Fast sampling of satisfying assignments from random k-SAT with applications to connectivity*

Zongchen Chen[†] Andreas Galanis [‡] Leslie Ann Goldberg [‡] Heng Guo [§]
Andrés Herrera-Poyatos [‡¶] Nitya Mani[†] Ankur Moitra[†]

January 2023

Abstract

We give a nearly linear-time algorithm to approximately sample satisfying assignments in the random k-SAT model when the density of the formula scales exponentially with k. The best previously known sampling algorithm for the random k-SAT model applies when the density $\alpha = m/n$ of the formula is less than $2^{k/300}$ and runs in time $n^{\exp(\Theta(k))}$ (Galanis, Goldberg, Guo and Yang, SIAM J. Comput., 2021). Here n is the number of variables and m is the number of clauses. Our algorithm achieves a significantly faster running time of $n^{1+o_k(1)}$ and samples satisfying assignments up to density $\alpha \leq 2^{0.039k}$.

The main challenge in our setting is the presence of many variables with unbounded degree, which causes significant correlations within the formula and impedes the application of relevant Markov chain methods from the bounded-degree setting (Feng, Guo, Yin and Zhang, J. ACM, 2021; Jain, Pham and Vuong, 2021). Our main technical contribution is a $o_k(\log n)$ bound of the sum of influences in the k-SAT model which turns out to be robust against the presence of high-degree variables. This allows us to apply the spectral independence framework and obtain fast mixing results of a uniform-block Glauber dynamics on a carefully selected subset of the variables. The final key ingredient in our method is to take advantage of the sparsity of logarithmic-sized connected sets and the expansion properties of the random formula, and establish relevant connectivity properties of the set of satisfying assignments that enable the fast simulation of this Glauber dynamics.

Our results also allow us to conclude that, with high probability, a random k-CNF formula with density at most $2^{0.227k}$ has a giant component of solutions that are connected in a graph where solutions are adjacent if they have Hamming distance $O_k(\log n)$. We are also able to deduce looseness results for random k-CNFs in the same regime.

[†]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 20139, USA

[‡]Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK.

[§]School of Informatics, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK. HG has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947778).

[¶]This author is supported by an Oxford-DeepMind Graduate Scholarship and a EPSRC Doctoral Training Partnership.

1 Introduction

The random k-SAT model is a foundational model in the study of randomised algorithms. For integers $k, n, m \geq 2$, the random formula $\Phi = \Phi(k, n, m)$ is a k-CNF formula chosen uniformly at random from the set of formulae with n Boolean variables and m clauses, where each clause has k literals (repetitions allowed). Here, we consider the sparse regime where the density of the formula, $\alpha = m/n$, is bounded by an absolute constant. An important question is determining the probability that the random formula is satisfiable as a function of its density (in the limit $n \to \infty$). Interestingly, for all sufficiently large k, the probability that Φ is satisfiable drops abruptly from 1 to 0 when the density α crosses a certain threshold $\alpha_\star(k)$. Recently there has been tremendous progress in establishing this phase transition, concluding that $\alpha_\star(k) = 2^k \log 2 - \frac{1}{2}(1+\log 2) + o_k(1)$ as $k \to \infty$ [19, 16]. Despite the good progress on pinning down this phase transition, finding satisfying assignments for densities up to α^* poses severe challenges. In fact, the best known algorithm [12] for finding a satisfying assignment of a random formula Φ succeeds up to densities $(1 + o_k(1))\frac{2^k}{k} \log k$, and going beyond such densities is a major open problem with links to phase transitions [1].

Lately there has been significant interest in the related computational problem of sampling satisfying assignments of Φ uniformly at random. This problem is closely connected to the problem of estimating the number of satisfying assignments of Φ, also known as the value of the partition function of the model. From a probabilistic viewpoint, the analysis of the partition function depends on subtle properties of the solution set $\Omega = \Omega_\Phi$ consisting of the satisfying assignments of Φ [2, 14, 45, 40]. In this direction, there has been substantial work on finding the so-called free energy of the model, i.e., the asymptotic value of the quantity $\frac{1}{n}\mathbf{E}[\log(1 + |\Omega|)]$. Computing the k-SAT free energy is a difficult problem which is still open (roughly, the difficulty comes from the asymmetry of the model and the unbounded degrees), but there have been results for closely related models including the permissive version of the model [14, 40, 17], the regular k-SAT model [18], and the regular NAE-SAT model [44, 45]. Very recently, a formula for the free energy of the 2-SAT model was given in [2].

Regarding the algorithmic problem of sampling satisfying assignments uniformly at random, in the random k-SAT model progress has been slower relative to other well-studied models on random graphs (such as k-colourings or independent sets). One of the main reasons for this is that the usual distribution properties that are typically used to obtain fast algorithms (such as correlation decay and spatial mixing) fail to hold for densities as low as $\alpha = o_k(1)$ [40]. These issues are in fact present already in the bounded-degree k-SAT setting, where the formulae are worst-case but every variable is constrained to have a bounded-number of occurrences. For random formulae, these issues are further aggravated by the fact that the degrees of a linear number of variables are unbounded. Very recently, the authors of [24] gave an approximate counting algorithm (FPTAS) for the number of satisfying assignments of Φ when k is large enough and $\alpha \lesssim 2^{k/300}$ (where $\lesssim$ hides a polynomial factor in $1/k$). This algorithm elevates Moitra's counting method for bounded-degree k-SAT [39] to the random formula setting, and is the first polynomial-time approximate-counting algorithm to achieve an exponential-in-k bound on α. However, its running time is $n^{\exp(\Theta(k))}$ because the algorithm repeatedly has to enumerate local structures (including solving LPs as a subroutine), which does not scale well with k. Hence, the problem of finding a fast algorithm for sampling the satisfying assignments in the random k-SAT model has remained open.

In this work we give a fast algorithm that in time $n^{1+o_k(1)}$ approximately samples satisfying assignments of a random k-SAT formula of density $\alpha \leq 2^{0.039k}$, within arbitrarily small polynomial error. Our work also delves into the connections between the solution space geometry of k-CNF Φ and algorithms for efficiently sampling from the solutions of Φ.

A unifying theme of previous approaches to counting and sampling CSP solutions is a tool called marking, first introduced in [39], which finds a set of "marked" variables such that the set of satisfying assignments projected on these variables is connected. Marking is also an essential step in the developing of our sampling algorithm. Our algorithm first runs a Markov chain to sample assignments of a judiciously-chosen subset of marked variables of Φ (from the relevant marginal distribution), and subsequently extending this random assignment to all the variables. This has the advantage that it avoids the enumeration of local structures, and in fact achieves a nearly-linear running time. We give a high-level overview of the techniques developed in our proofs in Section 2. Roughly, our Markov chain is a uniform-block Glauber dynamics which, interestingly, mixes quickly despite the presence of high-degree variables in the random formula. The main point of departure from similar approaches that have been applied to the bounded-degree setting is that we completely circumvent sophisticated coupling arguments that have been used there and which are unfortunately severely constricted by the unbounded degrees in our setting (and made inapplicable). Instead, our main technical contribution is to show that the stationary distribution of our chain is $(c^k \log n)$-spectrally independent for some constant $c \in (0,1)$, allowing us to apply recently-developed tools in the analysis of Markov chains. Unlike most applications of spectral independence, our proof does not rely on correlation decay (which, as we mentioned, fails to hold for densities exponential in k). We show our spectral-independence bounds by relating the probabilistic properties of the solution space with the structure of the formula using coupling techniques, so that we can exploit local sparsity properties of random k-SAT.

To formally state our main result, we say that an event $\mathcal{E}$ regarding the choice of the random formula Φ holds with high probability (abbreviated w.h.p.) if $\Pr(\mathcal{E}) = 1 - o(1)$ as $n \to \infty$. The total variation distance between two probability distributions μ and ν over the same space Ω is given by $\frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|$ and is denoted by $d_{\mathrm{TV}}(\mu,\nu)$. Our main result can now be stated as follows.

Theorem 1. For any real $\theta \in (0,1)$, there is $k_0 \geq 3$ with $k_0 = O(\log(1/\theta))$ such that, for any integers $k \geq k_0$ and $\xi \geq 1$, and for any positive real $\alpha \leq 2^{0.039k}$, the following holds.

There is an efficient algorithm to sample from the satisfying assignments of a random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$ within $n^{-\xi}$ total variation distance of the uniform distribution. The algorithm runs in time $O(n^{1+\theta})$, and succeeds w.h.p. over the choice of Φ.

Using standard techniques from the literature, this $O(n^{1+\theta})$ uniform sampling algorithm can be used to obtain a randomised approximation scheme for counting satisfying assignments of Φ in time $O(n^{2+\theta}/\varepsilon^2)$, where ε is the multiplicative error, see [21, Section 7] and Remark 56 for details.

Our results can be applied to analyse the solution space geometry of random k-CNF formulae for the densities under consideration. Many involved heuristics in statistical physics make predictions about the geometry of the solution space of a random k-CNF instance, often depicted in diagrams like Figure 1. Some phases and transitions in this diagram are precisely understood. For example, as mentioned above, the satisfiability threshold (pictured in the transition to the rightmost image in Figure 1) was determined by [19]. Another transition of interest is the clustering threshold, above which the solution space of a random k-CNF shatters into exponentially many linearly separated connected components, each of which contains an exponentially small fraction of the satisfying assignments of the formula, as rigorously understood in [15, 3, 37, 41].

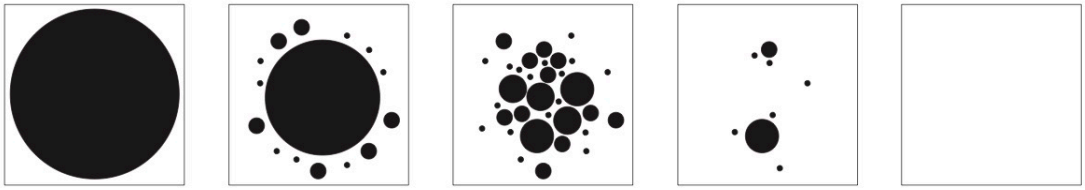

Figure 1: Heuristic phase diagrams such as above [36] depict the predicted evolution of the structure of the solution space of a random k-CNF as the density α of the formula increases from left to right. We primarily study the leftmost regime.

In the lower-density regime, the solution space geometry of random k-CNFs appears poorly understood. It is widely believed that beneath a critical clause density, the solution space of a random k-CNF is "connected." However, from the literature, it is not even clear what "connected" means. Connectivity is sometimes used in the statistical physics literature as a characterization of the entropy or energy profile of the solution space of a random k-CNF formula as in [48]. In such settings, connectivity is often characterized by an absence of clustering behavior, leaving somewhat of a mystery as to the graphical properties of the solution space of a low density random k-CNF.

Conjectures about connectivity take different forms, and different notions of what connectivity might mean are articulated in [48, 36, 15]. The most common precise notion of connectivity is with respect to Hamming distance, i.e. understanding connectivity properties of the graph of solutions to a random k-CNF, where solutions are $f(n)$-connected if their Hamming distance is at most $f(n)$. At lower densities, random k-CNFs still can have isolated solutions far in Hamming distance from other satisfying assignments. However, the prevailing belief is that below some threshold, the overwhelming majority of solutions to a random k-CNF lie in a giant component that is $o(n)$-connected.

Much more is known about related notions and local versions of connectivity, like looseness, which characterises how rigid a particular satisfying assignment is. Roughly speaking, a satisfying assignment to a formula is $f(n)$-loose if any variable can be flipped to yield a new satisfying assignment by changing at most $f(n)$ additional variable assignments. In [1], the authors showed $o(n)$-looseness holds in the connectivity regime for related, simpler random models, random q-coloring, and hypergraph 2-coloring, conjecturing that $o(n)$-looseness holds for random k-CNF instances below the clustering threshold. This conjecture was partially resolved in [15], where in an analysis of the decimation process for random k-SAT, the authors observed that with high probability over formulae and satisfying assignments, at least 99% of the variables were $O(\log n)$-loose. Looseness, however, is a local notion, not a global one. The set of elements in $\{0,1\}^n$ that have Hamming weight at least $2n/3$ or at most $n/3$ is 1-loose, but $\Omega(n)$-connected.

We will concern ourselves with the following precise notion of connectivity.

Definition 2 (D-Connectivity). Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. For any assignment $\Lambda \colon \mathcal{V} \to \{\mathsf{F}, \mathsf{T}\}$, let $\|\Lambda\|_1$ be the number of variables Λ assigns to be T. Throughout, we implicitly consider variable assignments in $\mathbb{F}_2^n$, so $\|\cdot\|_1$ encodes Hamming weight and $\|\Lambda_1 - \Lambda_2\|_1$ encodes Hamming distance.

We say a sequence of satisfying assignments $\zeta_0 \leftrightarrow \zeta_1 \leftrightarrow \cdots \leftrightarrow \zeta_\ell$ of Φ is a D-path if $\|\zeta_i - \zeta_{i-1}\|_1 \leq D$ for each $i \in [t]$. We say two satisfying assignments of Φ, $\Lambda, \Lambda' \in \Omega$, are D-connected if there exists a D-path connecting Λ and Λ' (that is, $\zeta_0 = \Lambda$ and $\zeta_\ell = \Lambda'$).

Marking-based deterministic and MCMC algorithms are mysterious at first glance, as they enable counting and sampling of k-CNF solutions even in regimes where the solution space is

disconnected (i.e. not 1-connected). In this work, we leverage the idea of marking in a novel way to construct paths that certify global connectivity properties of the solution space of k-CNFs at densities close to where counting algorithms are known.

Theorem 3. There is $k_0 \geq 3$ and a polynomial $p(k)$ with non-negative integer coefficients such that, for any integer $k \geq k_0$, and for any positive real $\alpha \leq 2^{0.227k}$, the following claim holds with high probability over the choice of a random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Two satisfying assignments chosen uniformly at random are $p(k)\log(n)$-connected with probability at least $1 - 1/n$.

In fact, we show it suffices to take $p(k) = 2k^5$. Our new applications of marking also have implications for other, more local, structural properties of the k-CNF solution space, like looseness.

Definition 4. Given a k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$ and a satisfying assignment Λ, a variable $v \in \mathcal{V}$ is $f(n)$-loose with respect to Λ if there exists satisfying assignment to Φ, $\tau \in \Omega$, with $\tau(v) \neq \Lambda(v)$ and $\|\Lambda - \tau\|_1 \leq f(n)$.

For a random k-CNF formula $\Phi = \Phi(k, n, m)$ and a satisfying assignment Λ chosen uniformly at random, we say that Φ is $f(n)$-loose if with high probability over (Φ, Λ), all variables $v \in V$ are $f(n)$-loose with respect to Λ.

We observed earlier that looseness does not imply connectivity; in fact, the other direction of implication is also false as looseness is an incomparable goal to connectivity. Looseness requires that locally, we are able to flip any variable and get to a nearby solution rather than merely the existence of a path away from a solution. Nonetheless, we are able to deduce some nontrivial results about the looseness of the solution space of random k-CNFs.

Theorem 5. There is $k_0 \geq 3$ such that, for any integer $k \geq k_0$, and for any positive real $\alpha \leq 2^{0.227k}$, the random k-CNF formula $\Phi(k, n, \lfloor \alpha n \rfloor)$ is $\mathrm{poly}(k)\log(n)$-loose.

We note here that, independently of this work, He, Wu, and Wang [28] also obtained sampling algorithms for random k-CNF formulae. The approach of [28] is based on bounding chains following the recursive sampler method developed in [6, 27, 26]. Their algorithm works up to densities roughly equal to $2^{k/3}$ and samples satisfying assignments within ε total variation distance of the uniform distribution in time $(n/\varepsilon)^{1+O(k^{-5})}$.

2 Proof outline

Our nearly linear-time sampling algorithm is based on running a Markov chain; this is a standard technique in approximate counting, where typically one runs a Markov chain on the whole state space that converges to the desired distribution. The twist in k-SAT is that the state space of the Markov chain needs to be carefully selected in order to avoid certain bottleneck phenomena that impede fast convergence. This approach has been recently applied to bounded-degree k-CNF formulae [21, 22, 31] building on the work of Moitra [39] (see also [32]) and using the Markov chain known as single-site Glauber dynamics. The main difficulties in all of these works are that the usual distribution properties that are typically used to obtain fast algorithms (such as correlation decay and spatial mixing) fail on the set of all SAT solutions, and in fact even ensuring a connected state space is a major problem. Working around this is one of the main challenges for us too, and in the random k-SAT setting it is further aggravated by the fact that a linear number of variables have degrees much higher than average. In fact, w.h.p., a good portion of vertices have degrees depending on n. with the maximum degree of the formula scaling as $\log n / \log\log n$.

This poses several new challenges for the Markov chain approach to work in our setting. First of all, we have to ensure that the set of satisfying assignments that our Markov chain considers has good connectivity properties. We address this problem in Section 2.1 of this proof outline, where we find a suitable subset of marked variables where we can run the Glauber dynamics; this part is inspired by Moitra's "marking" approach, though here we need to add an extra layer of marking to facilitate later the analysis of the Markov chain. Second and more importantly, state-of-the-art arguments for bounding the mixing time of the single-site Glauber dynamics on k-CNF formulae, such as [21, 31] break under the presence of high-degree variables. We focus on this in Section 2.2, where we outline a novel argument that analyses the mixing time of the uniform-block Glauber dynamics using recent advances in spectral independence [5, 34, 7, 10]. This is the first application of the spectral-independence framework for k-CNF formulae, where the absence of correlation decay limits the application of standard techniques (based on self-avoiding walk trees [7, 10]). To obtain our spectral-independence bounds we need to combine the probabilistic structure of satisfying assignments with the local sparsity properties of the random formula. The third challenge in our approach is simulating the individual steps of the uniform-block Glauber dynamics since they involve updating a linear number of variables, making the computation of the transition probabilities more challenging. To this end, we need to initialise our block Glauber dynamics to random values (instead of an arbitrary assignment that is typically used as initialisation), and show that the formula breaks into small tree-like connected components that allows us to do the relevant computations throughout the algorithm's execution (cf. Section 2.3). Based on these pieces, the full algorithm is presented in Section 2.4.

The fact that the formula breaks into small tree-like connected components when marked variables are assigned random values will also allow us to analyse the geometry of the space of satisfying assignment of the random formula, and we will delve into this connection in Section 2.3.

2.1 Marking variables in the random k-SAT model

In order to ensure good connectivity properties which are essential for fast convergence of the relevant Markov chain, our algorithm runs Glauber dynamics on a large subset $\mathcal{V}_{\mathrm{m}}$ of so-called "marked" variables of the random formula, leaving the rest of the variables unassigned. The variables in $\mathcal{V}_{\mathrm{m}}$ are chosen in a way that ensures that their marginals are near $1/2$, which is important for ensuring rapid mixing. Moitra [39] introduced a random "marking" procedure to identify such a subset of variables in the bounded-degree case. The presence of high-degree variables impedes a direct application of this technique in the random-formula setting, but in [24] the authors show that by temporarily removing a small linear number of "bad" clauses that contain high-degree variables, one can also achieve marginals near $1/2$ for an appropriate set of variables in the random k-SAT model. Here, we further refine these arguments, as we need more control over the high-degree variables of the formula in order to conclude rapid mixing of the Glauber dynamics. Recall that the degree of a variable v is the number of occurrences of literals involving the variable v in Φ and that the maximum degree of the formula Φ is the maximum degree among its variables. The following important definitions will be used throughout the paper. We usually use $\mathcal{V}$ to denote the set of variables and $\mathcal{C}$ to denote the set of clauses of a k-CNF formula Φ. For any $c \in \mathcal{C}$ we denote by $\mathrm{var}(c)$ the set of variables appearing in c, and for any $S \subseteq \mathcal{C}$ we denote $\mathrm{var}(S) = \bigcup_{c \in S} \mathrm{var}(c)$.

Definition 6 (high-degree, Δ_r). Let $r \in (0,1)$ and let $k \geq 3$ be an integer. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. We say that a variable $v \in \mathcal{V}$ is high-degree if the degree of v is at least $\Delta_r := \lceil 2^{rk} \rceil$.

We refer to Section 4 for details on our procedure to determine the bad variables/clauses of the formula Φ. Roughly, bad variables consist of high-degree variables (as in Definition 6), plus those

variables that appear in a clause with at least two other bad variables (recursively); bad clauses are those clauses that contain at least three bad variables. We use $\mathcal{V}_{\text{bad}}(r)$ and $\mathcal{C}_{\text{bad}}(r)$ to denote the sets of bad variables and clauses. We use $\mathcal{V}_{\text{good}}(r) = \mathcal{V} \setminus \mathcal{V}_{\text{bad}}(r)$ to denote the set of good variables, and $\mathcal{C}_{\text{good}}(r) = \mathcal{C} \setminus \mathcal{C}_{\text{bad}}(r)$ to denote the set of good clauses. The following proposition, proved in Section 4, summarises the main properties of the above sets.

Proposition 7. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. For any $c \in \mathcal{C}_{\text{good}}(r)$, we have $|\text{var}(c) \cap \mathcal{V}_{\text{bad}}(r)| \leq 2$, and for any $c \in \mathcal{C}_{\text{bad}}(r)$, we have $|\text{var}(c) \cap \mathcal{V}_{\text{good}}(r)| = 0$. Moreover, every good variable has degree less than Δ_r. There is a procedure to determine $\mathcal{C}_{\text{bad}}$ that runs in time $O(n + mk)$, where n is the number of variables of Φ and m is the number of clauses of Φ.

