}} \frac{\|\Lambda\zeta\|_1}{\|\zeta\|_\infty}.$$

Definition 40. *Let $\delta > 0$. We say that (Φ, σ) is (δ, t)-**quasirandom** if Φ satisfies* **Q0** *and $\Phi_{t,\sigma}$ satisfies* **Q1–Q4** *below.*

Q0. *There are no more than $\ln\ln n$ redundant clauses. Moreover, no variable occurs in more than $\ln n$ clauses of Φ.*

Q1. *No more than $10^{-5}\delta\theta n$ variables occur in clauses of length less than $\theta k/10$ or greater than $10\theta k$. Moreover, there are at most $10^{-4}\delta\theta n$ variables $x \in V_t$ such that*

$$(\theta k)^3\delta \cdot \textstyle\sum_{b \in N(x)} 2^{-|N(b)|} > 1.$$

Q2. *If $Q \subset V_t$ has size $|Q| \leq \delta\theta n$, then there are no more than $10^{-4}\delta\theta n$ variables x such that either*

$$\sum_{b \in N(x) : |N(b) \cap Q \setminus \{x\}| = 1} 2^{-|N(b)|} > \rho(\theta k)^5\delta, \text{ or} \tag{47}$$

$$\sum_{b \in N(x) : |N(b) \cap Q \setminus \{x\}| > 1} 2^{|N(b) \cap Q \setminus \{x\}| - |N(b)|} > \frac{\delta}{\theta k}, \text{ or} \tag{48}$$

$$\left| \sum_{b \in N_{\leq 1}(x, Q)} \frac{\mathrm{sign}(x, b)}{2^{|N(b)|}} \right| > \frac{\delta}{1000}. \tag{49}$$

Q3. *For any $0.01 \leq z \leq 1$ and any set $Q \subset V_t$ of size $0.01\delta(n-t) \leq |Q| \leq 100\delta(n-t)$ we have*

$$\sum_{b : |N(b) \cap Q| \geq z|N(b)|} |N(b)| \leq 1.01|Q|/z.$$

Q4. *For any set $Q \subset V_t$ of size $|Q| \leq 10\delta(n-t)$ the linear operator*

$$\Lambda_Q : \mathbf{R}^{V_t} \to \mathbf{R}^{V_t},\ \Gamma \mapsto \left(\sum_{b \in N_{\leq 1}(x, Q)} \sum_{y \in N(b) \setminus \{x\}} 2^{-|N(b)|} \cdot \mathrm{sign}(x, b)\mathrm{sign}(y, b)\Gamma_y \right)_{x \in V_t} \tag{50}$$

has norm $\|\Lambda_Q\|_\square \leq \delta^4\theta n$.

With respect to **Q0**, we have

Lemma 41 ([8]). *The random formula $\boldsymbol{\Phi}$ satisfies condition* **Q0** *w.h.p., for any density $0 < r = m/n \leq 2^k \ln 2$.*

Let Φ be a k-CNF and let $\delta > 0$. For a number $\delta > 0$ and an index $l > t$ we say that x_l is (δ, t)-*biased* if the result $\mu_{x_l}(\Phi_{t,\sigma}, \omega)$ of the BP computation on $\Phi_{t,\sigma}$ differs from $\frac{1}{2}$ by more than δ, i.e.,

$$|\mu_{x_l}(\Phi_{t,\sigma}, \omega) - 1/2| > \delta.$$

Moreover, (Φ, σ) is (δ, t)-*balanced* if no more than $\delta\theta n$ variables are (δ, t)-biased.

Theorem 42 ([8]). *There is $\rho_0 > 0$ such that for any k, r satisfying $\rho_0 \cdot 2^k/k \leq r \leq 2^k \ln 2$ and n sufficiently large the following is true. Suppose (Φ, σ) is (δ_t, t)-quasirandom for some $1 \leq t \leq T = (1 - \ln(\rho)/(c^2 k))n$. Then (Φ, σ) is (δ_t, t)-balanced.*

At the end of this section, we will verify that random formulas chosen from the distribution $\mathcal{P}'_k(n, m)$ are indeed quasirandom.

Proposition 43. *There exists a constant $\rho_0 > 0$ such that for any k, r satisfying $\rho_0 \cdot 2^k/k \leq r \leq 2^k \ln 2$ there is $\xi = \xi(k, r) > 0$ so that for n large and δ_t, T as in Theorem 42 the following is true. Let (Φ, σ) be a pair chosen from the planted model $\mathcal{P}'_k(n, m)$, given that $\sigma = \mathbf{1}$ is the all-true assignment. Then*

$$\mathrm{P}\left[(\Phi, \sigma) \text{ is } (\delta_t, t)\text{-quasirandom} | \mathbf{Q0}\right] \geq 1 - \exp\left[-\rho 2^{1-k} n\right]$$

for any $1 \leq t \leq T$.

Finally, Theorem 10 follows by combining Corollary 13, Theorem 42, and Proposition 43.

Proof of Proposition 43: Let $\boldsymbol{\Phi}' = \boldsymbol{\Phi}'_k(n, m)$ be a random formula obtained by including each possible clause with probability $p = m/(2^k \binom{n}{k})$ independently.

Proposition 44 ([8, Appendix E]). *There exists a constant $\rho_0 > 0$ such that for any k, r satisfying $\rho_0 \cdot 2^k/k \leq r \leq 2^k \ln 2$ for n large and δ_t, T as in (45) the following properties hold for a random formula $\boldsymbol{\Phi}'$ with probability at least $1 - \exp\left[-10 \sum_{s \leq t} \delta_s\right]$ for any $1 \leq t \leq T$, given that $\boldsymbol{\Phi}'$ satisfies* **Q0**.

1. **Q1** *and* **Q3** *are satisfied.*
2. *For any set Q of size $|Q| \leq \delta\theta n$ there are at most $10^{-5}\delta\theta n$ variables x that satisfy either (47), (48), or*

$$\left| \sum_{b \in N_{\leq 1}(x, Q)} \frac{\mathrm{sign}(x, b)}{2^{|N(b)|}} \right| > \frac{\delta}{2000}. \tag{51}$$

3. *For any Q the operator Λ_Q from (50) satisfies $\|\Lambda_Q\|_\square \leq \delta^4(n - t)/2$*

Let $\boldsymbol{\Phi}_t$ be the formula obtain from $\boldsymbol{\Phi}$ by substituting the value 'true' for $x_1, \ldots, x_{t-1}$ and simplifying. Since the δ_s form a geometric sequence, we have

$$\Sigma_t = \sum_{s \leq t} \delta_s \sim \frac{n}{ck \exp(c\theta k)}.$$

Observe that

$$\theta\delta n > 10^{15} \Sigma_t$$

if $\rho \geq \rho_0$ is chosen sufficiently large.

Lemma 45. *There exists a constant $\rho_0 > 0$ such that for any k, r satisfying $\rho_0 \cdot 2^k/k \leq r \leq 2^k \ln 2$ the following is true for the random formula $\boldsymbol{\Phi}'$ with probability at least $1 - \exp(-\rho 2^{2-k} n)$.*

1. *The total number of all-negative clauses is bounded by $2^{1-k} m$.*
2. *For each variable $x \in V_t$ let N_x be the number of all-negative clauses in which x appears. Then the number of variables $x \in V_t$ with $N_x > 2^{0.01\theta k}$ is bounded by $\delta^2 \theta n$.*

Proof. The first assertion simply follows from Chernoff bounds. With respect to the second assertion, assume that the first claim occurs, i.e., the total number of all-negative clauses is bounded by $2^{1-k}m = 2\rho n/k$. Then for each variable the average number of occurrences in such clauses is bounded by 2ρ. Therefore, the total number of variables that occur more than $2^{0.01\theta k}$ times is bounded by $2\rho \cdot 2^{-0.01\theta k}n$. By symmetry, the number of such variables that are amongst the last θn variables is (asymptotically) binomially distribution with mean $2\rho \cdot 2^{-0.01\theta k}\theta n$. Therefore, the second assertion follows from Chernoff bounds. □

Proof (Proposition 43). Let (Φ, σ) be a random pair chosen from the distribution $\mathcal{P}'_k(n, m)$. We may assume without loss of generality that σ is the all-true assignment. Thus, the formula Φ is obtained by including each clause that does not consist of negative literals only with probability $p = m/((2^k - 1)\binom{n}{k})$ independently. Now, let Φ' be the formula obtained by addition to Φ each of the $\binom{n}{k}$ all-negative clauses independently with probability p. Then Φ' has distribution $\boldsymbol{\Phi}'_k(n, m \cdot \frac{2^k}{2^k-1})$. Thus, with probability at least $1 - \exp[-10\sum_{s\leq t}\delta_s]$ the formula Φ' has the properties 1.–3. from Proposition 44. Let us condition on this event.
Since Φ' contains Φ as a sub-formula, the fact that Φ' enjoys properties **Q1** and **Q3** implies that the same is true of Φ. Furthermore, any variable x for which either (47) or (48) is true in Φ has the same property in Φ' (because the expressions on the left hand side are monotone with respect to the addition of clauses). With respect to the expression in (49), we decompose the sum for the pair (Φ, σ) as