It turns out that, w.h.p. over the choice of Φ, most clauses (and variables) in the random formula Φ are good, see Lemma 20 for a precise statement. At this stage, it would be natural to try to rework the Markov chain approach of [21]. To do this, we would split the set of good variables into marked variables and control variables in such a way that marked variables have marginals close to $1/2$. Then we run the Glauber dynamics on the set of marked variables. However, as we explain in Section 2.2, the state-of-the-art techniques used to analyse the mixing time of the single-site Glauber dynamics on bounded-degree formulae do not generalise to the random k-SAT setting; the main reason for this is that they fail to capture the effect that the high-degree variables have on the marginal probabilities of other variables. Therefore, we need to develop an alternative approach that is robust against the presence of high-degree variables. Our main contribution is an argument to apply the spectral independence framework [10, 11] to the random k-SAT model that leads to nearly linear sampling algorithms. To do this, it is important to introduce a third type of good variables, which we call the auxiliary variables. This motivates the following definition of marking.

Definition 8 (ρ-distributed, $(r, r_{\text{m}}, r_{\text{a}}, r_{\text{c}})$-marking, r_0, r_1, δ). Let $r \in (0, 1)$. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula and let V be a subset of $\mathcal{V}_{\text{good}}(r)$. We say that V is ρ-distributed if for each $c \in \mathcal{C}_{\text{good}}(r)$ we have $|\text{var}(c) \cap V| \geq \rho(k-3)$. An $(r, r_{\text{m}}, r_{\text{a}}, r_{\text{c}})$-marking of Φ is a partition $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ of the variables of Φ such that

1. the set of good variables $\mathcal{V}_{\text{m}}$ is r_{m}-distributed;
2. the set of good variables $\mathcal{V}_{\text{a}}$ is r_{a}-distributed.
3. $\mathcal{V}_{\text{c}}$ contains all the bad variables and the set $\mathcal{V}_{\text{c}} \setminus \mathcal{V}_{\text{bad}}(r)$ is r_{c}-distributed;

The variables in $\mathcal{V}_{\text{m}}$ are called marked variables, the variables in $\mathcal{V}_{\text{a}}$ are called auxiliary variables, and the variables in $\mathcal{V}_{\text{c}}$ are called control variables.

In our sampling algorithm we work with $r = r_0 - \delta$ for $r_0 := 0.117841$ and $\delta := 0.00001$, and work with an $(r, r_0, r_0, 2r_0)$-marking. In our connectivity results (Theorems 3 and 5) we choose $r = r_1 - \delta$ for $r_1 := 0.227092$ and work with an $(r, r_1, 0, r_1)$-marking in order to achieve the larger density threshold.

In Section 5 we show that random k-CNF formulae have $(r_0 - \delta, r_0, r_0, 2r_0)$-markings when the density α is below the threshold $2^{(r_0-\delta)k}/k^3$, and that the marginals of good variables are close to $1/2$; this is where the value of r_0 becomes important in the argument. We also show that random k-CNF formulae have $(r_1 - \delta, r_1, 0, r_1)$-markings when the density α is below the threshold $2^{(r_1-\delta)k}/k^3$. We state this result for r_0 in Proposition 10 below; first we give some relevant definitions.

Definition 9 (Ω^*, μ_A, Ω, Φ^Λ, $\mathcal{C}^\Lambda$, $\mathcal{V}^\Lambda$, Ω^Λ). Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. Let Ω^* be the set of all assignments $\mathcal{V} \to \{\mathsf{F}, \mathsf{T}\}$. Given any subset $A \subseteq \Omega^*$, let μ_A be the uniform distribution on A. Let Ω be the set of satisfying assignments of Φ. For any partial assignment Λ we denote by Φ^Λ the formula obtained by simplifying Φ under Λ, i.e., removing the clauses which are already satisfied by Λ, and removing false literals from the remaining clauses. We denote by $\mathcal{C}^\Lambda$ and $\mathcal{V}^\Lambda$ the sets of clauses and variables of Φ^Λ. Moreover, we denote by Ω^Λ the set of satisfying assignments of Φ^Λ.

Proposition 10. There is an integer k_0 such that for any $k \geq k_0$ and any density α with $\alpha \leq 2^{(r_0-\delta)k}/k^3$ the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. There exists an $(r_0 - \delta, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{a}, \mathcal{V}_\mathrm{c})$ of Φ. Moreover, for any such marking, for any $v \in \mathcal{V}_\mathrm{good}(r_0 - \delta)$, any $V \subseteq \mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}$ with $v \notin V$, and any $\Lambda \colon V \to \{\mathsf{F}, \mathsf{T}\}$, we have

$$\max\left\{\Pr_{\mu_{\Omega^\Lambda}}(v \mapsto \mathsf{F}), \Pr_{\mu_{\Omega^\Lambda}}(v \mapsto \mathsf{T})\right\} \leq \frac{1}{2}\exp\left(\frac{1}{k2^{r_0 k}}\right).$$

Proof. This follows directly by combining Lemmas 26 and 28, which are stated and proved in Section 5. □

We note that the density threshold of Theorem 1 is $2^{0.039k}$, which is significantly smaller than the threshold $2^{(r_0-\delta)k}/k^3$ in Proposition 10. The bottleneck for the threshold Theorem 1 comes from our mixing time results, see Section 2.2.

The bound given in Proposition 10 on the marginal probabilities of the marked and auxiliary variables is exploited several times in this work, and we will explain some of these applications in this proof outline. We remark that the bound on the marginals of good variables holds for any pinning of any subset of marked and auxiliary variables, which will be relevant in the spectral independence argument.

Definition 11 ($\mu|_V$). Let $\mathcal{V}$ be a finite set and let $\Omega \subseteq \{\mathsf{F}, \mathsf{T}\}^\mathcal{V}$. Let μ be a distribution over Ω. For a set $V \subseteq \mathcal{V}$, we denote by $\mu|_V$ the marginal distribution of μ on V.

Proposition 10 implies that the distribution $\mu_\Omega|_{\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}}$ is very close to the uniform distribution over all assignments $\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a} \to \{\mathsf{F}, \mathsf{T}\}$. This concept is formalised in the following definition.

Definition 12 (ε-uniform). Let V be a set of variables and μ be a probability distribution over the assignments $V \to \{\mathsf{F}, \mathsf{T}\}$. Let $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ be an assignment of some subset of variables $S \subseteq V$. We denote by $\Pr_\mu(\Lambda)$ the probability under μ of the event that the variables in S are assigned values according to Λ, and by $\Pr_\mu(\cdot|\Lambda)$ the corresponding conditional distribution of μ.

For $\varepsilon \in (0, 1)$, we say that the distribution μ is ε-uniform if for any variable $v \in V$ and any partial assignment $\Lambda \colon V \setminus \{v\} \to \{\mathsf{F}, \mathsf{T}\}$, we have

$$\max\left\{\Pr_\mu(v \mapsto \mathsf{F} | \Lambda), \Pr_\mu(v \mapsto \mathsf{T} | \Lambda)\right\} \leq \frac{1}{2}e^\varepsilon.$$

From Proposition 10, it follows that the distribution $\mu_\Omega|_{\mathcal{V}_\mathrm{m}}$ is ε-uniform for $\varepsilon = (2^{-r_0 k}/k)$, so for any $\Lambda \colon \mathcal{V}_\mathrm{m} \to \{\mathsf{F}, \mathsf{T}\}$, the probability that the assignment of the marked variables is Λ is at least $(1 - e^\varepsilon/2)^{|\mathcal{V}_\mathrm{m}|}$. The ε-uniform property also (trivially) guarantees that the space of assignments $\Lambda \colon \mathcal{V}_\mathrm{m} \to \{\mathsf{F}, \mathsf{T}\}$ with $\Pr_{\mu_\Omega}(\Lambda) > 0$ is connected via single-variable updates, so we can indeed consider the Glauber dynamics over $\mathcal{V}_\mathrm{m}$. This leads to the main challenge of this work: does this chain mix rapidly?

2.2 Mixing time of the Glauber dynamics on the marked variables

Recently, there has been significant progress in showing that the single-variable Glauber dynamics on appropriately chosen subsets of variables mixes quickly for k-CNF formulae with bounded degree [21, 31]. These approaches carefully execute a union bound over paths of clauses connecting marked variables in order to bound the coupling time between two copies of the chain. However, these union bound arguments break under the presence of high-degree variables that are present in random k-SAT; this is because the number of paths connecting marked variables is very sensitive to the max degree of the formula and in particular grows too fast in our setting. We give a more detailed discussion in Section 8.1.

Instead, we apply the spectral independence framework to show rapid mixing of a uniform-block Glauber dynamics, which we review briefly below. Applications of spectral independence usually exploit decay of correlations to show that the spectral independence condition holds, see [7, 10, 8] for examples. As we have mentioned in the introduction, correlation decay fails to hold for densities exponential in k in the random k-SAT model [40] and therefore, we have to develop a different approach to conclude that the spectral-independence condition holds in our setting. This is our main contribution in this work; we show that the marginal distribution on the marked variables, i.e., $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$, is $(\varepsilon \log n)$-spectrally independent for some $\varepsilon > 0$ that can be made arbitrarily small for sufficiently large k. Our argument builds on the coupling idea of Moitra [39] (as refined in [24] for random k-SAT) and relates the spectral independence condition to the expected number of failed clauses in this coupling process. This allows us to exploit the local sparsity properties of the random k-SAT model to analyse the mixing time of the Glauber dynamics.

A caveat here is that the spectral independence of $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is not enough on its own to conclude fast mixing of the single-site Glauber dynamics. The most direct way to work around this is to analyse instead the so-called ρ-uniform-block Glauber dynamics that updates ρ vertices at a time for some ρ that scales linearly in n; the main missing ingredient there is to show that the modified chain can be implemented efficiently which we discuss in Section 2.3. We next give a quick overview of the relevant ingredients of the spectral-independence literature that we will need.

2.2.1 The ρ-uniform-block Glauber dynamics, spectral independence, and the mixing time

Let V be a finite set of size M and μ be a distribution over the assignments $V \to \{\mathsf{F}, \mathsf{T}\}$. Let Ω be the set of assignments $V \to \{\mathsf{F}, \mathsf{T}\}$ with positive probability under μ. For an integer $\rho \in \{1, 2, \ldots, |V|\}$, the ρ-uniform-block Glauber dynamics for μ is a Markov chain X_t where $X_0 \in \Omega$ is an arbitrary configuration and, for $t \geq 1$, X_t is obtained from X_{t-1} by first picking a subset $S \subseteq V$ of size ρ uniformly at random, letting Λ_t be the restriction of X_t to $V \setminus S$, and updating the configuration on S according to the probability distribution $\mu(\cdot|\Lambda_t)$. This chain satisfies the detailed balance equation for μ. Hence, when the chain is irreducible, for $\varepsilon > 0$, we can consider its mixing time $T_{\mathrm{mix}}(\rho, \varepsilon) = \max_{\sigma \in \Omega} \min\{t : d_{\mathrm{TV}}(X_t, \mu) \leq \varepsilon \mid X_0 = \sigma\}$. We say that μ is b-marginally bounded if for all $v \in V$, $S \subseteq V \setminus \{v\}$, $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $\Pr_\mu(\Lambda) > 0$, and $\omega \in \{\mathsf{F}, \mathsf{T}\}$, it either holds that $\Pr_\mu(v \mapsto \omega|\Lambda) = 0$ or $\Pr_\mu(v \mapsto \omega|\Lambda) \geq b$. Spectral independence results have recently been used in the b-marginally bounded setting to obtain fast mixing time of the uniform-block Glauber dynamics [9, 11]. For $S \subset V$, $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $\Pr_\mu(\Lambda) > 0$, and $u, v \in V$ with $u \notin S$ and $0 < \Pr_\mu(u \mapsto \mathsf{T}|\Lambda) < 1$, the influence of u on v (under μ and Λ) is defined as

$$\mathcal{I}^\Lambda(u \to v) = \Pr_\mu\left(v \mapsto \mathsf{T} | u \mapsto \mathsf{T}, \Lambda\right) - \Pr_\mu\left(v \mapsto \mathsf{T} | u \mapsto \mathsf{F}, \Lambda\right). \tag{1}$$

The influence matrix conditioned on Λ is the (two-dimensional) matrix whose entries consist of $\mathcal{I}^\Lambda(u \to v)$ over all relevant u and v. We denote by $\mathcal{I}^\Lambda$ the matrix and by $\lambda_1(\mathcal{I}^\Lambda)$ its largest

eigenvalue in absolute value. For a real $\eta > 0$, we say that μ is η-spectrally independent if for all $S \subset V$ and $\Lambda\colon S \to \{\mathsf{F},\mathsf{T}\}$ with $\Pr_\mu(\Lambda) > 0$ we have $\lambda_1(\mathcal{I}^\Lambda) \leq \eta$. From the results of [11], one can conclude the following bound for the mixing time of the uniform-block Glauber dynamics, see Appendix B for details.

Lemma 13. The following holds for any reals $b, \eta > 0$, any $\kappa \in (0,1)$ and any integer M with $M \geq \frac{2}{\kappa}(4\eta/b^2 + 1)$. Let V be a set of size M, let μ be a distribution over the assignments $V \to \{\mathsf{F},\mathsf{T}\}$, let $\Omega = \{\Lambda\colon V \to \{\mathsf{F},\mathsf{T}\} : \mu(\Lambda) > 0\}$ and let $\mu_{\min} = \min_{\Lambda\in\Omega} \mu(\Lambda)$. If μ is b-marginally bounded and η-spectrally independent, then, for $\rho = \lceil \kappa M \rceil$ and $C_\rho = (2/\kappa)^{4\eta/b^2+1}$, we have

$$T_{\mathrm{mix}}(\rho,\varepsilon) \leq \left\lceil C_\rho \frac{M}{\rho}\left(\log\log\frac{1}{\mu_{\min}} + \log\frac{1}{2\varepsilon^2}\right)\right\rceil.$$

We are going to consider the uniform-block Glauber dynamics on the marked variables of Φ, so $V = \mathcal{V}_{\mathrm{m}}$, and the set of states coincides with the set of assignments $\mathcal{V}_{\mathrm{m}} \to \{\mathsf{F},\mathsf{T}\}$ as all of them have positive probability. In this setting, the target distribution is $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$. The distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(1/e)$-marginally-bounded as a straightforward consequence of the fact that it is $(1/k)$-uniform, see Remark 54 for details. Hence, in order to conclude rapid mixing it remains to establish spectral independence. For this, we are going to use the well-known fact (see for instance [10]) that, for $S \subset V$ and $\Lambda\colon S \to \{\mathsf{F},\mathsf{T}\}$, we have

$$\lambda_1(\mathcal{I}_\Lambda) \leq \max_{u\in V\setminus S} \sum_{v\in V\setminus S} |\mathcal{I}^\Lambda(u \to v)|. \tag{2}$$

2.2.2 Spectral independence in the random k-SAT model

In this section we state our spectral independence results in the random k-SAT model. The results stated in this section are proved in Section 8. Our main technical result is the following.

Lemma 14. There is an integer $k_0 \geq 3$ such that for any integer $k \geq k_0$ and any density α with $\alpha \leq 2^{r_0 k/3}/k^3$ the following holds. W.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for any $(r_0 - \delta, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ of Φ, the distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(2^{-(r_0-\delta)k}\log n)$-spectrally independent.

We are going to describe some of the ideas behind the proof of Lemma 14. First, we highlight the fact that, due to the presence of high-degree variables (which form logarithmically-sized connected components), current techniques seem unable to conclude η-spectral independence with $\eta = O(1)$. This has also been the case in recent work on 2-spin systems on random graphs [8], where instead correlation decay is exploited to prove η-spectral independence for some $\eta = o(\log n)$. Here, our η-spectral independence bound for $\eta = o_k(\log n)$ will be based on an appropriate coupling. Note, in light of Lemma 13, $\eta = O(\log n)$ is good enough for proving polynomial mixing time of the uniform-block Glauber dynamics, but we need the improved bound of Lemma 14 in order to conclude the following fast mixing-time result from Lemma 13 (as illustrated Section 8).

Lemma 15. There is a function $k_0(\theta) = \Theta(\log(1/\theta))$ such that, for any $\theta \in (0,1)$, for any integer $k \geq k_0(\theta)$ and any density α with $\alpha \leq 2^{0.039k}$ the following holds. W.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for any $(r_0 - \delta, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ of Φ and for $\rho = \lceil 2^{-k-1}|\mathcal{V}_{\mathrm{m}}| \rceil$, the ρ-uniform-block Glauber dynamics for updating the marked variables has mixing time $T_{\mathrm{mix}}(\rho, \varepsilon/2) \leq T := \lceil 2^{2k+3} n^\theta \log \frac{2n}{\varepsilon^2} \rceil$.

Lemma 15 is stated for the block size $\rho = \lceil 2^{-k-1}|\mathcal{V}_{\mathrm{m}}|\rceil$, but it could be proved more generally when $\rho = c|\mathcal{V}_{\mathrm{m}}|$ and $c \in (0,1)$. The fact that $\rho \leq |\mathcal{V}_{\mathrm{m}}|/2^k$ in the statement will be relevant in implementing efficiently the dynamics, discussed in Section 2.3.

We remark that the more restrictive density threshold $\alpha \leq 2^{r_0 k/3}/k^3$ in the statement of Lemma 14 arises in the union bound given in the proof of this lemma, and that for large enough k we have $2^{0.039k} \leq 2^{r_0 k/3}/k^3$, the former being the density threshold given in Lemma 15 and Theorem 1.

Our approach to prove η-spectral independence significantly differs from those that in two-spin systems, where it is enough to study sum of influences over trees (thanks to the tree of self-avoiding walks) and exploit decay of correlations in this setting (very roughly, the further away two vertices are in the tree, the smaller the influence that one vertex has in the other). Here we relate influences to the structure of the dependency graph G_Φ by running a coupling process on the auxiliary variables, and we state this connection in the upcoming Lemma 45. First we define more formally the dependency graph G_Φ.

Definition 16 (G_Φ). Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. We define the graph G_Φ as follows. The vertex set of G_Φ is $\mathcal{C}$ and two clauses c_1 and c_2 are adjacent if and only if $\mathrm{var}(c_1) \cap \mathrm{var}(c_2) \neq \emptyset$. A set $C \subseteq \mathcal{C}$ is connected if C is connected in the graph G_Φ. We say that two variables u and v are connected in Φ if there is a path $c_1, c_2, \ldots, c_\ell$ in G_Φ with $u \in \mathrm{var}(c_1)$ and $v \in \mathrm{var}(c_\ell)$.

Let $u \in \mathcal{V}_{\mathrm{m}}$, $S \subset \mathcal{V}_{\mathrm{m}}$ and $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$. The aim of the coupling process is bounding the sum $\sum_{v \in \mathcal{V}_{\mathrm{m}} \setminus (S \cup \{u\})} |\mathcal{I}^\Lambda(u \to v)|$ in terms of the expected size of a connected set of failed clauses, where the expectation is over the choices made in the coupling process. We refer to Section 8 for a definition of failed clauses, as it is not relevant in this discussion. Here we give a brief overview of how the coupling process on the auxiliary variables works. First, we start with two assignments $X = \Lambda \cup (u \mapsto \mathsf{T})$ and $Y = \Lambda \cup (u \mapsto \mathsf{F})$, where $\Lambda \cup (u \mapsto \omega)$ denotes the assignment defined on $S \cup \{u\}$ that agrees with Λ on S and sends u to ω. The process progressively extends X and Y on some auxiliary variables $v_1, v_2, \ldots$ following the optimal coupling between the marginals $\Pr_{\mu_\Omega}(v \mapsto \cdot | X)$ and $\Pr_{\mu_\Omega}(v \mapsto \cdot | Y)$, see Section 8 for the definition of optimal coupling. The main property of this process is that with high probability over the choices made, at some point the graphs G_{Φ^X} and G_{Φ^Y} factorise in small connected components in spite of the presence of bad variables and, on top of that, Φ^X and Φ^Y share most of these connected components. Then we can bound influences between marked variables by analysing the connected components where Φ^X and Φ^Y differ, which turn out to be $\mathrm{poly}(k) \log n$ in size after enough steps of the process.

One of the key ideas behind our analysis is exploiting the fact that, in the random k-SAT model, w.h.p. over the choice of the random formula Φ, any logarithmic-sized set of clauses Z that is connected in G_Φ has constant tree-excess, that is, the number of edges connecting a pair of clauses in Z is $|Z| + O(1)$. This saves a factor of $\Delta_{r_0 - \delta}$ in the spectral independence bound by ensuring that there is a large independent set of clauses in the set of failed clauses. We also obtain improved analysis by restricting the coupling process to auxiliary variables. This enables us to get exponentially small bounds (in k) on the influences between marked variables, which leads to our $(2^{-(r_0-\delta)k} \log n)$-spectral independence result.

2.3 Analysis of the connected components of Φ^Λ. Applications to connectivity and looseness

In this section we deal with the third challenge mentioned at the beginning of Section 2: can we determine the transition probabilities of the Glauber dynamics so that we can actually simulate this Markov chain? In fact, simulating the single-site Glauber dynamics on the marked variables was one of the main challenges even in the bounded-degree case. In that case this was resolved using a

method that is restricted to the bounded-degree setting (and whose bottleneck is the analysis of a rejection sampling procedure). A different procedure is required for the random k-SAT setting.

One of the key ideas to simulate this chain is starting the chain on an assignment $X_0 \colon \mathcal{V}_{\mathrm{m}} \to \{\mathsf{F}, \mathsf{T}\}$ drawn from the uniform distribution over all assignments of $\mathcal{V}_{\mathrm{m}}$. Since the distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(1/k)$-uniform (Proposition 10), the transition probabilities of the Glauber dynamics are close to uniform. This allows us to show that the probability distribution of the assignment X_t that is output by the uniform-block Glauber dynamics after t steps is also $(1/k)$-uniform (Corollary 29), which will be important in what follows.

In order to run the ρ-uniform-block Glauber dynamics we need to be able to sample from the distribution μ_{Ω^Λ} for any set $S \subseteq \mathcal{V}_{\mathrm{m}}$ with $|S| = \rho$ and any assignment $\Lambda \colon \mathcal{V}_{\mathrm{m}} \setminus S \to \{\mathsf{F}, \mathsf{T}\}$ that arises. Unless we can restrict Λ, sampling from μ_{Ω^Λ} could potentially be as hard as sampling from μ_Ω. Fortunately for us, the assignment Λ is not completely arbitrary; Λ is determined by the random choice of S and the current state of the Glauber dynamics (which follows a $(1/k)$-uniform distribution as discussed above). We show that we can efficiently sample from μ_{Ω^Λ} w.h.p. over the choice of Λ. An important observation is that we can efficiently sample from μ_{Ω^Λ} when the connected components of G_{Φ^Λ} are logarithmic in size, for example, by applying brute force. This raises the following question: does G_{Φ^Λ} break into small connected components w.h.p. over the choice of Λ? Lemma 17 gives a positive answer when $0 \leq \rho \leq |V|/2^k$. Here the reader can see V as the set of marked variables. The proof of Lemma 17 exploits sparsity properties of logarithmic-sized connected sets of clauses in random formulae in conjunction with the fact that μ is $(1/k)$-uniform. Lemma 17 is stated with an added layer of generality, as we will also apply it to analyse the geometry of the space of satisfying assignments of Φ with $r = r_1 - \delta$. In our sampling algorithm setting we consider $r = r_0 - \delta$. Recall that $r_0 = 0.117841$, $r_1 = 0.227092$ and $\delta = 0.00001$. The restriction $r \in (2\delta, 1/(2\log 2)]$ in the statement of Lemma 17 is not optimal, but it is enough for our purposes.

Lemma 17. Let $r \in (2\delta, 1/(2\log 2)]$. There is an integer $k_0 \geq 3$ such that, for any integer $k \geq k_0$, any density $\alpha \leq 2^{(r-2\delta)k}$, and any real number b with $a := 2k^4 < b$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$.

Let L be an integer satisfying $a \log n \leq L \leq b \log n$. Let V be a set of good variables of Φ that is $(r+\delta)$-distributed (Definition 8), let μ be a $(1/k)$-uniform distribution over the assignments $V \to \{\mathsf{F}, \mathsf{T}\}$, and let ρ be an integer with $0 \leq \rho \leq |V|/2^k$. Consider the following experiment. First, draw $S \subseteq V$ from the uniform distribution τ over subsets of V with size ρ. Then, sample an assignment Λ from $\mu|_{V \setminus S}$. Denote by $\mathcal{F}$ the event that that there is a connected set of clauses Y of Φ with $|Y| \geq L$ such that all clauses in Y are unsatisfied by Λ. Then $\Pr_{S \sim \tau}\left(\Pr_{\Lambda \sim \mu|_{V \setminus S}}(\mathcal{F}) \leq 2^{-\delta k L}\right) \geq 1 - 2^{-\delta k L}$.

Proof sketch. The proof is in Section 6. For the sake of exposition, we first sketch the proof in the case $\rho = 0$, where the conclusion in the statement reads $\Pr_{\Lambda \sim \mu|_V}(\mathcal{F}) \leq 2^{-\delta k L}$. At the end of this proof sketch we explain how we extend the proof to any ρ with $0 \leq \rho \leq |V|/2^k$.

The first step is exploiting local sparsity properties of random k-CNF formulae to find many variables from V in any sufficiently large connected set of clauses. Our sparsity results hold for connected sets of clauses with size at least $2k^4 \log n$, and let us conclude the following result (stated as Lemma 33 in Section 6): w.h.p. over the choice of Φ, for every connected set of clauses $Z \subseteq \mathcal{C}$ we have

$$\text{if } \ 2k^4 \log(n) \leq |Z| \leq b \log(n), \quad \text{then } \ |\mathrm{var}(Z) \cap V| \geq rk|Z|. \tag{3}$$

The proof of Lemma 33 counts the variables from V in Z by using the fact that Z does not contain many bad clauses (Lemma 20, which gives the restriction on r) and the fact that there

are not many edges joining clauses in Z. In fact, for such a set Z, we show that the number of edges is of order $|Z| + O(1)$, that is, Z has constant tree-excess (Lemma 31). We also need the following result on random k-CNF formulae. For each clause $c \in \mathcal{C}$, let $\mathcal{Z}(c, L) = \{Z \subseteq \mathcal{C} : c \in Z, Z \text{ is connected in } G_\Phi, |Z| = L\}$. Then, w.h.p. over the choice of Φ, [24, Lemma 40] shows that, as long as $L \geq \log n$,

$$\text{for any clause } c \in \mathcal{C} \text{ we have } |\mathcal{Z}(c, L)| \leq (9k^2\alpha)^L. \tag{4}$$

Once we have established (3) and (4), the proof exploits the fact that μ is close to the uniform distribution. First, we introduce some notation. Let L be an integer with $a \log n \leq L \leq b \log n$. Let $S = \emptyset$ as we are dealing with the case $\rho = 0$. For $c \in \mathcal{C}$ and $Z \in \mathcal{Z}(c, L)$, we denote by $\mathcal{E}_1(Z, S)$ the event that none of the clauses of Z are satisfied by assignment Λ (Definition 9), where Λ is drawn from $\mu|_{V \setminus S}$, see Definition 11. We keep track of S in the notation here as this is relevant in the general case. The first observation is that the event $\mathcal{F}$ from the statement satisfies $\mathcal{F} = \bigcup_{c \in \mathcal{C}, Z \in \mathcal{Z}(c,L)} \mathcal{E}_1(Z, S)$. We then claim that for any $c \in \mathcal{C}$ and $Z \in \mathcal{Z}(c, L)$ we have

$$\Pr_{\Lambda \sim \mu|_{V \setminus S}} (\mathcal{E}_1(Z, S)) \leq \frac{2^{-\delta k L}}{|\mathcal{C}| \cdot |\mathcal{Z}(c, L)|}, \tag{5}$$

so the result would follow from a union bound over c and Z. Let us give some insight on how we prove (5). Let $c \in \mathcal{C}$ and $Z \in \mathcal{Z}(c, L)$. The main idea is that, if all clauses in Z are unsatisfied by Λ then, when we sampled $\Lambda \sim \mu|_{V \setminus S}$, for each variable v in $\text{var}(Z) \cap (V \setminus S)$ we picked the value that does not satisfy the clauses of Z containing v. Thus, we can bound the probability that all clauses in Z are unsatisfied as a product, over the variables in $\text{var}(Z) \cap (V \setminus S)$, of probabilities, each factor corresponding to the probability that a variable is assigned a certain value (under some careful conditioning, see the proof in Section 6 for details). Since the distribution μ is $(1/k)$-uniform, each one of these factors can be bounded by $\exp(1/k)/2$, obtaining

$$\Pr_{\Lambda \sim \mu|_{V \setminus S}} (\mathcal{E}_1(Z, S)) \leq \left(\frac{1}{2} \exp\left(\frac{1}{k}\right) \right)^{|\text{var}(Z) \cap (V \setminus S)|}. \tag{6}$$

In (3) we gave a lower bound on $|\text{var}(Z) \cap V|$, which can be applied in conjunction with (4) to conclude, after some calculations, that the bound given in (5) holds.

The case $\rho > 0$ is more technical and one has to be more careful in these calculations. We show that (5) holds when S does not contain many variables in $\text{var}(Z) \cap V$. A slightly different argument is needed when going from (6) to (5); here we have to bound $|\text{var}(Z) \cap (V \setminus S)|$ instead of $|\text{var}(Z) \cap V|$. It turns out that, as long as the bound $|\text{var}(Z) \cap V \cap S| \leq |\text{var}(Z) \cap V|/k$ holds, the calculations to go from (6) to (5) also hold in this setting. Finally, we show that the probability that $|\text{var}(Z) \cap V \cap S| \leq |\text{var}(Z) \cap V|/k$ occurs when picking S is at least $1 - 2^{\delta k L}$. The proof of this fact is purely combinatorial, and requires the hypothesis $\rho \leq |V|/2^k$, see Section 6 for details. □

Once we have established Lemma 17, we can use it to implement the ρ-uniform-block Glauber dynamics on the marked variables for $0 < \rho \leq |\mathcal{V}_\text{m}|$ and complete our sampling algorithm, which we explicitly state in Section 2.4.

Before concluding this section, we mention how we apply Lemma 17 to analyse the geometry of the space of satisfying assignments of Φ in order to conclude the $O(\log n)$-connectivity and $O(\log n)$-looseness results given in Theorems 3 and 5. First, we need the following definition.

Definition 18 (H_Φ). Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. We define the graph H_Φ as follows. The vertex set of H_Φ is $\mathcal{V}$ and two variables v_1 and v_2 are adjacent in H_Φ if there is a clause $c \in \mathcal{C}$ with $v_1, v_2 \in \text{var}(c)$.

We apply Lemma 17 with $r = r_1 - \delta$ and a density $\alpha \leq 2^{(r_1-3\delta)k}/k^3$. For an $(r, r_1, 0, r_1)$-marking $(\mathcal{V}_\mathrm{m}, \emptyset, \mathcal{V}_\mathrm{c})$ of Φ, we let $V = \mathcal{V}_\mathrm{m}$ and $\mu = \mu_\Omega|_{\mathcal{V}_\mathrm{m}}$. In this setting, for $\rho = 0$, Lemma 17 allows us to conclude that, w.h.p. over the choice of $\Lambda \sim \mu_\Omega|_{\mathcal{V}_\mathrm{m}}$, the graph G_{Φ^Λ} consists of connected components with size at most $O(\log n)$. Thus, the connected components of H_{Φ^Λ} have size at most $O(\log n)$ as each clause contains at most k variables. This leads to the main idea behind the proof of Theorem 3: we can construct $O(\log n)$-paths between satisfying assignments by progressively updating the variables in each one of the connected components of H_{Φ^Λ}. As an example, let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$ be these connected components and let σ_1 and σ_2 be two satisfying assignments that agree with Λ on $\mathcal{V}_\mathrm{m}$. Then we can find an $O(\log n)$-path $\sigma_1 = \zeta_0 \leftrightarrow \zeta_1 \leftrightarrow \cdots \leftrightarrow \zeta_t = \sigma_2$ as follows: the assignment ζ_j is the satisfying assignment that agrees with Λ, agrees with σ_1 on the variables in $\mathcal{V} \setminus \left(\bigcup_{i=1}^{j} \mathcal{E}_j\right)$ and agrees with σ_2 on the variables in $\bigcup_{i=1}^{j} \mathcal{E}_j$. The case when σ_1 and σ_2 differ on some marked variables builds on the same idea though it is more technical and requires applying Lemma 17 with $\rho = 1$. We refer to Section 10.1 for this argument and the proof of Theorem 3.

The fact that the connected components of H_{Φ^Λ} are $O(\log n)$ in size with high probability over $\Lambda \sim \mu_\Omega|_{\mathcal{V}_\mathrm{m}}$ is also related to the looseness of the formula Φ. Let $v \in \mathcal{V} \setminus \mathcal{V}_\mathrm{m}$. For any satisfying assignment σ that agrees with Λ on the marked variables, we can construct a satisfying assignment τ with $\tau(v) \neq \sigma(v)$ and $\|\sigma - \tau\|_1 = O(\log n)$ by updating the variables in the connected component of v in H_{Φ^Λ}, provided that there is a way to satisfy this connected component when giving v the value $\tau(v)$. In Section 10.2 we formalise this idea and give all the details of this argument to prove Theorem 5.

2.4 The sampling algorithm

To complete this proof outline, we explicitly describe Algorithm 1, our algorithm for sampling satisfying assignments of k-CNF formulae. The algorithm uses a method Sample(Φ^Λ, S) to sample an assignment $\tau \colon S \to \{\mathsf{F}, \mathsf{T}\}$ from the distribution $\mu_{\Omega^\Lambda}|_S$. This method exploits the fact that logarithmic-sized connected set of clauses have constant tree-excess, which does not hold in the bounded-degree case. This tree-like property enables us to efficiently sample satisfying assignments on the connected components of Φ^Λ by a standard dynamic programming argument, see Section 7. Lemma 19 is our main result on Sample(Φ^Λ, S).

Lemma 19. There is an integer $k_0 \geq 3$ such that, for any integers $k \geq k_0$, $b \geq 2k^4$ and any density $\alpha > 0$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let V be a subset of variables and let $\Lambda \colon V \to \{\mathsf{F}, \mathsf{T}\}$ be a partial assignment such that all the connected components in G_{Φ^Λ} have size at most $b \log(n)$. Then, there is an algorithm that, for any $S \subseteq \mathcal{V} \setminus V$, samples an assignment from $\mu_{\Omega^\Lambda}|_S$ in time $O(|S| \log n)$.

The method Sample(Φ^Λ, S) is used in Algorithm 1 to implement each step of the ρ-uniform-block Glauber dynamics on the marked variables. It is also used to extend the assignment of marked variables computed by the Glauber dynamics to a satisfying assignment of Φ. As a design choice, this method returns error when the connected components of G_{Φ^Λ} have size larger than $2k^4(1 + \xi) \log(n)$. We remark that the probability that Sample(Φ^Λ, S) returns error is very small when running the Glauber dynamics thanks to Lemma 17. We can now introduce Algorithm 1, which has two parameters $\theta \in (0, 1)$ and $\xi \geq 1$ as in Theorem 1.

Algorithm 1 The approximate sampling algorithm for satisfying assignments of random k-CNF formulae.

Input: A k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$ with n variables

Compute the sets of bad/good variables and bad/good clauses for Φ as in Proposition 7.
Let $\varepsilon = n^{-\xi}$. Compute a marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ for Φ as in Lemma 26 with $p = \varepsilon/4$. This succeeds with probability at least $1 - \varepsilon/4$. If this does not succeed, the algorithm returns error.
For each $v \in \mathcal{V}_{\mathrm{m}}$, sample $X_0(v) \in \{\mathsf{F}, \mathsf{T}\}$ uniformly at random.
for t from 1 to $T := \lceil 2^{2k+3} n^{\theta} \log \frac{2n}{\varepsilon^2} \rceil$ do
    Choose uniformly at random a set of marked variables $S \subseteq \mathcal{V}_{\mathrm{m}}$ with size $\rho := \lceil 2^{-k-1} |\mathcal{V}_{\mathrm{m}}| \rceil$.
    Let Λ_t be the assignment X_{t-1} restricted to $\mathcal{V}_{\mathrm{m}} \setminus S$.
    $Y \leftarrow \mathrm{Sample}(\Phi^{\Lambda_t}, S)$.
    $X_t \leftarrow \Lambda_t \cup Y$.
end for
$Y \leftarrow \mathrm{Sample}(\Phi^{X_T}, \mathcal{V}_{\mathrm{a}} \cup \mathcal{V}_{\mathrm{c}})$.
return $X_T \cup Y$.

We remark here that Algorithm 1 only works for large enough k, and this hypothesis will be used several times in our arguments. The quantity T defined in this algorithm corresponds to the mixing time of the ρ-uniform-block Glauber dynamics given in Lemma 15.

3 Paper outline

The rest of this work is organised as follows. In Section 4 we introduce the procedure for determining bad clauses. In Section 5 we prove Proposition 10 on markings of random formulae. In Section 6 we prove our technical result on the connected components of Φ^{Λ}, Lemma 17. In Section 7 we give the method Sample and prove Lemma 19. In Section 8 we prove the results on spectral independence stated in Section 2.2 of the proof outline. In Section 9 we complete the proof of Theorem 1 by combining our mixing time results (Lemma 15), our algorithm to sample from small connected components (Lemma 19) and our result on the size of the connected components of Φ^{Λ} (Lemma 17). Finally, in Section 10 we prove Theorems 3 and 5 on the geometry of the space of satisfying assignments of Φ.

To help keep track of the notation and definitions introduced in this work, the reader is referred to the tables in Appendix C.

4 High-degree and bad variables in random CNF formulae

As we noted in the introduction, one of the keys to sampling satisfying assignments in the unbounded-degree setting is to "sacrifice" a few variables per clause (treating them separately in the sampling algorithm) and to (temporarily) remove a small linear number of clauses that contain these. The point of this is to ensure that the remaining ("good") clauses have mostly low-degree variables (at most two bad ones) and also that the rest of the clauses (the "bad" ones) form small connected components that interact with the good clauses in a manageable way.

Recall that, for $r \in (0, 1)$, high-degree variables were introduced in Definition 6 as those variables with at least $\Delta_r := \lceil 2^{kr} \rceil$ occurrences in the formula. In this work we consider two possible values for r here, $r = r_0 - \delta$ and $r = r_1 - \delta$, where $r_0 = 0.117841$, $r_1 = 0.227092$ and $\delta = 0.00001$. The values r_0 and r_1 arise as solutions of an optimisation problem in Section 5 when we establish the markings that we use in our proofs. The marking used in our algorithmic results requires

the more restrictive definition of high-degree variable with $r = r_0 - \delta$ than the marking used in our connectivity results with $r = r_1 - \delta$. Subtracting δ will make our calculations easier without affecting our results.

By standard arguments about random graphs, one can determine that, w.h.p. over the choice of Φ, the number of high-degree variables of Φ is bounded. We want to identify the clauses of Φ that have at most 2 high-degree variables, since clauses with a lot of high-degree variables will interfere with our sampling algorithms. This motivates the following construction. The bad variables and bad clauses of Φ are identified by running the process given in Algorithm 2. Here $\mathcal{V}_{\text{bad}}(r)$ denotes the set of bad variables and $\mathcal{C}_{\text{bad}}(r)$ denotes the set of bad clauses.

Algorithm 2 Computing bad variables and bad clauses for $r \in (0,1)$

Input: A k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$

$\mathcal{V}_0(r) \leftarrow$ the set of high-degree variables, i.e., variables with at least $\Delta_r = \lceil 2^{rk} \rceil$ occurrences in Φ.
$\mathcal{C}_0(r) \leftarrow$ the set of clauses with at least 3 variables in $\mathcal{V}_0(r)$
$i \leftarrow 0$
while $i = 0$ or $\mathcal{V}_i(r) \neq \mathcal{V}_{i-1}(r)$ do
    $i \leftarrow i + 1$
    $\mathcal{V}_i(r) \leftarrow \mathcal{V}_{i-1}(r) \cup \text{var}(\mathcal{C}_{i-1}(r))$
    $\mathcal{C}_i(r) \leftarrow \{c \in \mathcal{C} : |\text{var}(c) \cap \mathcal{V}_i(r)| \geq 3\}$
end while
$\mathcal{C}_{\text{bad}}(r) \leftarrow \mathcal{C}_i(r)$ and $\mathcal{V}_{\text{bad}} \leftarrow \mathcal{V}_i(r)$
return $\mathcal{V}_{\text{bad}}(r), \mathcal{C}_{\text{bad}}(r)$

We define the good clauses of Φ as $\mathcal{C}_{\text{good}}(r) = \mathcal{C} \setminus \mathcal{C}_{\text{bad}}(r)$ and the good variables of Φ as $\mathcal{V}_{\text{good}}(r) = \mathcal{C} \setminus \mathcal{V}_{\text{bad}}(r)$. The sets $\mathcal{V}_{\text{good}}(r), \mathcal{V}_{\text{bad}}(r), \mathcal{C}_{\text{good}}(r), \mathcal{C}_{\text{bad}}(r)$ depend on the parameter $r \in (0,1)$. The value of r here will be $r_0 - \delta$ except in Section 10 where we prove our connectivity results for $r = r_1 - \delta$, and in some of the marking results in Section 5. We will use the observations given in Proposition 7 several times in this work.

Proposition 7. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. For any $c \in \mathcal{C}_{\text{good}}(r)$, we have $|\text{var}(c) \cap \mathcal{V}_{\text{bad}}(r)| \leq 2$, and for any $c \in \mathcal{C}_{\text{bad}}(r)$, we have $|\text{var}(c) \cap \mathcal{V}_{\text{good}}(r)| = 0$. Moreover, every good variable has degree less than Δ_r. There is a procedure to determine $\mathcal{C}_{\text{bad}}$ that runs in time $O(n + mk)$, where n is the number of variables of Φ and m is the number of clauses of Φ.

Proof. In this proof we briefly explain the implementation of Algorithm 2. First, for each clause c we keep track of the number of bad variables in $\text{var}(c)$, denoted $\text{bad}(c)$. We also have a stack of bad variables $S_{\mathcal{V}}$ that are yet to be processed by the algorithm. At the start of the algorithm, we set $S_{\mathcal{V}} \leftarrow \mathcal{V}_0$. While $S_{\mathcal{V}}$ is non-empty, we take the variable v on the top of the stack and increase $\text{bad}(c')$ by 1 for those clauses c' where v appears. If any of these updates gives $\text{bad}(c') \geq 3$, we add $\text{var}(c')$ to the stack $S_{\mathcal{V}}$, set the variables in $\text{var}(c')$ as bad and set the clause c' as bad. At the end of this process, $S_{\mathcal{V}}$ is empty and we have found all the bad variables and bad clauses of Φ. As every variable is added to the stack at most once and the list $\text{bad}(\cdot)$ is updated at most mk times (once per literal in Φ), the running time is $O(n + mk)$. □

In our work we need a variation of result of [24] that controls the number of bad clauses in connected subgraphs of G_Φ. We state this result in Lemma 20 and prove it in Appendix A.

Lemma 20 (Modified version of [24, Lemma 8.16]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For every connected set of clauses Y in G_Φ such that $|\mathrm{var}(Y)| \geq 2k^4 \log n$, we have $|Y \cap \mathcal{C}_{\mathrm{bad}}(r)| \leq |Y|/k$.

We also need a bound on the number of bad clauses of Φ, which is also proved in Appendix A.