$$S_x(\Phi, \sigma) = S_x(\Phi', \sigma) - R_x,$$

where R_x sums over all clauses that are in Φ' but not in Φ. Due to **Q1**, we may assume that only clauses of length at least $0.1\theta k$ occur in the sum R_x. Thus, letting N_x denote the number of clauses in $\Phi' \setminus \Phi$ containing x, we get $|R_x| \leq 2^{-0.1\theta k}N_x$. The second part of Lemma 45 implies that for all but $\delta^2\theta n$ variables we have $N_x \leq 2^{0.01\theta k}$. Hence, R_x is tiny for all but $\delta^2\theta n$ variables. This shows that Φ satisfies **Q2**.
With respect to **Q4**, let D be the difference of the two linear operators for Φ and Φ'. Only clauses of length at least $0.1\theta k$ and at most $10\theta k$ contribute to D. Hence, letting N denote the number of all-negative clauses, we have

$$\|D\|_\square \leq 2^{-0.1\theta k}(10\theta k)^2 N.$$

Since $N \leq 2^{1-k}m = 2\rho n/k$ by Lemma 45, we thus get

$$\|D\|_\square \leq 200\theta n(\theta k)2^{-0.1\theta k}.$$

Hence, the third part of Proposition 44 implies that Φ satisfies **Q4**. □

References

1. D. Achlioptas, A. Coja-Oghlan: Algorithmic barriers from phase transitions. Proc. 49th FOCS (2008) 793–802.
2. D. Achlioptas, A. Coja-Oghlan, F. Ricci-Tersenghi: On the solution space geometry of random formulas. Random structures and algorithms, awaiting publication.
3. D. Achlioptas, C. Moore: Random k-SAT: two moments suffice to cross a sharp threshold. SIAM Journal on Computing **36** (2006) 740–762.
4. D. Achlioptas, Y. Peres: The threshold for random k-SAT is $2^k \ln 2 - O(k)$. Journal of the AMS **17** (2004) 947–973.
5. D. Achlioptas, F. Ricci-Tersenghi: Random formulas have frozen variables. SIAM J. Comput. **39** (2009) 260–280.
6. A. Braunstein, M. Mézard, R. Zecchina: Survey propagation: an algorithm for satisfiability. Random Structures and Algorithms **27** (2005) 201–226.
7. A. Coja-Oghlan: A better algorithm for random k-SAT. SIAM J. Computing **39** (2010) 2823–2864.
8. A. Coja-Oghlan: On belief propagation guided decimation for random k-SAT. Proc. 22nd SODA (2011) 957–966.
9. H. Daudé, M. Mézard, T. Mora, R. Zecchina: Pairs of SAT-assignments in random Boolean formulae. Theoretical Computer Science **393** (2008) 260–279.
10. A. Frieze, S. Suen: Analysis of two simple heuristics on a random instance of k-SAT. Journal of Algorithms **20** (1996) 312–355.
11. M. Hajiaghayi, G. Sorkin: The satisfiability threshold of random 3-SAT is at least 3.52. IBM Research Report RC22942 (2003).
12. S. Janson, T. Łuczak, A. Ruciński: Random Graphs, Wiley 2000.
13. A. Kaporis, L. Kirousis, E. Lalas: The probabilistic analysis of a greedy satisfiability algorithm. Random Structures and Algorithms **28** (2006) 444–480.
14. L. Kroc, A. Sabharwal, B. Selman: Message-passing and local heuristics as decimation strategies for satisfiability. Proc 24th SAC (2009) 1408–1414.

15. F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborova: Gibbs states and the set of solutions of random constraint satisfaction problems. Proc. National Academy of Sciences **104** (2007) 10318–10323.
16. S. Mertens, M. Mézard, R. Zecchina: Threshold values of random K-SAT from the cavity method. Random Struct. Alg. **28** (2006) 340–373.
17. M. Mézard, G. Parisi, R. Zecchina: Analytic and algorithmic solution of random satisfiability problems. Science **297** (2002) 812–815.
18. D. Mitchell, B. Selman, H. Levesque: Hard and easy distribution of SAT problems. Proc. 10th AAAI (1992) 459–465.
19. A. Montanari, F. Ricci-Tersenghi, G. Semerjian: Solving constraint satisfaction problems through Belief Propagation-guided decimation. Proc. 45th Allerton (2007).
20. F. Ricci-Tersenghi, G. Semerjian: On the cavity method for decimated random constraint satisfaction problems and the analysis of belief propagation guided decimation algorithms. J. Stat. Mech. (2009) P09001.