Lemma 21 (Modified version of [24, Lemma 8.12]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. We have $|\mathcal{C}_{\mathrm{bad}}(r)| \leq 2(\alpha/\Delta_r)n/2^{k^{10}}$ and $|\mathcal{V}_{\mathrm{bad}}(r)| \leq 2(k+1)(\alpha/\Delta_r)n/2^{k^{10}}$.

Lemmas 20 and 21 guarantee that, w.h.p. over the choice of Φ, bad clauses are a minority among all the clauses of Φ. This will be used to show that bad clauses do not affect significantly the behaviour of our sampling algorithm. We point out that the definitions of $\mathcal{V}_{\mathrm{good}}(r), \mathcal{V}_{\mathrm{bad}}(r), \mathcal{C}_{\mathrm{good}}(r)$ and $\mathcal{C}_{\mathrm{bad}}(r)$ given in [24] have $r = 1/300$ and, in Algorithm 2, use the condition $|\mathrm{var}(c) \cap \mathcal{V}_i(r)| \geq k/10$ instead of $|\mathrm{var}(c) \cap \mathcal{V}_i(r)| \geq 3$

Hence, our definitions of good clauses and good variables are more restrictive. However, it turns out that, with minor changes, the proof of Lemma 20 given in [24] can be extended to our setting. These changes are explained in Appendix A.

5 Identifying a set of "marked" variables with good marginals

A property that is useful for sampling satisfying assignments is having a high proportion of variables in each good clause such that the marginals of these variables are fairly close to 1/2. That is, having variables which are roughly equally likely to be true or false in a random satisfying assignment. The marginals of high-degree variables do vary. However, even in the random k-SAT model it turns out that there are enough variables with marginals near 1/2. Following the basic approach of Moitra [39], we partition the good variables of a random k-CNF formula into types. Here we have three types of variables (instead of two): marked, auxiliary and control variables. The high-level goal is to do this in such a way that each clause has a good proportion of each one of these types of variables. We call this construction a marking, see Definition 8 of the proof outline for the precise definition. For such a marking, we will show that as long as the control variables are left unassigned/unpinned, the marginals of the marked and auxiliary variables are all near 1/2 as a consequence the Lovász local lemma [20]. We first set up the notation and results that we need.

It is not difficult to show that in the random k-SAT model, w.h.p. over the choice of the formula Φ, two distinct clauses share at most 2 variables (see Lemma 22). Previous work on counting/sampling satisfying assignments of bounded degree formulae had to analyse subsets of disjoint clauses in order to deal with the fact that small sets of clauses might share most of their variables. The restriction to disjoint subsets imposes further restrictions on the maximum degree of the formula and on the density of the formula in the random k-SAT model setting. Here we manage to exploit Lemma 22 to avoid these restrictions.

Lemma 22. For any $k \geq 3$ and any density $\alpha > 0$ (possibly depending on k), the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. We have $|\mathrm{var}(c)| \geq k - 1$ and $|\mathrm{var}(c) \cap \mathrm{var}(c')| \leq 2$ for all $c, c' \in \mathcal{C}$ with $c \neq c'$.

Proof. First, let us prove that, for $k \geq 3$, w.h.p. over the choice of Φ, $|\mathrm{var}(c)| \geq k - 1$ for all $c \in \mathcal{C}$. Let us denote by $\mathcal{R}_c$ the event that a clause c has at least two repetitions among its variables, that

is, $|\mathrm{var}(c)| \leq k-2$. We claim that $\Pr(\mathcal{R}_c) \leq q(k)/n^2$, where $q = \binom{k}{3} + k(k-1)(k-2)(k-3)/4$. To prove this statement we note that the probability that a variable appears at least 3 times in c is at most $\binom{k}{3}n^{k-2}/n^k$, and the probability that two distinct variables are repeated in c is at most $p(k)n(n-1)n^{k-4}/n^k$ for $p(k) = k(k-1)(k-2)(k-3)/4$. Hence, by adding up both cases, we find that $\Pr(\mathcal{R}_c) \leq q(k)/n^2$, and $\Pr(\bigcup_{c\in\mathcal{C}} \mathcal{R}_c) \leq q(k)m/n^2 \leq q(k)\alpha/n = O(1/n)$, so the result follows.

Let $c, c' \in \mathcal{C}$ with $c \neq c'$. We study $|\mathrm{var}(c) \cap \mathrm{var}(c')|$,

$$\Pr\left(\left|\mathrm{var}(c) \cap \mathrm{var}(c')\right| \geq 3\right) \leq \frac{n(n-1)(n-2)n^{2(k-3)}(k(k-1)(k-2))^2}{n^{2k}} \leq \frac{k^6}{n^3}.$$

Therefore, the probability that there is a pair of clauses c, c' with $|\mathrm{var}(c) \cap \mathrm{var}(c')| \geq 3$ is bounded from above by $\frac{m(m-1)}{2}\frac{k^6}{n^3} \leq \frac{\alpha^2}{2}\frac{k^6}{n} = O\left(\frac{1}{n}\right)$, which finishes the proof. □

We will use the asymmetric version of the Lovász local lemma (LLL), proved by Lovász and originally published in [46]. Before stating this result, let us introduce some notation. Let $\mathcal{P}$ be a finite collection of mutually independent random variables. Let B an event that is a function of the random variables in $\mathcal{P}$. Let $\mathcal{A}$ be a collection of events that are a function of the random variables in $\mathcal{P}$. We define $\Gamma(B)$ as the set of events $A \in \mathcal{A}$ such that $A \neq B$ and A and B are not independent. In this setting, $\Pr_P(B)$ is the probability that the event B holds when sampling all the random variables in $\mathcal{P}$.

Theorem 23 (Asymmetric Lovász local lemma, [25, Theorems 1.1 and 2.1]). Let $\mathcal{P}$ be a finite collection of mutually independent random variables. Let $\mathcal{A}$ be a collection of events that are a function of the random variables in $\mathcal{P}$. If there exists a function $x : \mathcal{A} \to (0,1)$ such that, for all $A \in \mathcal{A}$, we have

$$\Pr_P(A) \leq x(A) \prod_{N\in\Gamma(A)} (1 - x(N)),$$

then $\Pr_P\left(\bigcap_{A\in\mathcal{A}} \overline{A}\right) > 0$. Furthermore, for any event B that is a function of the random variables in $\mathcal{P}$, we have

$$\Pr_P\left(B \,\middle|\, \bigcap_{A\in\mathcal{A}} \overline{A}\right) \leq \Pr_P(B) \prod_{A\in\Gamma(B)} \left(1 - x(A)\right)^{-1}.$$

We are going to apply the LLL in Lemma 26 to find an $(r_0 - \delta, r_0, r_0, 2r_0)$-marking of Φ (Definition 8), w.h.p. over the choice of the random formula, for some appropriate $r_0 \in (0,1)$. Before proving Lemma 26, let us highlight how strong the properties of a marking are. First, the fact that a set of marked variables is ρ-distributed (Definition 8) will allow us to find, w.h.p. over the choice of Φ, a good amount of marked variables in any set of clauses, even if the set includes bad clauses, see Lemma 33 for a precise statement. This result is an essential ingredient in our proofs. Secondly, as long as the control variables are left unassigned, the marginals of the marked and auxiliary variables will be near 1/2 as a consequence of the LLL, as we show later in this section (Lemma 28). We remark that, in the definition of ρ-distributed set of variables, we ask for $|\mathrm{var}(c) \cap V| \geq \rho(k-3)$ instead of $|\mathrm{var}(c) \cap V| \geq \rho k$ to account for the fact that w.h.p. a good clause has at most a repeated variable (Lemma 22) and at most two bad variables (Proposition 7), which will come up in the proofs presented in this section. First, we need the following definition.

Definition 24 ($\Phi_{\mathrm{good}}(r)$, $\Phi_{\mathrm{bad}}(r)$). Let $r \in (0,1)$. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula. Let $\Phi_{\mathrm{good}}(r) = (\mathcal{V}_{\mathrm{good}}(r), \mathcal{C}_{\mathrm{good}}(r))$ be the CNF formula obtained by taking the good clauses of Φ and ignoring the bad variables appearing in them. Let $\Phi_{\mathrm{bad}}(r)$ be the k-CNF formula with variables $\mathcal{V}_{\mathrm{bad}}(r)$ and clauses $\mathcal{C}_{\mathrm{bad}}(r)$.

Note that in $G_{\Phi_{\text{good}}(r)}$ two clauses c_1 and c_2 in $\mathcal{C}_{\text{good}}$ are adjacent if and only if $\text{var}(c_1) \cap \text{var}(c_2) \cap \mathcal{V}_{\text{good}} \neq \emptyset$. By definition of good variables, the maximum degree in $G_{\Phi_{\text{good}}(r)}$ is at most $k(\Delta_r - 1)$, which will be important when applying the LLL. We also need the following version of Chernoff's bounds.

Lemma 25 (Chernoff's bounds - [43, Theorem 2.1 and Corollary 4.1]). Let $n \in \mathbb{N}$, $p \in [0,1]$, and let $X_1, \ldots, X_n$ be n independent random variables with $X_j \in \{0,1\}$ and $\Pr(X_j = 1) = p$ for all $j = 1, \ldots, n$. Let $X = \sum_{j=1}^{n} X_j$. Then, for any $t \in (p,1)$ and any $s \in (0,p)$, we have $\Pr(X \geq tn) \leq e^{-D(t,p)n}$ and $\Pr(X \leq sn) \leq e^{-D(s,p)n}$, where, for reals $x, y \in (0,1)$, $D(x,y) := x \log(x/y) + (1-x)\log((1-x)/(1-y))$ is the Kullback-Leibler divergence.

We can now state the main result of this section. The Lovász local lemma ideas in the proof of Lemma 26 are standard in the literature since the work of Moitra [39] but the quantities involved are adapted to our setting.

Lemma 26. There is a positive integer k_0 such that for any $k \geq k_0$ and any density α with $\alpha \leq 2^{(r_0-\delta)k}/k^3$ the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$:

1. there exists a partial assignment of bad variables that satisfies all bad clauses;
2. there exists an $(r_0 - \delta, r_0, r_0, 2r_0)$-marking of Φ. Furthermore, for any $p \in (0,1)$, such an $(r_0 - \delta, r_0, r_0, 2r_0)$-marking can be computed with probability at least $1-p$ in time $O(n \log(1/p))$.

Proof. In this proof we set $r = r_0 - \delta$. We note that for any $k \geq 4$ our density $\alpha \leq 2^{(r_0-\delta)k}/k^3$ is below the threshold $c_k > 1.3836 \cdot 2^k/k$ established in [23, Theorem 1.3]. For densities below this threshold, w.h.p. over the choice of Φ, there is a satisfying assignment for Φ. When Φ is satisfiable, we claim that there is an assignment of the bad variables that satisfies all bad clauses. Indeed, all the variables in bad clauses are bad (Proposition 7) and, thus, the restriction of a satisfying assignment to $\mathcal{V}_{\text{bad}}(r)$ must satisfy all the bad clauses. In the rest of this proof we show that assertion 2 also holds.

In view of Lemma 22, we may assume that $|\text{var}(c)| \geq k-1$ for all $c \in \mathcal{C}$. Let us find the $(r, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$. If all clauses are bad, then we set $\mathcal{V}_{\text{c}} = \mathcal{V}$, $\mathcal{V}_{\text{m}} = \emptyset$ and $\mathcal{V}_{\text{a}} = \emptyset$. This is trivially an $(r, r_0, r_0, 2r_0)$-marking for Φ. In the rest of the proof we assume that there are good variables. We study the following probability space. For each good variable v, we set v as "marked" with probability $\beta \in (0, 1/2)$, "auxiliary" with probability β and "control" with probability $1 - 2\beta$. This decision is made independently for each good variable. Each bad variable is set as "control". Let $\mathcal{P}$ be the set $\{P_v : v \in \mathcal{V}_{\text{good}}(r)\}$, where P_v is the random choice made in this experiment for v. Let $\mathcal{V}_{\text{m}}$ be the set of marked variables, let $\mathcal{V}_{\text{a}}$ be the set of auxiliary variables, and let $\mathcal{V}_{\text{c}}$ be the set of control variables obtained by running this experiment. For each clause $c \in \mathcal{C}_{\text{good}}(r)$, let A_c be the event that c has less than $r_0(k-3)$ marked variables or less than $r_0(k-3)$ auxiliary variables or less than $2r_0(k-3)$ good control variables. We are going to apply the LLL on the formula $\Phi_{\text{good}}(r)$ so as to show that $\Pr(\bigcap_{c \in \mathcal{C}_{\text{good}}(r)} \overline{A_c}) > 0$. For each $c \in \mathcal{C}_{\text{good}}(r)$, in view of Proposition 7 and the fact that $|\text{var}(c)| \geq k-1$, we have $|\text{var}(c) \cap \mathcal{V}_{\text{good}}(r)| \geq k-3$. Hence, we can apply the Chernoff bound given in Lemma 25 with $n = |\text{var}(c) \cap \mathcal{V}_{\text{good}}(r)|$, $p = \beta$ and $s = r_0$ to obtain, for any choice $V \in \{\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}\}$,

$$\Pr_P(|\text{var}(c) \cap V| < r_0(k-3)) \leq e^{-D(r_0,\beta)(k-3)}.$$

When $V = \mathcal{V}_{\text{c}} \setminus \mathcal{V}_{\text{bad}}$, $n = |\text{var}(c) \cap \mathcal{V}_{\text{good}}(r)|$, $p = 1 - 2\beta$ and $s = 2r_0$ we obtain

$$\Pr_P(|\text{var}(c) \cap V| < 2r_0(k-3)) \leq e^{-D(2r_0,1-2\beta)(k-3)}.$$

We have chosen r_0 to be as large as possible under the restrictions that $D(r_0, \beta) \geq r_0 \log 2$ and $D(2r_0, 1-2\beta) \geq r_0 \log 2$. The values $\beta = 0.571027$ and $r_0 = 0.117841$ satisfy these restrictions. We conclude that

$$\Pr_P(A_c) \leq 2 \cdot e^{-D(r_0,\beta)(k-3)} + e^{-D(2r_0,1-2\beta)(k-3)} \leq 3 \cdot 2^{-r_0(k-3)}.$$

Let $\Delta' = 2^{r_0(k-3)}/(3e^2k)$ and let $x(A_c) = 1/(k\Delta')$ for all $c \in \mathcal{C}_{\text{good}}(r)$. We check that x satisfies the condition of the LLL for $\mathcal{P}$ and $\mathcal{A} = \{A_c : c \in \mathcal{C}_{\text{good}}(r)\}$. For $k \geq 43$, $1/(k\Delta') \in (0,1)$ and thus $x(A_c) \in (0,1)$ for all $c \in \mathcal{C}_{\text{good}}(r)$. We note that $\Gamma(A_c) = \{A_{c'} : c' \in \mathcal{C}_{\text{good}}(r), c' \neq c, \text{var}(c') \cap \text{var}(c) \cap \mathcal{V}_{\text{good}}(r) \neq \emptyset\}$. The graph $G_{\Phi_{\text{good}}(r)}$, given in Definition 16, has maximum degree at most $k(\Delta_r - 1)$, so $|\Gamma(A_c)| \leq k(\Delta_r - 1) \leq k\Delta'$, where the latter inequality holds for large enough k as $\Delta_r = \lceil 2^{rk} \rceil$ and $r = r_0 - \delta$. Therefore, we have

$$x(A_c) \prod_{N \in \Gamma(A_c)} (1 - x(N)) \geq \frac{1}{k\Delta'} \left(1 - \frac{1}{k\Delta'}\right)^{k\Delta'} \geq \frac{1}{e^2 k\Delta'} = 3 \cdot 2^{-r_0(k-3)}, \tag{7}$$

where we used $(1 - 1/z)^z \geq e^{-2}$ for all $z \geq 2$ in the second inequality. Thus,

$$x(A_c) \prod_{N \in \Gamma(A_c)} (1 - x(N)) \geq 3 \cdot 2^{-r_0(k-3)} \geq \Pr_P(A_c).$$

We conclude that, by the LLL, $\Pr_P\left(\bigcap_{c \in \mathcal{C}_{\text{good}}(r)} \overline{A_c}\right) > 0$, so there exists a partition $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ of the variables of Φ such that $\mathcal{V}_{\text{bad}}(r) \subseteq \mathcal{V}_{\text{c}}$ and each good clause contains at least $r_0(k-3)$ marked variables, $r_0(k-3)$ auxiliary variables and $2r_0(k-3)$ good control variables. That is, $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ satisfies Definition 8 for $r = r_0 - \delta$, $r_{\text{m}} = r_0$, $r_{\text{a}} = r_0$, and $r_{\text{c}} = 2r_0$. Moreover, with probability at least $1 - \delta$, this partition can be computed in $4n\alpha\Delta' k \log(1/\delta)$ steps with the algorithm of Moser and Tardos [42]. □

We now give the marking result that we use in our connectivity results, which holds for densities at most $2^{(r_1-\delta)k}/k^3$, where $r_1 = 0.227092$. The larger density threshold comes from the fact that the marking result is less strong – we do not require auxiliary variables nor a high number of good control variables in every clause.

Lemma 27. There is a positive integer k_0 such that for any $k \geq k_0$ and any density α with $\alpha \leq 2^{(r_1-\delta)k}/k^3$ the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$:

1. there exists a partial assignment of bad variables that satisfies all bad clauses;
2. there exists an $(r_1 - \delta, r_1, 0, r_1)$-marking of Φ. Furthermore, for any $p \in (0,1)$, such an $(r_1 - \delta, r_1, 0, r_1)$-marking can be computed with probability at least $1 - p$ in time $O(n \log(1/p))$.

Proof. The proof is analogous to that of Lemma 26. Here we explain the main differences. First, we set $r = r_1 - \delta$ instead of $r = r_0 - \delta$. The second difference is that we study the following probability space: each good variable v is set as "marked" with probability β and "control" with probability $1 - \beta$. We let A_c be the event that c has less than $r_1(k-3)$ marked variables or less than $r_1(k-3)$ good control variables. A Chernoff bound as in the proof of Lemma 26 gives

$$\Pr_P(A_c) \leq e^{-D(r_1,\beta)(k-3)} + e^{-D(r_1,1-\beta)(k-3)} \leq 2 \cdot 2^{-r_1(k-3)},$$

where we chose r_1 as large as possible so that $D(r_1, \beta) \geq r_1 \log 2$ and $D(r_1, 1-\beta) \geq r_1 \log 2$. The choices $\beta = 1/2$ and $r_1 = 0.227092$ satisfy these restrictions. We let $\Delta' = 2^{r_1(k-3)}/(3e^2k)$ and let $x(A_c) = 1/(k\Delta')$ for all $c \in \mathcal{C}_{\text{good}}(r)$. It remains to check that we can apply the asymmetric LLL on the formula $\Phi_{\text{good}}(r)$ to conclude that $\Pr(\bigcap_{c \in \mathcal{C}_{\text{good}}(r)} \overline{A_c}) > 0$. This was done in equation (7) in Lemma 26. We note that the bound given in (7) also holds in our current setting if we replace r_0 by r_1. We find that $x(A_c) \prod_{N \in \Gamma(A_c)} (1 - x(N)) \geq 3 \cdot 2^{-r_1(k-3)} \geq \Pr_P(A_c)$ and, thus, there exists a partition $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ of the variables of Φ such that $\mathcal{V}_{\text{bad}}(r) \subseteq \mathcal{V}_{\text{c}}$, $\mathcal{V}_{\text{a}} = \emptyset$, and each good clause contains at least $r_1(k-3)$ marked variables and at least $r_1(k-3)$ good control variables. □

In the remaining of this section we bound the marginals of μ_Ω (recall that μ_Ω is the uniform distribution over the satisfying assignments of the formula Φ, Definition 9) on any marked and auxiliary variable. In fact, we prove the stronger result that the marginal distribution of μ_Ω on $\mathcal{V}_{\text{m}} \cup \mathcal{V}_{\text{a}}$ is ε-uniform, i.e., very close to the uniform distribution, see Definition 12. We give a bound for each one of the markings established in Lemmas 26 and 27. Here we write $\Lambda_1 \cup \Lambda_2$ for the combined assignment of Λ_1 and Λ_2.

Lemma 28. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a satisfiable k-CNF formula. The following claims hold.

1. Let $r = r_0 - \delta$ and let $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ be a $(r, r_0, r_0, 2r_0)$-marking of Φ. Then for any satisfying assignment Λ_{bad} of $\Phi_{\text{bad}}(r)$, any assignment $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ where $S \subseteq \mathcal{V}_{\text{m}} \cup \mathcal{V}_{\text{a}}$, and any $v \in \mathcal{V}_{\text{good}}(r) \setminus S$ we have

$$\max\left\{\Pr_{\mu_\Omega}\left(v \mapsto \mathsf{F} \middle| \Lambda \cup \Lambda_{\text{bad}}\right), \Pr_{\mu_\Omega}\left(v \mapsto \mathsf{T} \middle| \Lambda \cup \Lambda_{\text{bad}}\right)\right\} \leq \frac{1}{2} \exp\left(\frac{1}{k 2^{r_0 k}}\right).$$

   In particular, the distribution $\mu_\Omega|_{\mathcal{V}_{\text{m}} \cup \mathcal{V}_{\text{a}}}$ is $(2^{-r_0 k}/k)$-uniform.

2. Let $r = r_1 - \delta$ and let $(\mathcal{V}_{\text{m}}, \emptyset, \mathcal{V}_{\text{c}})$ be a $(r, r_1, \emptyset, r_1)$-marking of Φ. Then, for any satisfying assignment Λ_{bad} of $\Phi_{\text{bad}}(r)$, any assignment $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ where $S \subseteq \mathcal{V}_{\text{m}}$, and any $v \in \mathcal{V}_{\text{good}}(r) \setminus S$ we have

$$\max\left\{\Pr_{\mu_\Omega}\left(v \mapsto \mathsf{F} \middle| \Lambda \cup \Lambda_{\text{bad}}\right), \Pr_{\mu_\Omega}\left(v \mapsto \mathsf{T} \middle| \Lambda \cup \Lambda_{\text{bad}}\right)\right\} \leq \frac{1}{2} \exp\left(\frac{1}{k}\right).$$

   In particular, the distribution $\mu_\Omega|_{\mathcal{V}_{\text{m}}}$ is $(1/k)$-uniform.

Proof. We prove each one of the claims separately. The proofs are analogous so for the second claim we only highlight the differences in the proof.

1. Here $r = r_0 - \delta$. Let Λ_{bad} be an assignment of bad variables that satisfies all bad clauses. Let $S \subseteq \mathcal{V}_{\text{m}} \cup \mathcal{V}_{\text{a}}$, let Λ be an assignment of S to $\{\mathsf{F}, \mathsf{T}\}$, and let $v \in \mathcal{V}_{\text{good}}(r) \setminus S$. We note that $\Pr_{\mu_{\Omega^\tau}}(\cdot) = \Pr_{\mu_\Omega}(\cdot \,|\tau)$ for any assignment τ of some variables. In light of this observation, we are going to prove that

$$\max\left\{\Pr_{\mu_{\Omega^{\Lambda \cup \Lambda_{\text{bad}}}}}\left(v \mapsto \mathsf{F}\right), \Pr_{\mu_{\Omega^{\Lambda \cup \Lambda_{\text{bad}}}}}\left(v \mapsto \mathsf{T}\right)\right\} \leq \frac{1}{2} \exp\left(\frac{1}{k 2^{r_0 k}}\right). \tag{8}$$

   We apply the LLL to the formula $\Phi' := \Phi^{\Lambda \cup \Lambda_{\text{bad}}}$ as follows. Let $\mathcal{V}'$ and $\mathcal{C}'$ be the sets of variables and clauses of Φ'. Note that, $\mathcal{V}' \subseteq \mathcal{V}_{\text{good}}(r)$, $\mathcal{C}' \subseteq \mathcal{C}_{\text{good}}(r)$ and $G_{\Phi'}$ is a subgraph of $G_{\Phi_{\text{good}}(r)}$ as all bad variables have been assigned a value and all bad clauses have been satisfied. We set $P_v = \sigma(v)$ for all $v \in \mathcal{V}'$, where $\sigma \colon \mathcal{V}' \to \{\mathsf{F}, \mathsf{T}\}$ is chosen uniformly at

random from the set of assignments $\mathcal{V}' \to \{\mathsf{F},\mathsf{T}\}$, and $\mathcal{P} = \{P_v : v \in \mathcal{V}'\}$. We define the set $\mathcal{A}$ as the set containing for all $c \in \mathcal{C}'$ the event $A_c =$ "the clause c is not satisfied by the random assignment σ". By the definition of $(\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{a}, \mathcal{V}_\mathrm{c})$, there are at least $2r_0(k-3)$ good control variables in c. Since good control variables are not assigned a value by $\Lambda \cup \Lambda_\mathrm{bad}$ and, thus, they are in $\mathcal{V}'$, we have $\Pr_P(A_c) \leq 2^{-2r_0(k-3)}$. Recall that $\Delta_r = \lceil 2^{(r_0-\delta)k} \rceil$ (Definition 6). Let $\Delta' = 2^{2r_0(k-3)}/(e^2k)$ and let $x(A_c) = \frac{1}{k\Delta_0}$ for all $c \in \mathcal{C}'$. Let us show that x satisfies the LLL condition in this setting. In view of $\Gamma(A_c) = \{A_{c'} : c' \in \mathcal{C}', c' \neq c, \mathrm{var}(c) \cap \mathrm{var}(c') \cap \mathcal{V}' \neq \emptyset\}$, which can be identified with a subset of the neighbours of c in $G_{\Phi_\mathrm{good}(r)}$, and $|\Gamma(A_c)| \leq k\Delta_r \leq k\Delta'$ for large enough k, we find that

$$x(A_c) \prod_{N \in \Gamma(A_c)} (1 - x(N)) \geq \frac{1}{k\Delta'} \left(1 - \frac{1}{k\Delta'}\right)^{k\Delta'} \geq \frac{1}{e^2 k\Delta'} = 2^{-2r_0(k-3)} \geq \Pr_P(A_c),$$

where we used $(1-1/z)^z \geq e^{-2}$ for all $z \geq 2$. Let $A = \{v \mapsto \mathsf{T}\} := \{\sigma \colon \mathcal{V}' \to \{\mathsf{F},\mathsf{T}\}$ with $\sigma(v) = \mathsf{T}\}$. In Φ', we have $\Gamma(A) = \{A_c : c \in \mathcal{C}', v \in \mathrm{var}(c)\}$, so $|\Gamma(A)| < \Delta_r$. By the LLL, we obtain

$$\Pr_P\left(v \mapsto \mathsf{T} \,\middle|\, \bigcap\nolimits_{c \in \mathcal{C}'} \overline{A_c}\right) \leq \frac{1}{2} \prod_{N \in \Gamma(A)} \big(1 - x(N)\big)^{-1} \leq \frac{1}{2}\left(1 - \frac{1}{k\Delta'}\right)^{-(\Delta_r - 1)}.$$

For $x > 1$, we have $(1-1/x)^{-1} = 1 + 1/(x-1) \leq \exp(1/(x-1))$. We find that

$$\Pr_P\left(v \mapsto \mathsf{T} \,\middle|\, \bigcap\nolimits_{c \in \mathcal{C}'} \overline{A_c}\right) \leq \frac{1}{2} \exp\left(\frac{\Delta_r - 1}{k\Delta' - 1}\right) \leq \frac{1}{2} \exp\left(\frac{1}{k2^{r_0 k}}\right),$$

where in the latter inequality we used $(p-j)/(q-j) \leq p/q$ for all $0 < j < p \leq q$ and the fact that $\Delta_r = \lceil 2^{(r_0-\delta)k} \rceil \leq 2^{-r_0 k} \cdot 2^{2r_0(k-3)}/(e^2k) = 2^{-r_0 k}\Delta'$ for large enough k. We note that $\Pr_{\mu_{\Omega^{\Lambda \cup \Lambda_\mathrm{bad}}}}(\cdot) = \Pr_P\left(\cdot \,|\, \bigcap_{c \in \mathcal{C}'} \overline{A_c}\right)$, which completes the proof of one of the upper bounds of (8). The other upper bound is proved analogously by applying the LLL with $A = \{v \mapsto \mathsf{F}\}$. Finally, we conclude that the distribution $\mu_\Omega|_{\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}}$ is $(2^{-r_0 k}/k)$-uniform by the arbitrary choice of Λ_bad and the law of total probability, see Definition 12.

2. The proof is analogous. The only changes are $r = r_1 - \delta$, $\Delta' = 2^{r_1(k-3)}/(e^2k)$, and the fact that, since each good clause has at least $r_1(k-3)$ good control variables, we have $\Pr_P(A_c) \leq 2^{-r_1(k-3)}$. This time we have $x(A_c)\prod_{N \in \Gamma(A_c)}(1 - x(N)) \geq \frac{1}{e^2 k\Delta'} \geq \Pr_P(A_c)$, which justifies our choice of Δ'. Thus, we can apply the LLL, and the conclusion this time becomes

$$\Pr_P\left(v \mapsto \mathsf{T} \,\middle|\, \bigcap\nolimits_{c \in \mathcal{C}'} \overline{A_c}\right) \leq \frac{1}{2} \exp\left(\frac{\Delta_r - 1}{k\Delta' - 1}\right) \leq \frac{1}{2} \exp\left(\frac{1}{k}\right),$$

where in the latter inequality we used $(p-j)/(q-j) \leq p/q$ for all $0 < j < p \leq q$ and the fact that $\Delta_r = \lceil 2^{(r_1-\delta)k} \rceil \leq 2^{r_1(k-3)}/(e^2k) = \Delta'$ for large enough k. □

The $(1/k)$-uniform property proved in Lemma 28 is remarkably strong: as long as the control variables are left unassigned, the rest of the variables have marginals close to $1/2$, even if some of the marked and auxiliary variables are pinned / have already been assigned a value. This property is used several times in this work and will allow us to prove that, for any pinning of some marked variables, the influences between marked variables are bounded. In the following corollary we extend Lemma 28 to the distributions computed by the Glauber dynamics on the marked variables.

Corollary 29. Let $r = r_0 - \delta$. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a satisfiable k-CNF formula that has an $(r, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{a}, \mathcal{V}_\mathrm{c})$. Let ρ be an integer with $1 \leq \rho < |\mathcal{V}_\mathrm{m}|$. Let t be a non-negative integer and let X_t be the (random) assignment obtained after running the ρ-uniform-block Glauber dynamics on the marked variables for t steps, starting on an assignment X_0 that is chosen uniformly at random. Then the probability distribution of X_t is $(2^{-r_0 k}/k)$-uniform.

Proof. Let $\varepsilon = (2^{-r_0 k}/k)$. Let $V_1, V_2, \ldots,$ be a possible choice of sets of marked variables to be updated when running the ρ-uniform-block Glauber dynamics. We are going to prove that, conditioning on this choice of sets of variables, the probability distribution of X_t is ε-uniform. Note that by the law of total probability and the fact that the choice of $V_1, V_2, \ldots$ is arbitrary, this is enough to conclude the result. We carry out the proof by induction on t. Let π_t be the probability distribution of X_t. As π_0 is the uniform distribution over assignments on $\mathcal{V}_\mathrm{m}$, the claim holds for $t = 0$. Let us now assume that π_{t-1} is ε-uniform and let us prove that this is also the case for π_t. To show the desired uniformity of π_t (cf. Definition 12), consider arbitrary $v \in \mathcal{V}_\mathrm{m}$ and $\Lambda \colon \mathcal{V}_\mathrm{m} \setminus \{v\} \to \{\mathsf{F}, \mathsf{T}\}$, we need to bound $\Pr_{\pi_t}(v \mapsto \mathsf{F} | \Lambda)$ and $\Pr_{\pi_t}(v \mapsto \mathsf{T} | \Lambda)$. We distinguish two cases:

- Case $v \in V_t$. By definition of the Glauber dynamics, the values of X_t on V_t are obtained by sampling from the distribution μ_Ω conditioned on the restriction of X_{t-1} to $\mathcal{V}_\mathrm{m} \setminus V_t$. Thus, we have $\Pr_{\pi_t}(v \mapsto \mathsf{F} | \Lambda) = \Pr_{\mu_{\Omega^\Lambda}}(v \mapsto \mathsf{F})$ since the conditioning involving Λ sets all the marked variables other than v. As $\mu_\Omega|_{\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}}$ is ε-uniform by Lemma 28, we conclude that $\Pr_{\pi_t}(v \mapsto \mathsf{F} | \Lambda) = \Pr_{\mu_{\Omega^\Lambda}}(v \mapsto \mathsf{F}) \leq \frac{1}{2}\exp(\varepsilon)$. The same bound holds for $v \mapsto \mathsf{T}$.
- Case $v \notin V_t$. If v is not updated in steps 1 through t, then $\Pr_{\pi_t}(v \mapsto \mathsf{F} | \Lambda) = \Pr_{\pi_0}(v \mapsto \mathsf{F}) = 1/2$. Otherwise, let j be the largest integer with $j < t$ such that $v \in V_j$. Let Λ_j be the restriction of Λ to $\mathcal{V}_\mathrm{m} \setminus \bigcup_{i \in \{j+1, j+2, \ldots, t\}} V_i$. By the induction hypothesis, $\Pr_{\pi_t}(v \mapsto \mathsf{F} | \Lambda) = \Pr_{\pi_j}(v \mapsto \mathsf{F} | \Lambda_j) \leq (1/2)\exp(\varepsilon)$. The same bound holds for $v \mapsto \mathsf{T}$.

As both cases are exhaustive, the proof is concluded. □

Previous work on counting/sampling satisfying assignments of k-CNF formulae does not require the use of auxiliary variables, so the marking used is of the form $(\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{c})$. Here auxiliary variables play an essential role in bounding the influences between marked variables as we illustrated in Section 2. In order for this approach to be successful, we have to show that a large proportion of the variables are marked. We conclude this section with the following bound on the size of $\mathcal{V}_\mathrm{m}$.

Corollary 30. Let $r \in (0, 1/(2\log 2))$. There is an integer k_0 such that for any $k \geq k_0$ and any density α with $\alpha \leq \Delta_r / k^3$ the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For any $\rho \in (0, 1)$ and any set of good variables V that is ρ-distributed we have $|V| \geq (\rho - \delta)(k\alpha/\Delta_r)n$.

Proof. W.h.p. over the choice of Φ, by Lemma 21 we have $|\mathcal{C}_\mathrm{bad}(r)| \leq 2(\alpha/\Delta_r)n/2^{k^{10}} \leq \alpha n/4^k$, so $|\mathcal{C}_\mathrm{good}(r)| \geq |\mathcal{C}| - \alpha n/4^k \geq \alpha n - 1 - \alpha n/4^k = \alpha n(1 - 1/4^k) - 1$. Since V is ρ-distributed, counting repetitions, there are at least $\rho(k-3)|\mathcal{C}_\mathrm{good}(r)|$ occurrences of the variables of V in the good clauses of Φ. Each good variable occurs in at most Δ_r good clauses, so we find that

$$|V| \geq \frac{\rho(k-3)|\mathcal{C}_\mathrm{good}(r)|}{\Delta_r} \geq \frac{\rho(k-3)}{\Delta_r}\left(\alpha n\left(1 - \frac{1}{4^k}\right) - 1\right) \geq \frac{\rho(k-4)}{\Delta_r}(\alpha n - 1),$$

which is at least $(\rho - \delta)(k\alpha/\Delta_r)n$ for large enough k. □

6 Analysis of the connected components of Φ^Λ

In this section we prove Lemma 17, which bounds the size of the connected components of Φ^Λ, where Λ is drawn from a $(1/k)$-uniform distribution over an $(r+\delta)$-distributed set of good variables. In order to carry out this proof, we have to understand the structure of logarithmic-sized sets of clauses of the random k-CNF formula Φ. Section 6.1 is devoted to this purpose. In Section 6.2 we apply the results of Section 6.1 to obtain a lower bound of the number of marked/auxiliary variables in logarithmic-sized sets of clauses. Finally, in Section 6.3 we complete the proof of Lemma 17.

6.1 Logarithmic-sized sets of clauses in the random k-SAT model

A connected graph $H = (V, E)$ has tree-excess $c \in \mathbb{Z}_{\geq 0}$ if $|E| = c + |V| - 1$. It turns out that, w.h.p. over the choice of Φ, small connected sets of clauses of Φ have tree-excess bounded by a quantity that only depends on k and the density α. This property is established in Lemma 31 and is essential to our proofs.

Lemma 31. Let $k \geq 3$ be an integer. Let $b > 0$ and $\alpha > 0$ be real numbers. W.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, every connected subset of clauses with size at most $b \log(n)$ has tree-excess at most $c := \max\{1, 2b \log(ek^2\alpha)\}$.

Proof. Let n be the number of variables and m be the number of clauses of Φ, so $m/n \leq \alpha$. Note that the probability that two clauses of Φ are not disjoint is at most k^2/n. Let $\ell \in \{1, 2, \ldots, \lfloor b \log(n) \rfloor\}$. We upper bound the probability that there is a connected subset of clauses of size ℓ with tree-excess at least $c + 1$ by

$$\binom{m}{\ell} \ell^{\ell-2} \binom{\ell(\ell-1)/2}{c+1} \left(\frac{k^2}{n}\right)^{\ell+c}, \tag{9}$$

where the factors appearing are the following ones:

- $\binom{m}{\ell}$ is the number of subsets of clauses of size ℓ;
- $\ell^{\ell-2}$ is the number of trees on ℓ labelled vertices;
- $\binom{\ell(\ell-1)/2}{c+1}$ is the number of ways to pick $c + 1$ pairs of distinct clauses of a set of size ℓ;
- $\left(k^2/n\right)^{\ell+c}$ is an upper bound of the probability that all the edges chosen in the two previous items appear in the graph G_Φ.

We are going to show that the probability given in (9) is $O(n^{-c/4})$, where the hidden constant only depends on k. If this holds, by a union bound over $\ell \in \{1, 2, \ldots, \lfloor b \log(n) \rfloor\}$, we would find that the probability that there is a connected subset of clauses of Φ with size at most $b \log(n)$ and tree-excess at least $c + 1$ is $O(b \log(n) n^{-c/4}) = o(1)$. This would complete the proof. Using the inequality $\binom{p}{q} \leq (ep/q)^q$ and $m/n \leq \alpha$ we can bound (9) by

$$\begin{aligned}
\left(\frac{em}{\ell}\right)^\ell \ell^{\ell-2} \left(\frac{e\ell(\ell-1)/2}{c+1}\right)^{c+1} \left(\frac{k^2}{n}\right)^{\ell+c} &\leq \left(\frac{em}{\ell}\right)^\ell \ell^{\ell-2} \left(\frac{e\ell^2/2}{c+1}\right)^{c+1} \left(\frac{k^2}{n}\right)^{\ell+c} \\
&= \left(\frac{e}{2c+2}\right)^{c+1} \left(\frac{emk^2}{n}\right)^\ell \left(\frac{k^2\ell^2}{n}\right)^c \\
&\leq \left(\frac{e}{2c+2}\right)^{c+1} (ek^2\alpha)^\ell \left(\frac{k^2\ell^2}{n}\right)^c.
\end{aligned} \tag{10}$$

Now we distinguish two cases:

- Case when $ek^2\alpha \leq 1$. We have $c = 1$ by definition. Thus, (10) can be further bounded by

$$\left(\frac{e}{2c+2}\right)^{c+1}\left(\frac{k^2\ell^2}{n}\right)^{c} = O\left(\frac{(\log n)^2}{n}\right) = O\left(n^{-c/4}\right)$$

  as we wanted.

- Case when $ek^2\alpha > 1$. Then, as $\ell \leq b\log n$ and $b\log(ek^2\alpha) \leq c/2$ by definition, we have

$$\left(ek^2\alpha\right)^{\ell} \leq \left(ek^2\alpha\right)^{b\log n} = n^{b\log(ek^2\alpha)} \leq n^{c/2}.$$

  We conclude that (10) can be further bounded by

$$\left(\frac{e}{2c+2}\right)^{c+1}\left(\frac{k^2\ell^2}{\sqrt{n}}\right)^{c} = \left(\frac{e}{2c+2}\right)^{c+1}\left(\frac{k^4\ell^4}{n}\right)^{c/2} = O\left(n^{-c/4}\right)$$

  as we wanted, where we used $c > 0$. □

Recall that in Lemma 20 we established that, in sets of clauses that have at least $2k^4 \log n$ variables, the number of bad clauses of Φ is not too large. We aim to apply Lemma 20 to logarithmic-sized sets of clauses. In general, $|Y|$ might be significantly larger than $|\mathrm{var}(Y)|$, so it is not clear how to apply Lemma 20. However, in the random k-CNF formula setting the following holds.

Lemma 32. Let $k \geq 3$ be an integer and let $a > 0$ and $\alpha > 0$ be real numbers. W.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for every set of clauses Y with $|Y| \geq a \log n$, we have $|\mathrm{var}(Y)| \geq a \log n$.

Proof. Let $\ell := \lceil a \log n \rceil - 1$ and let $m = \lfloor \alpha n \rfloor$. We prove the equivalent statement that, w.h.p. over the choice of Φ, for every set of clauses Y with $|\mathrm{var}(Y)| \leq \ell$, we have $|Y| \leq \ell$. We note that if there is a set of clauses Y with $|\mathrm{var}(Y)| \leq \ell$ and $|Y| > \ell$, then for any subset Y' of Y with $|Y'| = \ell + 1$ we have $|\mathrm{var}(Y')| \leq |\mathrm{var}(Y)| \leq \ell$. Hence, it suffices to prove that there is no set Y of clauses with $|\mathrm{var}(Y)| \leq \ell$ and $|Y| = \ell + 1$. We can assume n is large enough so that $\ell \leq e \cdot n$.

Let $\mathcal{E}$ be the event that there is a set of clauses Y of size $\ell + 1$ and a set of variables X of size ℓ such that all clauses in Y have all variables in X. Then by a union bound

$$\Pr(\mathcal{E}) \leq \binom{m}{\ell+1}\binom{n}{\ell}\left(\frac{\ell}{n}\right)^{(\ell+1)k},$$

where the first factor is the number of sets Y, the second factor is the number of sets X and the third factor is the probability that all variables in the clauses of Y are in X. From the well-known bound $\binom{p}{q} \leq (ep/q)^q$, we obtain

$$\begin{aligned}\Pr(\mathcal{E}) &\leq \left(\frac{em}{\ell+1}\right)^{\ell+1}\left(\frac{en}{\ell}\right)^{\ell}\left(\frac{\ell}{n}\right)^{(\ell+1)k} \leq \left(\frac{em}{\ell}\right)^{\ell+1}\left(\frac{en}{\ell}\right)^{\ell+1}\left(\frac{\ell}{n}\right)^{(\ell+1)k} \\ &\leq \left(\frac{e\alpha n}{\ell}\right)^{\ell+1}\left(\frac{en}{\ell}\right)^{\ell+1}\left(\frac{\ell}{n}\right)^{(\ell+1)k} = \left(e^2\alpha\frac{\ell^{k-2}}{n^{k-2}}\right)^{\ell+1},\end{aligned}$$

which is $O(\log(n)/n)$ because $k \geq 3$ and $\ell = O(\log n)$. □

6.2 Number of marked variables in logarithmic-sized sets of clauses

Our results on random k-CNF formulae can now be combined to give a lower bound on the number of marked / auxiliary variables in logarithmic-sized sets of clauses. We prove this result in a more general setting by considering a set of good variables V that is r'-distributed for the formula Φ. The reader can think of V as the set of marked variables or the set of auxiliary variables for one of the markings established in Section 5.

Lemma 33. Let $r \in (0, 1/(2\log 2)]$, $r' \in (0,1)$ and $\hat{\delta} \in (0, r)$. There is a positive integer k_0 such that, for any integer $k \geq k_0$, any density $\alpha \leq \Delta_r/k^3$ and any real number b with $2k^4 < b$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let V be a set of good variables that is r'-distributed. Then, for every set of clauses Y that is connected in G_Φ such that $2k^4 \log(n) \leq |Y| \leq b\log(n)$, we have $|\mathrm{var}(Y) \cap V| \geq (r' - \hat{\delta})k|Y|$.

Proof. Let $a = 2k^4$. We apply Lemma 20 to find that there is k_1 such that for $k \geq k_1$, w.h.p. over the choice of Φ, for every set of clauses Y that is connected in G_Φ,

$$\text{if } \ |\mathrm{var}(Y)| \geq a\log(n), \ \text{ then } |Y \cap \mathcal{C}_{\mathrm{bad}}(r)| \leq |Y|/k. \tag{11}$$

We apply Lemma 32 with $a = 2k^4$ to find that, w.h.p. over the choice of Φ, for every set of clauses Y, we have

$$\text{if } \ |Y| \geq a\log(n), \text{then } |\mathrm{var}(Y)| \geq a\log(n). \tag{12}$$

Finally, for any $b > 0$, we apply Lemma 31, obtaining that, w.h.p. over the choice of Φ, for every set of clauses Y that is connected in G_Φ,

$$\text{if } |Y| \leq b\log n, \ \text{ then } Y \text{ has tree-excess at most } c = \max\{1, 2b\log(ek^2\alpha)\} = O(1). \tag{13}$$

Let Y be a set of clauses that is connected in G_Φ such that $a\log(n) \leq |Y| \leq b\log(n)$. Then, by (12) and (11), we have $|Y \cap \mathcal{C}_{\mathrm{good}}(r)| \geq |Y|(1-1/k)$. By definition of r'-distributed (Definition 8), each good clause has at least $r'(k-3)$ variables in V. As there are at most $|Y|-1+c$ edges in G_Φ joining clauses in Y, see (13), and two distinct clauses only share at most two variables by Lemma 22, we have

$$\begin{aligned} |\mathrm{var}(Y) \cap V| &\geq r'(k-3)\left(1 - \frac{1}{k}\right)|Y| - 2(|Y| + c - 1) \\ &\geq (r'(k-4) - 2)|Y| - 2(c-1). \end{aligned}$$

There is $k_0 \geq k_1$ such that for $k \geq k_0$, we find that, for any set of clauses Y that is connected in G_Φ and has $a\log(n) \leq |Y| \leq b\log(n)$, $|\mathrm{var}(Y) \cap V| \geq (r' - \hat{\delta}/2)k|Y| - 2(c-1)$. Therefore, using $2(c-1) = O(1)$, for large enough n we conclude that $|\mathrm{var}(Y) \cap V| \geq (r' - \hat{\delta})k|Y|$ and the result follows. □

6.3 Proof of Lemma 17

We use the following result of [24] on the number of connected sets of clauses in G_Φ.

Lemma 34 ([24, Lemma 8.6]). Let $\alpha > 0$. W.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for any clause c, the number of connected sets of clauses in G_Φ with size $\ell \geq \log n$ containing c is at most $(9k^2\alpha)^\ell$.

We can now complete the proof of Lemma 17. Recall that we will apply this result with $r = r_0 - \delta$ or $r = r_1 - \delta$, where $\delta = 0.00001$.

Lemma 17. Let $r \in (2\delta, 1/(2\log 2)]$. There is an integer $k_0 \geq 3$ such that, for any integer $k \geq k_0$, any density $\alpha \leq 2^{(r-2\delta)k}$, and any real number b with $a := 2k^4 < b$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$.

Let L be an integer satisfying $a \log n \leq L \leq b \log n$. Let V be a set of good variables of Φ that is $(r+\delta)$-distributed (Definition 8), let μ be a $(1/k)$-uniform distribution over the assignments $V \to \{\mathsf{F}, \mathsf{T}\}$, and let ρ be an integer with $0 \leq \rho \leq |V|/2^k$. Consider the following experiment. First, draw $S \subseteq V$ from the uniform distribution τ over subsets of V with size ρ. Then, sample an assignment Λ from $\mu|_{V \setminus S}$. Denote by $\mathcal{F}$ the event that that there is a connected set of clauses Y of Φ with $|Y| \geq L$ such that all clauses in Y are unsatisfied by Λ. Then $\Pr_{S\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{F}) \leq 2^{-\delta k L}\right) \geq 1 - 2^{-\delta k L}$.

Proof. We apply Lemma 33 with our choices of b and with $\hat{\delta} = \delta$ to conclude that, w.h.p. over the choice of Φ, for every connected set of clauses $Z \subseteq \mathcal{C}$ we have

$$\text{if} \quad a\log(n) \leq |Z| \leq b\log(n), \quad \text{then} \quad |\mathrm{var}(Z) \cap V| \geq rk|Z|. \tag{14}$$

We also need the following result on random k-CNF formulae. For each clause $c \in \mathcal{C}$, let

$$\mathcal{Z}(c, L) = \{Z \subseteq \mathcal{C} : c \in Z, Z \text{ is connected in } G_\Phi, |Z| = L\}.$$

Then, w.h.p. over the choice of Φ, Lemma 34 shows that, as long as $L \geq \log n$,

$$\text{for any clause } c \in \mathcal{C} \text{ we have } |\mathcal{Z}(c, L)| \leq (9k^2\alpha)^L. \tag{15}$$

The facts that we have just established using Lemma 33 and Lemma 34 are all the properties of random formulae that we need in this proof. The hypothesis $\alpha \leq \Delta_r$ is used when calling Lemma 20 in the proof of Lemma 33.

Let L be an integer with $a \log n \leq L \leq b \log n$. First, we are going to fix $S \subseteq V$ with $|S| = \rho$ and study the event $\mathcal{F}$ described in the statement. For $c \in \mathcal{C}$ and $Z \in \mathcal{Z}(c, L)$, we denote by $\mathcal{E}_1(Z, S)$ the event that $Z \subseteq \mathcal{C}^\Lambda$, where Λ is drawn from $\mu|_{V\setminus S}$, see Definition 11. Recall that $Z \subseteq \mathcal{C}^\Lambda$ means that none of the clauses in Z are satisfied by the assignment Λ (Definition 9). We note that $\mathcal{F} = \bigcup_{c\in\mathcal{C}, Z\in\mathcal{Z}(c,L)} \mathcal{E}_1(Z, S)$. We are going to show that, for large enough n,

$$\Pr_{S\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus S}}\left(\bigcup_{c\in\mathcal{C}, Z\in\mathcal{Z}(c,L)} \mathcal{E}_1(Z, S)\right) > 2^{-\delta k L}\right) \leq 2^{-\delta k L}, \tag{16}$$

which is equivalent to the result stated in this lemma. We note that the left-hand side of (16) can be upper bounded by

$$\begin{aligned}&\Pr_{S\sim\tau}\left(\exists c \in \mathcal{C}, Z \in \mathcal{Z}(c, L) : \Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z, S)) > \frac{2^{-\delta k L}}{|\mathcal{C}| \cdot |\mathcal{Z}(c, L)|}\right) \leq \\ &\sum_{c\in\mathcal{C}, Z\in\mathcal{Z}(c,L)} \Pr_{S\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z, S)) > \frac{2^{-\delta k L}}{|\mathcal{C}| \cdot |\mathcal{Z}(c, L)|}\right).\end{aligned} \tag{17}$$

We are going to show that, for any $c \in \mathcal{C}$ and $Z \in \mathcal{Z}(c, L)$,

$$\Pr_{S\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z, S)) > \frac{2^{-\delta k L}}{|\mathcal{C}| \cdot |\mathcal{Z}(c, L)|}\right) \leq \left(2ek \cdot 2^{-rk}\right)^L. \tag{18}$$

Before proving (18), let us complete the proof assuming that this inequality holds. In light of (15), we have $|\mathcal{Z}(c, L)| \leq (9k^2 2^{(r-2\delta)k})^L$. We use the following observation,

$$\text{for } k > 1/(\delta \log 2) \text{ and for large enough } n, \quad |\mathcal{C}| \leq n\alpha \leq n^{\delta k^5 \log 2} \leq 2^{(\delta/2)kL}. \tag{19}$$

Combining (17), (18) and (19), we conclude that, for large enough k, the left-hand size of (16) is bounded above by

$$\sum_{c\in\mathcal{C},Z\in\mathcal{Z}(c,L)}\left(2ek\cdot 2^{-rk}\right)^L \leq n\alpha\cdot\left(9k^2 2^{(r-2\delta)k}\right)^L\cdot\left(2ek\cdot 2^{-rk}\right)^L = n\alpha\left(18ek^3 2^{-2\delta k}\right)^L \leq 2^{-\delta kL},$$

which completes the proof of (16), and hence the proof of the lemma, subject to (18).

To prove (18), we are going to find many S for which $\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S)) \leq 2^{-\delta kL}/(|\mathcal{C}|\cdot|\mathcal{Z}(c,L)|)$ holds. With this in mind, we introduce an event that may occur when sampling S:

$$\begin{aligned}\mathcal{E}_2(Z) :=&\text{“the random set } S\subseteq V \text{ that we select contains fewer}\\ &\text{than } \ell := \lceil|\mathrm{var}(Z)\cap V|/k\rceil \text{ variables in } \mathrm{var}(Z)\cap V\text{”}.\end{aligned} \tag{20}$$

We will show (in equation (24)) that the event $\mathcal{E}_2(Z)$ holds for most choices of S. Before proving this claim, let us assume that $\mathcal{E}_2(Z)$ holds for S and let us prove that $\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S)) \leq 2^{-\delta kL}/(|\mathcal{C}|\cdot|\mathcal{Z}(c,L)|)$. If there are $c_1,c_2\in Z$ and $v\in\mathrm{var}(c_1)\cap\mathrm{var}(c_2)\cap(V\setminus S)$ such that $c_1\neq c_2$ and the literal of v in c_1 is the negation of the literal of v in c_2, then at least one of c_1 and c_2 is satisfied by the assignment $\Lambda\colon V\setminus S\to\{\mathsf{F},\mathsf{T}\}$. In this case we have $\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S))=0$. Let us now consider the complementary case:

$$\begin{aligned}&\text{for all } c_1,c_2\in Z \text{ with } c_1\neq c_2 \text{ and } v\in\mathrm{var}(c_1)\cap\mathrm{var}(c_2)\cap(V\setminus S),\\ &\text{the literal of } v \text{ in } c_1 \text{ is the same as the literal of } v \text{ in } c_2.\end{aligned} \tag{21}$$

In this setting, we call $\omega(v)$ the value of v that does not satisfy the clauses in Z that contain v. Note that $\omega(v)$ is well-defined by assumption (21). Let $u_1,u_2,\ldots,u_t$ be the list of variables in $(\mathrm{var}(Z)\cap V)\setminus S$. We denote by $\mathcal{W}_j$ the event that u_j is assigned the value $\omega(u_j)$ by Λ when sampling $\Lambda\sim\mu|_{V\setminus S}$. Then, by definition of $\mathcal{W}_j$, we have

$$\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S)) = \prod_{j=1}^{t}\Pr_{\Lambda\sim\mu|_{V\setminus S}}\left(\mathcal{W}_j\,\middle|\,\bigcap_{i=1}^{j-1}\mathcal{W}_i\right).$$

As μ is $(1/k)$-uniform, we find that $\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{W}_j|\bigcap_{i=1}^{j-1}\mathcal{W}_i)\leq(1/2)\exp(1/k)$ for all $j\in\{1,2,\ldots,t\}$. We conclude that

$$\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S)) \leq \left(\frac{1}{2}\exp\left(\frac{1}{k}\right)\right)^t.$$

From (14) and the fact that $\mathcal{E}_2(Z)$ holds for S, we have

$$t=|\mathrm{var}(Z)\cap(V\setminus S)|\geq|\mathrm{var}(Z)\cap V|-\lceil|\mathrm{var}(Z)\cap V|/k\rceil\geq|\mathrm{var}(Z)\cap V|(1-1/k)-1\geq rL(k-1)-1.$$

It follows that

$$\begin{aligned}\Pr_{\Lambda\sim\mu|_{V\setminus S}}(\mathcal{E}_1(Z,S)) &\leq \left(\frac{1}{2}\exp\left(\frac{1}{k}\right)\right)^{r(k-1)L-1}\\ &\leq 2\left(2\cdot 2^{-rk}\exp\left(\frac{r(k-1)}{k}\right)\right)^L\\ &\leq \left(4e\cdot 2^{-rk}\right)^L,\end{aligned}$$

where we used that $1/2 \leq (1/2)\exp(1/k) < 1$ in the second and third inequality. For large enough k, we find that

$$\left(4e \cdot 2^{-rk}\right)^L = \left(\frac{9 \cdot 4ek^2 \cdot \alpha \cdot 2^{-rk}}{9k^2\alpha}\right)^L \leq \left(\frac{9 \cdot 4ek^2 \cdot 2^{-2\delta k}}{9k^2\alpha}\right)^L \leq \frac{2^{-(3/2)\delta kL}}{|\mathcal{Z}(c,L)|} \leq \frac{2^{-\delta kL}}{|\mathcal{C}| \cdot \mathcal{Z}(c,L)|}, \quad (22)$$

where in the second to last inequality we applied $9 \cdot 4ek^2 \leq 2^{(\delta/2)k}$ and the bound on the size of $\mathcal{Z}(c,L)$ given in (15), and in the last inequality we used (19). As S was picked as any subset of V with $|S| = \rho$ such that $\mathcal{E}_2(Z)$ holds, it follows that

$$\Pr_{S\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus S}}\left(\mathcal{E}_1(Z,S)\right) > \frac{2^{-\delta kL}}{|\mathcal{C}| \cdot |\mathcal{Z}(c,L)|}\right) \leq \Pr_{S\sim\tau}\left(\overline{\mathcal{E}_2(Z)}\right). \quad (23)$$

In order to prove (18), which finishes the proof, we need to show $\Pr_{S\sim\tau}\left(\overline{\mathcal{E}_2(Z)}\right) \leq (2ek \cdot 2^{-rk})^L$. The probability of $\overline{\mathcal{E}_2(Z)}$ can be bounded as follows. Recall that $|S| = \rho$. If $\rho < \ell$, then, by the definition of $\mathcal{E}_2(Z)$ in (20), we obtain $\Pr_{S\sim\tau}(\mathcal{E}_2(Z)) = 1$. Otherwise, the number of choices of S (with $|S| = \rho$) such that $|S \cap \mathrm{var}(Z) \cap V| \geq \ell$ is at most $\binom{|\mathrm{var}(Z)\cap V|}{\ell}\binom{|V|-\ell}{\rho-\ell}$. Hence, we have

$$\begin{aligned}
\Pr_{S\sim\tau}\left(\overline{\mathcal{E}_2(Z)}\right) &\leq \binom{|V|}{\rho}^{-1}\binom{|\mathrm{var}(Z)\cap V|}{\ell}\binom{|V|-\ell}{\rho-\ell} \\
&= \frac{\rho(\rho-1)\cdots(\rho-\ell+1)}{|V|(|V|-1)\cdots(|V|-\ell+1)}\binom{|\mathrm{var}(Z)\cap V|}{\ell} \\
&\leq \left(\frac{\rho}{|V|}\right)^\ell \left(\frac{e|\mathrm{var}(Z)\cap V|}{\ell}\right)^\ell \leq \left(\frac{\rho}{|V|}ek\right)^\ell,
\end{aligned}$$

where we used $\ell := \lceil |\mathrm{var}(Z)\cap V|/k \rceil \geq |\mathrm{var}(Z)\cap V|/k$, $(p-i)/(q-i) \leq p/q$ for any $0 < i < p < q$ and $\binom{p}{q} \leq (ep/q)^q$. Combining this with the hypothesis $\rho \leq |V|/2^k$ and the bound $\ell \geq rL$, see (14), we obtain

$$\Pr_{S\sim\tau}\left(\overline{\mathcal{E}_2(Z)}\right) \leq \left(ek2^{-k}\right)^\ell \leq \left((ek)^r \cdot 2^{-rk}\right)^L \leq \left(2ek \cdot 2^{-rk}\right)^L. \quad (24)$$

The bound (18) follows from combining (23) and (24), which completes the proof. □

7 Sampling from small connected components

In this section we prove Lemma 19. Recall that Lemma 19 claims the existence of a procedure to sample from marginals of the uniform distribution on the satisfying assignments of Φ^Λ when the connected components of G_{Φ^Λ} have small size. Here we make this procedure explicit. Our algorithm exploits the fact that the tree-excess of logarithmic-sized subsets of G_Φ is bounded by a constant depending only on k, see Lemma 31, and the fact that when G_Φ is acyclic, we can exactly count and sample satisfying assignments efficiently via a dynamic programming algorithm (Proposition 35).

Proposition 35. There is an algorithm that, for any k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$ such that G_Φ is a tree, computes the number of satisfying assignments of Φ in time $O(4^k|\mathcal{C}|)$.

Proof. We give an algorithm based on dynamic programming. Let us fix a vertex / clause c of G_Φ as the root and consider the corresponding directed tree structure $T := (G_\Phi, c)$. For any clause c' of Φ, let $T_{c'}$ be the subtree of T hanging from c'. For any assignment $\sigma\colon \mathrm{var}(c') \to \{\mathsf{F}, \mathsf{T}\}$, let

$\mathrm{sa}(c', \sigma)$ denote the number of satisfying assignments of the formula determined by $T_{c'}$ that extend σ. Our goal is computing the number of satisfying assignments of Φ, which, under this notation, is equal to

$$\mathrm{sa}(\Phi) := \sum_{\sigma\colon\ \mathrm{var}(c) \to \{\mathsf{F},\mathsf{T}\}} \mathrm{sa}(c, \sigma). \tag{25}$$

We do this by computing $\mathrm{sa}(c', \sigma)$ for any clause c' and any assignment $\sigma\colon\ \mathrm{var}(c') \to \{\mathsf{F}, \mathsf{T}\}$. Using the tree structure of T, we show that $\mathrm{sa}(c', \sigma)$ satisfies a recurrence. There are two cases:

1. c' is a leaf. Then $\mathrm{sa}(c', \sigma) = 1$ if c' is satisfied by σ and 0 otherwise.

2. c' is not a leaf. Let $T_1, \ldots, T_l$ be the trees hanging from c' in T and let $c_1, \ldots, c_l$ be their roots. Then, since $T_1, \ldots, T_l$ do not share variables as G_Φ is acyclic, we have

$$\mathrm{sa}(c', \sigma) = \prod_{j=1}^{l} \sum_{\tau \in A(c_j, \sigma)} \mathrm{sa}(c_j, \tau),$$

   where $A(c_j, \sigma)$ is the set of assignments of the variables in $\mathrm{var}(c_j)$ that agree with σ on $\mathrm{var}(c') \cap \mathrm{var}(c_j)$.

We can apply this recurrence with dynamic programming to compute $\mathrm{sa}(c, \sigma)$ for any assignment $\sigma\colon\ \mathrm{var}(c) \to \{\mathsf{F}, \mathsf{T}\}$. More explicitly, we compute $\mathrm{sa}(c', \sigma)$ by levels of the tree, starting from the deepest level, where all nodes are leaves, and ending at the root c. This involves computing at most 2^k entries $\mathrm{sa}(c', \cdot)$ per clause c' of Φ. After computing all the entries appearing in this recurrence, we compute the number of satisfying assignments of Φ, $\mathrm{sa}(\Phi)$, as in equation (25). The overall procedure takes at most $O(4^k|\mathcal{C}|)$ steps since each entry $\mathrm{sa}(c', \sigma)$ is accessed at most 2^k times when computing the corresponding entries for the parent of c', and there are at most $2^k|\mathcal{C}(T)|$ entries. □

In Algorithm 3 we give an algorithm based on Proposition 35 to count satisfying assignments of a k-CNF formula. Recall the folklore fact that if we can count satisfying assignments then we can sample from the marginal of μ_Ω on v by counting the satisfying assignments of $\Phi^{v \mapsto \mathsf{F}}$ and $\Phi^{v \mapsto \mathsf{T}}$.

Algorithm 3 Counting satisfying assignments via trees

Input: a k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$
Output: The number of satisfying assignments of Φ.

1: Find a spanning forest T of G_Φ.
2: Let $\mathcal{V}_T$ be the set of variables that gives rise to edges of G_Φ that are not in T.
3: $count \leftarrow 0$.
4: for all $\Lambda\colon \mathcal{V}_T \to \{\mathsf{F}, \mathsf{T}\}$ do
5: Note that the graph G_{Φ^Λ} is acyclic. Hence, we can count the number of satisfying assignments of Φ^Λ in time $O(4^k|\mathcal{C}(\Phi^\Lambda)|)$ by applying Proposition 35 to each connected component of G_{Φ^Λ} and taking the product of the numbers obtained. Let $\mathrm{sa}(\Phi^\Lambda)$ be the result of this computation.
6: $count \leftarrow count + \mathrm{sa}(\Phi^\Lambda)$.
7: end for
8: return $count$

Proposition 36. Let $\Phi = (\mathcal{V}, \mathcal{C})$ be a k-CNF formula and let c be the tree-excess of G_Φ. Then Algorithm 3 counts the number of satisfying assignments of Φ in time $O(2^{k(c+2)}|\mathcal{C}|)$.

Proof. We note that, in the execution of Algorithm 3, we have $|\mathcal{V}_T| \leq kc$. Hence, there are at most 2^{kc} iterations of the for loop and each one takes $O(4^k|\mathcal{C}|)$ steps, so the running time follows. The fact that the algorithm is correct follows from the correctness of the procedure presented in Proposition 35. □

Even though the running time of Algorithm 3 is not polynomial in the size of the formula Φ (in fact, it is exponential in general), we obtain linear running time when the formulae considered have constant tree-excess. As shown in Lemma 31, this is the case for logarithmic-sized subsets of clauses of random formulae. We can now finish the proof of Lemma 19.

Lemma 19. There is an integer $k_0 \geq 3$ such that, for any integers $k \geq k_0$, $b \geq 2k^4$ and any density $\alpha > 0$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let V be a subset of variables and let $\Lambda \colon V \to \{\mathsf{F}, \mathsf{T}\}$ be a partial assignment such that all the connected components in G_{Φ^Λ} have size at most $b \log(n)$. Then, there is an algorithm that, for any $S \subseteq \mathcal{V} \setminus V$, samples an assignment from $\mu_{\Omega^\Lambda}|_S$ in time $O(|S| \log n)$.

Proof. We apply Lemma 31, so, w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, any connected set of clauses in G_Φ with size at most $b \log(n)$ has tree-excess at most $c = \max\{1, 2b \log(e\alpha k^2)\} = O(1)$. First, we give an algorithm for the case $|S| = 1$. Let Φ, V and Λ as in the statement, and let $S = \{v\}$. Let H be the connected component of the clauses that contain v in G_{Φ^Λ}, and let $\Phi' = (\mathcal{V}', \mathcal{C}')$ be the subformula of Φ^Λ with $G_{\Phi'} = H$. The formula Φ' has size at most $b \log(n)$. Moreover, the graph $G_{\Phi'} = H$ has tree-excess at most c as H is a subgraph of G_Φ with size at most $b \log(n)$. Thus, we can apply Proposition 36 to count the number of satisfying assignments of $\Phi'^{v \to \mathsf{F}}$ and $\Phi'^{v \to \mathsf{T}}$ in time $O(2^{k(c+2)}|\mathcal{C}'|) = O(\log n)$. Let these numbers be t_0 and t_1 respectively. It is straightforward to use t_0 and t_1 to sample from the marginal of the distribution μ_{Ω^Λ} for v; we only have to sample an integer $t \in [0, t_0 + t_1)$ and output F if $t < t_0$ and T otherwise. The whole process takes time $O(\log n)$.

Finally, we argue how to extend this algorithm to the case $|S| > 1$. For this, first, we give an order to the variables in S, say $u_1, u_2, \ldots, u_\ell$. We then call the algorithm described in the paragraph above once for each variable in $u_1, u_2, \ldots, u_\ell$. The inputs of the algorithm in the j-th call are the variable u_j and the assignment $\Lambda_j = \Lambda \cup \tau_{j-1}$, where τ_{j-1} is the assignment obtained in the previous calls for $u_1, \ldots, u_{j-1}$. After this process, τ_ℓ is an assignment of all the variables in S that follows the distribution $\mu_{\Omega^\Lambda}|_S$. This assignment has been computed in $O(|S| \log n)$ steps as we wanted. □

8 Mixing time of the Markov chain

In this section we study the mixing time of the ρ-uniform-block Glauber dynamics on the marked variables and prove Lemma 15. As explained in Section 2.2, in order to conclude rapid mixing of this Markov chain we apply the spectral independence framework, which has recently been extended to the ρ-uniform-block Glauber dynamics [11]. Traditionally in path coupling or spectral independence arguments one has to bound a sum of influences by a constant in order to obtain rapid mixing of the single-site Glauber dynamics. However, due to the presence of high-degree variables, an $O(1)$ upper bound seems unattainable in the random k-SAT formula setting; in the worst case paths of high-degree variable may significantly affect influences. This seems also to be the case for other random models, such as the hardcore model on random graphs [8]. Here we show that that sums of influences are at most $\varepsilon \log n$ for small ε (Lemma 14). Even though this is generally not enough to conclude rapid mixing of the single-site Glauber dynamics, it turns out to be enough to conclude rapid mixing of the ρ-uniform-block Glauber dynamics for $\rho = \Theta(n)$. An

essential ingredient in our argument is exploiting the auxiliary variables in introduced in Section 5. Therefore, in this section we will work with $r = r_0 - \delta$ and a $(r, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$. Since r is fixed, we drop it from the notation and write, for instance, $\mathcal{V}_{\mathrm{good}}$ instead of $\mathcal{V}_{\mathrm{good}}(r)$.

This section is divided as follows. In Section 8.1, we explain why bounded-degree methods to bound the mixing time of the Glauber dynamics fail to generalise from the bounded-degree k-SAT model to the random k-SAT model. In Section 8.2 we prove Lemmas 45 and 14. In Section 8.3 we prove Lemma 15.

8.1 Previous work on the Glauber dynamics for bounded-degree k-SAT formulae

In this section we explain why previously known arguments for showing rapid mixing of the Glauber dynamics on bounded-degree k-SAT formulae do not extend to the random k-SAT model. This section is not used in our work and may be skipped by a reader who just wants to understand our approach and result. The best result currently known on bounded-degree formulae is [31], where the authors show, for large enough k, how to efficiently sample satisfying assignments of k-CNF formulae in which their variables have maximum degree $\hat{\Delta} \leq C\, 2^{0.1742 \cdot k}/k^3$, where $C > 0$ is a constant that does not depend on k.[1] Their result actually holds in the more general setting of atomic constrain satisfaction problems (albeit with a different bound on $\hat{\Delta}$). As part of their work, they show that the single-site Glauber dynamics on a set of marked variables mixes quickly. Their argument is restricted to atomic CSPs with bounded-degree and strongly exploits the properties of the Glauber dynamics in this setting. They study the optimal coupling of the single-site Glauber dynamics, we refer to [38] for the definition of coupling of Markov chains. In such a coupling the goal is to show that two copies of the chain starting from truth assignments differing in at least a marked variable (a so-called discrepancy) can be coupled in a small number of steps. Here it is crucial that the marginals of the marked variables are near $1/2$, so the optimal coupling generates new discrepancies with small probability. At this stage, the high-level idea to conclude rapid mixing of the Glauber dynamics is bounding the probability that the dynamics has not coupled by a product of probabilities, each corresponding to the event that a clause is unsatisfied at a certain time, and aggregating over all possible discrepancy sequences.

The fundamental observation in [31], based on the work on monotone k-CNF formulae presented in [29], is that if there is an update of a marked variable that generates a discrepancy in the chains, then there is another marked variable where the chains disagree that is connected to the former variable through a path of clauses, where consecutive clauses in the path share at least a variable. Moreover, each one of the clauses in this path is unsatisfied by at least one of the two copies of the chain. As a consequence, from a discrepancy at time t one can find a sequence of discrepancies going back to time 0, and these discrepancies are joined by a path of clauses. Thus, the union bound over discrepancy sequences is essentially a union bound over paths of clauses with a particular time structure, where the same clause can be appear in the path several times. Extending this idea to the random k-SAT model presents two main issues. First of all, the number of discrepancy sequences of any given length may be too large due to the presence of bad clauses and the fact that they can repeatedly appear in the sequence. Moreover, it may be the case that these discrepancy sequences mostly consist of bad clauses, which are always unsatisfied in both chains and, thus, the probability that they are unsatisfied is not small. Interestingly, similar issues arise when directly extending the bounded-degree approach based on the coupling process of [39, 21] to our setting. In [21] the mixing time argument only succeeds when $\hat{\Delta} \leq 2^{k/20}/(8k)$ and is also based on a union bound over path of clauses that are unsatisfied or contain discrepancies after running a coupling process.

[1] In [31] the maximum degree $\hat{\Delta}$ of Φ is defined as the maximum over $c \in \mathcal{C}$ of the number of clauses that share a variable with c. Under this definition of $\hat{\Delta}$, their result holds for $\hat{\Delta} \leq C 2^{0.1742 \cdot k}/k^2$.

However, very importantly, these paths of clauses are simple (clauses are not repeated) and the combinatorial structures appearing in the coupling process are less complex than the discrepancy sequences of [31]. This allowed the authors of [24] to exploit the expansion properties of random k-CNF formulae to analyse the coupling process of [39] on the random setting. Here we incorporate novel ideas to the work of [24] in order to obtain a tighter analysis that leads to nearly linear running time of our sampling algorithm.

8.2 Spectral independence in the k-SAT model

In this section we prove Lemma 14. In order to bound the sum of influences of marked variables, we follow the coupling process technique that is standard in the literature [24, 39, 21]. In this work we introduce the concept of auxiliary variables in the coupling process and exploit the sparsity properties of logarithmic-sized sets of clauses, which allows us to conclude a $2^{-r_0 k}\log n$ spectral independence bound. The key idea is that if we progressively extend two assignments X and Y on auxiliary variables following the optimal coupling, with high probability over X and Y, at some point the formulae Φ^X and Φ^Y factorise in small connected components in spite of the presence of bad variables and, on top of that, Φ^X and Φ^Y share most of these connected components. Then we can bound influences between marked variables by analysing the connected components where Φ^X and Φ^Y differ. First, let us introduce the notation and results on couplings that we need.

Let μ and ν be two distributions over the same space $\widehat{\Omega}$. A coupling τ of μ and ν is a joint distribution over $\widehat{\Omega}\times\widehat{\Omega}$ such that the projection of τ on the first coordinate is μ and the projection on the second coordinate is ν. Recall that the total variation distance of μ and ν is defined by $d_{\mathrm{TV}}(\mu,\nu)=\frac{1}{2}\sum_{x\in\widehat{\Omega}}|\mu(x)-\nu(x)|$. If a random variable X has distribution μ, we also write $d_{\mathrm{TV}}(X,\nu)$ to mean $d_{\mathrm{TV}}(\mu,\nu)$. An important property of couplings is the coupling lemma.

Proposition 37 (Coupling lemma). Let τ be a coupling of μ and ν. Then $d_{\mathrm{TV}}(\mu,\nu)\leq \Pr_{(X,Y)\sim\tau}(X\neq Y)$. Moreover, there exists a coupling that achieves equality.

The coupling τ of μ and ν that minimises $\Pr_{(X,Y)\sim\tau}(X\neq Y)$ is called optimal. Let us now assume that μ and ν are Bernoulli distributions with parameters $0\leq p\leq q\leq 1$ respectively, so $\Pr_\mu(X=1)=p$ and $\Pr_\nu(Y=1)=q$. The monotone coupling τ of μ and ν is defined as follows. We pick U uniformly at random in $[0,1]$ and set $X=1$ only when $U\leq p$ and $Y=1$ only when $U\leq q$. For this coupling we have $\Pr_{(X,Y)\sim\tau}(X\neq Y)=q-p=d_{\mathrm{TV}}(X,Y)$ and, hence, the monotone coupling is optimal. This optimal coupling will come up in the coupling process when sampling from the marginals of auxiliary variables.

Before presenting our coupling process, we show how we can bound a sum of influences between marked variables with the help of the coupling lemma. In all this section we fix a k-CNF formula Φ and a $(r,r_0,r_0,2r_0)$-marking $(\mathcal{V}_{\mathrm{m}},\mathcal{V}_{\mathrm{a}},\mathcal{V}_{\mathrm{c}})$ of Φ. Given two assignments Λ_1 and Λ_2 on disjoint sets of variables, recall that we denote by $\Lambda_1\cup\Lambda_2$ the combined assignment on the union of their domains.

Proposition 38. Let $u\in\mathcal{V}_{\mathrm{m}}$ and $\Lambda\colon S\to\{\mathsf{F},\mathsf{T}\}$ with $S\subseteq\mathcal{V}_{\mathrm{m}}\setminus\{u\}$. Let (X,Y) be a coupling where X follows the distribution $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{T}}}|_{\mathcal{V}_{\mathrm{m}}}$ and Y follows the distribution $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{F}}}|_{\mathcal{V}_{\mathrm{m}}}$. Then

$$\sum_{v\in\mathcal{V}_{\mathrm{m}}\setminus(S\cup\{u\})}\left|\mathcal{I}^{\Lambda}(u\to v)\right|\leq\sum_{v\in\mathcal{V}_{\mathrm{m}}\setminus(S\cup\{u\})}\Pr\left(X(v)\neq Y(v)\right). \tag{26}$$

Proof. Let $v\in\mathcal{V}_{\mathrm{m}}$. Then for any $\omega\in\{\mathsf{F},\mathsf{T}\}$, we have $\Pr(v\mapsto\omega|\Lambda,u\mapsto\mathsf{T})=\Pr(X(v)=\omega)$ and $\Pr(v\mapsto\omega|\Lambda,u\mapsto\mathsf{F})=\Pr(Y(v)=\omega)$. Thus, by the coupling lemma,

$$\left|\mathcal{I}^{\Lambda}(u\to v)\right|=\left|\Pr(X(v)=\mathsf{T})-\Pr(Y(v)=\mathsf{T})\right|=d_{\mathrm{TV}}(X(v),Y(v))\leq\Pr\left(X(v)\neq Y(v)\right),$$

and the proof follows by adding over $v\in\mathcal{V}_{\mathrm{m}}\setminus(S\cup\{u\})$. □

For two assignments X and Y on a subset of variables V, we say that X and Y have a discrepancy at $v \in V$ when $X(v) \neq Y(v)$. In [21] the authors manage to bound (26) by a constant that does not depend on n when the considered formula has bounded degree. However, their argument breaks under the presence of high-degree variables due to the fact that we cannot control the number of bad clauses in a path of clauses unless the path has length at least $\Omega(\log n)$. Here instead we perform the coupling process developed in [24] over auxiliary variables, which accounts for the presence of bad clauses.

Before presenting our algorithm for the coupling process on auxiliary variables, let us describe some of the notation and structures that are used in this algorithm. Let $u \in \mathcal{V}_{\mathrm{m}}$ and $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $S \subseteq \mathcal{V}_{\mathrm{m}} \setminus \{u\}$. We start with two assignments $\widehat{X}$ and $\widehat{Y}$ that have a discrepancy at u and agree with Λ on S. In the coupling process we identify a set of failed clauses, denoted $\mathcal{F}_{\mathrm{d}} \cup \mathcal{F}_{\mathrm{u}}$. At each step of the process, we check if a clause is failed or extend the coupling to an auxiliary variable. It is important in our arguments that all clauses containing a discrepancy are failed, and that we make sure that the set of failed clauses is connected in G_Φ at all times. In order to achieve connectivity of failed clauses, at each step of the coupling process we only consider clauses that are adjacent to failed clauses in G_Φ. For ease of reading, here we present a list of the structures that appear in our algorithm.

1. $\mathcal{V}_{\mathrm{d}}$. Set of discrepancies, i.e., variables v with $\widehat{X}(v) \neq \widehat{Y}(v)$.
2. $\mathcal{F}_{\mathrm{d}}$. Set of all clauses containing a variable in $\mathcal{V}_{\mathrm{d}}$. These are failed clauses.
3. $\mathcal{V}_{\mathrm{set}}$. Set of variables that are assigned a value in the coupling.
4. $\mathcal{F}_{\mathrm{u}}$. Set of clauses that have been considered by the coupling process, and are either bad, or are unsatisfied by at least one of $\widehat{X}$ and $\widehat{Y}$ and have all their auxiliary variables in $\mathcal{V}_{\mathrm{set}}$. These are failed clauses.
5. $\mathcal{C}_{\mathrm{rem}}$. Set of clauses that have unassigned auxiliary variables or have not been explored yet.

Our coupling process on auxiliary variables is given in Algorithm 4.

Algorithm 4 The coupling process on auxiliary variables

Input: A k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$, an $(r, r_0, r_0, 2r_0)$-marking $\mathcal{M} = (\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{a}, \mathcal{V}_\mathrm{c})$, $u \in \mathcal{V}_\mathrm{m}$ and $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $S \subseteq \mathcal{V}_\mathrm{m} \setminus \{u\}$.

Output: a pair of assignments $\widehat{X}, \widehat{Y} \colon \mathcal{V}_\mathrm{set} \to \{\mathsf{F}, \mathsf{T}\}$ for some set of variables $\mathcal{V}_\mathrm{set}$ such that:

- $S \cup \{u\} \subseteq \mathcal{V}_\mathrm{set} \subseteq S \cup \{u\} \cup \mathcal{V}_\mathrm{a}$,
- $\widehat{X}$ and $\widehat{Y}$ agree with Λ on S, $\widehat{X}(u) = \mathsf{T}$ and $\widehat{Y}(u) = \mathsf{F}$.

We fix two total orders $\leq_\mathcal{V}$ and $\leq_\mathcal{C}$ over the variables and clauses of Φ. These are only relevant to have a pre-determined order in which clauses and variables are considered in this algorithm.
Initialise $\widehat{X}$ and $\widehat{Y}$ as Λ, and set $\widehat{X}(u) = \mathsf{T}$ and $\widehat{Y}(u) = \mathsf{F}$.
$\mathcal{V}_\mathrm{set} \leftarrow S \cup \{u\}$, $\mathcal{V}_\mathrm{d} \leftarrow \{u\}$, $\mathcal{F}_\mathrm{d} \leftarrow \{c \in \mathcal{C} : u \in \mathrm{var}(c)\}$, $\mathcal{F}_\mathrm{u} \leftarrow \emptyset$, $\mathcal{C}_\mathrm{rem} \leftarrow \mathcal{C}$.
while $\exists c \in \mathcal{C}_\mathrm{rem} : \mathrm{var}(c) \cap (\mathcal{V}_\mathrm{d} \cup \mathrm{var}(\mathcal{F}_\mathrm{u})) \neq \emptyset$ do
  Let c be smallest clause according to $\leq_\mathcal{C}$ with $\mathrm{var}(c) \cap (\mathcal{V}_\mathrm{d} \cup \mathrm{var}(\mathcal{F}_\mathrm{u})) \neq \emptyset$.
  if c is a bad clause then
    Remove c from $\mathcal{C}_\mathrm{rem}$ and add c to $\mathcal{F}_\mathrm{u}$.
  end if
  if c is a good clause and $(\mathrm{var}(c) \cap \mathcal{V}_\mathrm{a}) \setminus \mathcal{V}_\mathrm{set} = \emptyset$ then
    Remove c from $\mathcal{C}_\mathrm{rem}$ (as all auxiliary variables in c have been set).
    if c is unsatisfied by at least one of $\widehat{X}$ and $\widehat{Y}$ then
      Add c to $\mathcal{F}_\mathrm{u}$.
    end if
  end if
  if c is a good clause and $(\mathrm{var}(c) \cap \mathcal{V}_\mathrm{a}) \setminus \mathcal{V}_\mathrm{set} \neq \emptyset$ then
    Let v be the smallest variable in $(\mathrm{var}(c) \cap \mathcal{V}_\mathrm{a}) \setminus \mathcal{V}_\mathrm{set}$ (according to $\leq_\mathcal{V}$).
    Extend $\widehat{X}$ and $\widehat{Y}$ to v by sampling from the optimal coupling between the marginal distributions of $\mu_{\Omega^{\widehat{X}}}$ and $\mu_{\Omega^{\widehat{Y}}}$ on v, and add v to $\mathcal{V}_\mathrm{set}$.
    if $\widehat{X}(v) \neq \widehat{Y}(v)$ then
      Add v to $\mathcal{V}_\mathrm{d}$. Add all clauses containing v to $\mathcal{F}_\mathrm{d}$.
    end if
  end if
end while
return $(\widehat{X}, \widehat{Y})$.

First, we analyse the sets $\mathcal{V}_\mathrm{set}$, $\mathcal{V}_\mathrm{d}$, $\mathcal{F}_\mathrm{d}$, $\mathcal{F}_\mathrm{u}$ and $\mathcal{C}_\mathrm{rem}$ and prove the connectivity property of $\mathcal{F}_\mathrm{d} \cup \mathcal{F}_\mathrm{u}$. In the rest of this section we fix the inputs of Algorithm 4 unless stated otherwise.

Proposition 39 (Properties of the coupling process). The coupling process in Algorithm 4 terminates eventually and, at the end of the process, the sets $\mathcal{V}_\mathrm{set}$, $\mathcal{V}_\mathrm{d}$, $\mathcal{F}_\mathrm{d}$, $\mathcal{F}_\mathrm{u}$ and $\mathcal{C}_\mathrm{rem}$ present the following properties:

1. We have $S \cup \{u\} \subseteq \mathcal{V}_\mathrm{set} \subseteq \mathcal{V}_\mathrm{a} \cup S \cup \{u\}$, $\mathcal{V}_\mathrm{d} = \{v \in \mathcal{V}_\mathrm{set} : \widehat{X}(v) \neq \widehat{Y}(v)\}$, and $\mathcal{F}_\mathrm{d}$ is the set of clauses containing a variable in $\mathcal{V}_\mathrm{d}$.
2. For all $c \in \mathcal{F}_\mathrm{u}$ we have $\mathrm{var}(c) \cap \mathcal{V}_\mathrm{a} \subseteq \mathcal{V}_\mathrm{set}$ and c is unsatisfied by at least one of $\widehat{X}$ and $\widehat{Y}$.
3. For all $c \in \mathcal{C}_\mathrm{rem}$, we have $\mathrm{var}(c) \cap (\mathcal{V}_\mathrm{d} \cup \mathrm{var}(\mathcal{F}_\mathrm{u})) = \emptyset$.
4. For all $c \in \mathcal{C} \setminus (\mathcal{C}_\mathrm{rem} \cup \mathcal{F}_\mathrm{u})$, we have $\mathrm{var}(c) \cap (\mathcal{V}_\mathrm{d} \cup \mathrm{var}(\mathcal{F}_\mathrm{u})) \neq \emptyset$, $\mathrm{var}(c) \cap \mathcal{V}_\mathrm{a} \subseteq \mathcal{V}_\mathrm{set}$ and c is satisfied by $\widehat{X}$ and $\widehat{Y}$.
5. The set $\mathcal{F}_\mathrm{d} \cup \mathcal{F}_\mathrm{u}$ is connected in G_Φ.

Proof. Each iteration of the coupling procedure either removes a clause from $\mathcal{C}_{\text{rem}}$, or samples the values $\widehat{X}(v)$ and $\widehat{Y}(v)$ for an auxiliary variable v and adds v to $\mathcal{V}_{\text{set}} \subseteq \mathcal{V}$. As $\mathcal{C}_{\text{rem}}$ and $\mathcal{V}$ are finite, the coupling terminates after a finite number of iterations. We prove the five properties in the statement separately. First, we note that the sets $\mathcal{V}_{\text{set}}$, $\mathcal{V}_{\text{d}}$, $\mathcal{F}_{\text{d}}$, $\mathcal{F}_{\text{u}}$ never decrease in size during the execution of Algorithm 4, whereas the set $\mathcal{C}_{\text{rem}}$ never increases in size.

Property 1. Note that at the start of Algorithm 4 (line 3) this property holds. The result then follows from the fact that the sets $\mathcal{V}_{\text{set}}$, $\mathcal{V}_{\text{d}}$ and $\mathcal{F}_{\text{d}}$ are only updated from line 15 to line 20 of Algorithm 4, and these steps preserve Property 1.

Property 2. This follows from the facts that the set $\mathcal{F}_{\text{u}}$ is originally empty, it is only extended in lines 7 and 12, and bad clauses do not contain auxiliary variables.

Property 3. This property follows from the fact that clauses that satisfy $\text{var}(c) \cap (\mathcal{V}_{\text{d}} \cup \text{var}(\mathcal{F}_{\text{u}})) \neq \emptyset$ at some point are eventually removed from $\mathcal{C}_{\text{rem}}$ in either line 7 (if they are bad) or in line 10 (if they are good, once all the auxiliary variables of the clause are in $\mathcal{V}_{\text{set}}$).

Property 4. If $c \in \mathcal{C} \setminus (\mathcal{C}_{\text{rem}} \cup \mathcal{F}_{\text{u}})$, then c has been removed from $\mathcal{C}_{\text{rem}}$ in line 10 but it has not been added to $\mathcal{F}_{\text{u}}$ in line 12, which proves this property.

Property 5. We note that at the start of the coupling process (line 3) $\mathcal{F}_{\text{d}} \cup \mathcal{F}_{\text{u}}$ is connected. Let us analyse every line of the algorithm where the sets $\mathcal{F}_{\text{d}}$ and $\mathcal{F}_{\text{u}}$ are enlarged. When it comes to $\mathcal{F}_{\text{d}}$, this occurs in line 19 if this line is executed. Let c be the clause considered in that iteration of the coupling process and let v be the variable of c considered in line 16. We recall that $\text{var}(c) \cap (\mathcal{V}_{\text{d}} \cup \text{var}(\mathcal{F}_{\text{u}})) \neq \emptyset$ and $v \in (\text{var}(c) \cap \mathcal{V}_{\text{a}}) \setminus \mathcal{V}_{\text{set}}$. In line 19 we add all to $\mathcal{F}_{\text{d}}$ all the clauses containing v. Let C_v be the set of such clauses. Since $\emptyset \neq \text{var}(c) \cap (\mathcal{V}_{\text{d}} \cup \text{var}(\mathcal{F}_{\text{u}})) \subseteq \text{var}(c) \cap \text{var}(\mathcal{F}_{\text{d}} \cup \mathcal{F}_{\text{u}})$ and $c \in C_v$, we conclude that $\mathcal{F}_{\text{d}} \cup \mathcal{F}_{\text{u}} \cup C_v$ is connected as we wanted. When it comes to $\mathcal{F}_{\text{u}}$, we add clauses in lines 7 and 12. In this case, we add a clause c such that $\text{var}(c) \cap (\mathcal{V}_{\text{d}} \cup \text{var}(\mathcal{F}_{\text{u}})) \neq \emptyset$, so $\mathcal{F}_{\text{d}} \cup \mathcal{F}_{\text{u}} \cup \{c\}$ is connected in G_Φ. □

We can now prove our main result concerning the structure of $\Phi^{\widehat{X}}$ and $\Phi^{\widehat{Y}}$.

Lemma 40. Let $\widehat{X}$ and $\widehat{Y}$ be the assignments returned by Algorithm 4 and let $\mathcal{C}_{\text{rem}}$ and $\mathcal{F}_{\text{u}}$ be as in Proposition 39. There are sets of clauses $\mathcal{C}_1 \subseteq \mathcal{C}_{\text{rem}}$ and $\mathcal{C}_2, \mathcal{C}_3 \subseteq \mathcal{F}_{\text{u}}$ such that $\Phi^{\widehat{X}} = (\mathcal{V} \setminus \mathcal{V}_{\text{set}}, \mathcal{C}_1 \cup \mathcal{C}_2)$ and $\Phi^{\widehat{Y}} = (\mathcal{V} \setminus \mathcal{V}_{\text{set}}, \mathcal{C}_1 \cup \mathcal{C}_3)$, where the variables in $\mathcal{V}_{\text{set}}$ are removed from the clauses in $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$.

Proof. We determine the set of clauses that are unsatisfied by $\widehat{X}$ or $\widehat{Y}$ with the help of Proposition 39. We distinguish 3 disjoint cases:

- $c \in \mathcal{C}_{\text{rem}}$. Then $\text{var}(c) \cap \mathcal{V}_{\text{d}} = \emptyset$, so $\widehat{X}$ and $\widehat{Y}$ agree in all the variables in $\text{var}(\mathcal{C}_{\text{rem}}) \cap \mathcal{V}_{\text{set}}$. As a consequence, the restrictions of $\Phi^{\widehat{X}}$ and $\Phi^{\widehat{Y}}$ to $\mathcal{C}_{\text{rem}}$ give rise to the same CNF formula. Note that some of the clauses in $\mathcal{C}_{\text{rem}}$ might be satisfied by both $\widehat{X}$ and $\widehat{Y}$, but they are never satisfied by only one of the two assignments.

- $c \in \mathcal{F}_{\text{u}}$. Then c is unsatisfied by at least one of $\widehat{X}$ and $\widehat{Y}$ and, thus, it appears in at least one of $\Phi^{\widehat{X}}$ and $\Phi^{\widehat{Y}}$. The clause c may contain a variable $v \in \mathcal{V}_{\text{d}}$.

- $c \in \mathcal{C} \setminus (\mathcal{C}_{\text{rem}} \cup \mathcal{F}_{\text{u}})$. By Proposition 39, we have $\text{var}(c) \cap (\mathcal{V}_{\text{d}} \cup \text{var}(\mathcal{F}_{\text{u}})) \neq \emptyset$ and $\text{var}(c) \cap \mathcal{V}_{\text{a}} \subseteq \mathcal{V}_{\text{set}}$. Since $c \notin \mathcal{F}_{\text{u}}$, it follows that c is satisfied by both $\widehat{X}$ and $\widehat{Y}$ and, thus, c does not appear in any of the formulae $\Phi^{\widehat{X}}$ and $\Phi^{\widehat{Y}}$.

We conclude that we can write $\mathcal{C}^{\widehat{X}} = \mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}^{\widehat{Y}} = \mathcal{C}_1 \cup \mathcal{C}_3$, where $\mathcal{C}_1 \subseteq \mathcal{C}_{\text{rem}}$ and $\mathcal{C}_2, \mathcal{C}_3 \subseteq \mathcal{F}_{\text{u}}$ as we wanted. □

In order to further analyse the probability distribution of the output of Algorithm 4, we introduce the following definition.

Definition 41 (run, $\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$, $\tau_{\mathcal{R}}(\Phi,\mathcal{M},u,\Lambda)$, $\mathcal{V}_{\mathrm{set}}(R)$, $\mathcal{V}_{\mathrm{d}}(R)$, $\mathcal{F}_{\mathrm{u}}(R)$, $\mathcal{F}_{\mathrm{d}}(R)$, $\mathcal{C}_{\mathrm{rem}}(R)$). A run of Algorithm 4 is a sequence of all the random choices $(\widehat{X}(v),\widehat{Y}(v))$ made in line 17 when executing Algorithm 4. Let $\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$ be the set of all possible runs of Algorithm 4 for the inputs $\Phi,\mathcal{M},u,\Lambda$ and let $\tau_{\mathcal{R}}(\Phi,\mathcal{M},u,\Lambda)$ be the probability distribution that Algorithm 4 yields on $\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$. Each run $R\in\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$ determines the output $(\widehat{X},\widehat{Y})$ and the sets $\mathcal{V}_{\mathrm{set}}(R),\mathcal{V}_{\mathrm{d}}(R),\mathcal{F}_{\mathrm{u}}(R),\mathcal{F}_{\mathrm{d}}(R),\mathcal{C}_{\mathrm{rem}}(R)$ that are computed in Algorithm 4.

With the aim of applying Proposition 38, we extend the coupling $(\widehat{X},\widehat{Y})$ to all marked and auxiliary variables.

Definition 42 (The coupling (X,Y)). Let $R\in\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$ and let $(\widehat{X},\widehat{Y})$ be the corresponding output of the run R. Let $\leq_{\mathcal{V}}$ be a total order on the variables of Φ and let $v_1\leq_{\mathcal{V}}v_2\leq_{\mathcal{V}}\cdots\leq_{\mathcal{V}}v_t$ be the variables in $(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}$. We extend the assignments $\widehat{X},\widehat{Y}\colon\mathcal{V}_{\mathrm{set}}\to\{\mathsf{F},\mathsf{T}\}$ to $v_1,v_2,\ldots,v_t$ inductively (as follows) to obtain a coupling (X,Y) such that X follows the distribution $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{T}}|_{(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}}}$ and Y follows the distribution $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{F}}|_{(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}}}$. Assume that X and Y are defined on $\mathcal{V}_{\mathrm{set}}\cup\{v_1,v_2,\ldots,v_{j-1}\}$ for $j\in\{1,2,\ldots,t\}$. Then we sample $(X(v_j),Y(v_j))$ from the optimal/monotone coupling of the marginal distributions (on v_j) of μ_{Ω^X} and μ_{Ω^Y}.

Remark 43. When $R\in\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$ follows the probability distribution $\tau_{\mathcal{R}}(\Phi,\mathcal{M},u,\Lambda)$ (Definition 41), the pair of random assignments (X,Y) of Definition 42 is a coupling of the distributions $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{T}}|_{\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}}}}$ and $\mu_{\Omega^{\Lambda\cup u\mapsto\mathsf{F}}|_{\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}}}}$.

In Lemma 44 we bound the probabilities $\Pr(X(v)\neq Y(v)|R)$ for any $R\in\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$ and $v\in(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}(R)$.

Lemma 44. Let $R\in\mathcal{R}(\Phi,\mathcal{M},u,\Lambda)$. Let (X,Y) be the coupling of Definition 42. Then for any $v\in(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}(R)$ we have $\Pr(X(v)\neq Y(v)|R)\leq 2^{-r_0k+1}/k$.

Proof. Let $\widehat{X}$ and $\widehat{Y}$ be the output of Algorithm 4 for the run R. Let $v_1,v_2,\ldots,v_t$ be the variables in $(\mathcal{V}_{\mathrm{m}}\cup\mathcal{V}_{\mathrm{a}})\setminus\mathcal{V}_{\mathrm{set}}(R)$ in the order that they are considered in Definition 42. Let $j\in\{1,2,\ldots,t\}$ and let $\Lambda',\Lambda''\colon\mathcal{V}_{\mathrm{set}}(R)\cup\{v_1,v_2,\ldots,v_{j-1}\}\to\{\mathsf{F},\mathsf{T}\}$ be two assignments such that $\Lambda'|_{\mathcal{V}_{\mathrm{set}}}=\widehat{X}$ and $\Lambda''|_{\mathcal{V}_{\mathrm{set}}}=\widehat{Y}$. When X agrees with Λ' and Y agrees with Λ'', the values $X(v_j)$ and $Y(v_j)$ are sampled from the optimal/monotone coupling between the marginals on v_j of the distributions $\mu_{\Omega^{\Lambda'}}$ and $\mu_{\Omega^{\Lambda''}}$. Let us denote these marginals by ν_X and ν_Y respectively. Thus, by the coupling lemma (Proposition 37) and Proposition 10 (or Lemma 28) on the marginals of marked and auxiliary variables, we have

$$\begin{aligned}\Pr\left(X(v_j)\neq Y(v_j)|\Lambda',\Lambda''\right)&=d_{\mathrm{TV}}(\nu_X,\nu_Y)=\left|\Pr(X(v_j)=\mathsf{T}|\Lambda')-\Pr(Y(v_j)=\mathsf{T}|\Lambda'')\right|\\&\leq|\Pr(X(v_j)=\mathsf{T}|\Lambda')-1/2|+|1/2-\Pr(Y(v_j)=\mathsf{T}|\Lambda'')|\\&\leq\exp\left(\frac{1}{k2^{r_0k}}\right)-1.\end{aligned}$$

Applying the inequality $e^z\leq 1+2z$ for $z\in(0,1)$, we find that $\Pr\left(X(v_j)\neq Y(v_j)|\Lambda',\Lambda''\right)\leq 2^{-r_0k+1}/k$. Thus, from the arbitrary choice of Λ',Λ'' and the law of total probability we conclude that the bound $\Pr\left(X(v_j)\neq Y(v_j)|R\right)\leq 2^{-r_0k+1}/k$ holds. □

Combining all the results presented up to this stage in the current section allows us relate the sum $\sum_{v\in\mathcal{V}_{\mathrm{m}}\setminus(S\cup\{u\})}\left|\mathcal{I}^{\Lambda}(u\to v)\right|$ to the coupling process over auxiliary variables. In fact, we bound this sum of influences between marked variables by the expected number of failed clauses in the coupling process on auxiliary variables. Recall that here $r=r_0-\delta$.

Lemma 45. There is an integer k_0 such that for any $k \geq k_0$ and any density α with $\alpha \leq 2^{(r_0-\delta)k}/k^3$ the following holds w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let $(\mathcal{V}_\mathrm{m}, \mathcal{V}_\mathrm{a}, \mathcal{V}_\mathrm{c})$ be an $(r_0 - \delta, r_0, r_0, 2r_0)$-marking of Φ, and let $u \in \mathcal{V}_\mathrm{m}$ and $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $S \subseteq \mathcal{V}_\mathrm{m} \setminus \{u\}$. Then for a random run R of the coupling process on the auxiliary variables (Algorithm 4), we have

$$\sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \left| \mathcal{I}^\Lambda(u \to v) \right| \leq 2^{-r_0 k + 1} \mathbb{E}\left[|\mathcal{F}_\mathrm{u}(R)| \right].$$

Proof. Let (X, Y) be the coupling in Definition 42 for a (random) run $R \sim \tau_\mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)$ of Algorithm 4. We are going to show that

$$\Pr(X(v) = Y(v) | R) = 1 \text{ for all } v \in V := (\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}) \setminus (\mathcal{V}_\mathrm{set}(R) \cup \mathrm{var}(\mathcal{F}_\mathrm{u}(R))). \tag{27}$$

Let $\widehat{X}, \widehat{Y} \colon \mathcal{V}_\mathrm{set}(R) \to \{\mathsf{F}, \mathsf{T}\}$ be the output of Algorithm 4 for the run R. By Lemma 40 we conclude that we can write $\mathcal{C}^{\widehat{X}} = \mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}^{\widehat{Y}} = \mathcal{C}_1 \cup \mathcal{C}_3$, where $\mathcal{C}_1 \subseteq \mathcal{C}_\mathrm{rem}(R)$ and $\mathcal{C}_2, \mathcal{C}_3 \subseteq \mathcal{F}_\mathrm{u}(R)$. Thus, the variables in V (see (27) for a definition of V) either appear in a clause in $\mathcal{C}_1$ or they are not present in any of the formulae $\Phi^{\widehat{X}}$ and $\Phi^{\widehat{Y}}$. Moreover, by Proposition 39, we have $\mathrm{var}(c) \cap \mathrm{var}(c') = \emptyset$ for all $c \in \mathcal{C}_\mathrm{rem}(R)$ and $c' \in \mathcal{F}_\mathrm{u}(R)$. We conclude that the distributions $\mu_{\Omega^{\widehat{X}}}\big|_V$ and $\mu_{\Omega^{\widehat{Y}}}\big|_V$ agree – both are the uniform distribution over the satisfying assignments of the CNF formula $(V, \mathcal{C}_1)$. Let $v_1, v_2, \ldots, v_t$ be the variables in V in the order they are considered in the definition of the coupling (X, Y). By induction on $j \in \{1, 2, \ldots, t\}$, the marginals on v_j in Definition 42 are the same when coupling $X(v_j)$ and $Y(v_j)$. Thus, we have $X(v_j) = Y(v_j)$ for all $j \in \{1, 2, \ldots, t\}$.

Since $S \cup \{u\} \subseteq \mathcal{V}_\mathrm{set}(R) \subseteq S \cup \{u\} \cup \mathcal{V}_\mathrm{a}$, we have $\mathcal{V}_\mathrm{m} \setminus V = S \cup \{u\} \cup (\mathcal{V}_\mathrm{m} \cap \mathrm{var}(\mathcal{F}_\mathrm{u}(R)))$. In light of Lemma 44 and (27), we find that

$$\sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \Pr(X(v) \neq Y(v) | R) \leq \sum_{v \in \mathcal{V}_\mathrm{m} \cap \mathrm{var}(\mathcal{F}_\mathrm{u}(R))} \Pr(X(v) \neq Y(v) | R) \leq \frac{2}{k} 2^{-r_0 k} |\mathrm{var}(\mathcal{F}_\mathrm{u}(R))|.$$

From $|\mathrm{var}(\mathcal{F}_\mathrm{u}(R))| \leq k |\mathcal{F}_\mathrm{u}(R)|$ we conclude that

$$\sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \Pr(X(v) \neq Y(v) | R) \leq 2^{-r_0 k + 1} |\mathcal{F}_\mathrm{u}(R)|. \tag{28}$$

In the rest of this proof we are going to aggregate (28) over $R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)$ with the aim of applying Proposition 38. Let (X, Y) be the coupling in Definition 42 for a (random) run $R \sim \tau_\mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)$ of Algorithm 4. We have

$$\begin{aligned}
\sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \Pr(X(v) \neq Y(v)) &= \sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \sum_{R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)} \Pr(R) \Pr(X(v) \neq Y(v) | R) \\
&= \sum_{R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)} \Pr(R) \sum_{v \in \mathcal{V}_\mathrm{m} \setminus (S \cup \{u\})} \Pr(X(v) \neq Y(v) | R) \\
&\leq 2^{-r_0 k + 1} \sum_{R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)} \Pr(R) |\mathcal{F}_\mathrm{u}(R)| \\
&= 2^{-r_0 k + 1} \mathbb{E}\left[|\mathcal{F}_\mathrm{u}(R)| \right].
\end{aligned}$$

Finally, we note that we can indeed apply Proposition 38 to the restriction of X and Y on $\mathcal{V}_\mathrm{m}$ as (X, Y) is a coupling of the distributions $\mu_{\Omega^{\Lambda \cup u \mapsto \mathsf{T}}}\big|_{\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}}$ and $\mu_{\Omega^{\Lambda \cup u \mapsto \mathsf{F}}}\big|_{\mathcal{V}_\mathrm{m} \cup \mathcal{V}_\mathrm{a}}$ (Remark 43). This finishes the proof. □

In the remainder of this section we bound $\mathbb{E}[|\mathcal{F}_{\mathrm{u}}(R)|]$, which would complete our proof of Lemma 14 when combined with Lemma 45. In order to do this we exploit the fact that $\mathcal{F}_{\mathrm{u}}(R) \cup \mathcal{F}_{\mathrm{d}}(R)$ is connected in G_Φ (Proposition 39), the local sparsity properties of random CNF formulae and the properties of the marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$. It is important that the bound on $\mathbb{E}[|\mathcal{F}_{\mathrm{u}}(R)|]$ is $\mathrm{poly}(k)\log n$ in order to conclude fast mixing time of the ρ-uniform-block Glauber dynamics when applying the spectral independence framework. First, we bound the probability that some good clauses are failed in Algorithm 4. At first glance this seems to be a straightforward task thanks to the fact that the marginals of marked and auxiliary variables are close to 1/2 (see Proposition 10). However, for any good clauses c_1 and c_2, the events that $c_1 \in \mathcal{F}_{\mathrm{d}}(R) \cup \mathcal{F}_{\mathrm{u}}(R)$ and $c_2 \in \mathcal{F}_{\mathrm{d}}(R) \cup \mathcal{F}_{\mathrm{u}}(R)$ may not be independent; any value given to the variables in c_1 may affects the marginals of the variables in c_2 and whether these variables are considered by the coupling process or not. However, we show that, as long as c_1 and c_2 do not share good variables, these dependencies are not very strong and we can indeed bound the probability that $c_1, c_2 \in \mathcal{F}_{\mathrm{d}}(R) \cup \mathcal{F}_{\mathrm{u}}(R)$ with a careful probability argument that analyses the coupling process step by step, see Lemma 49. With this in mind, we introduce the following definitions.

Definition 46 ($\mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$, $\mathcal{A}_{\leq t}$). For a positive integer t, we let $\mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$ be the set containing for each $R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)$ a tuple with the first $\min\{t, \mathrm{length}(R)\}$ entries of the sequence R. That is, $\mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$ is the set containing all possible sequences of the first t choices that Algorithm 4 makes in line 17. Note that if $R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)$ has $\mathrm{length}(R) \leq t$, then $R \in \mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$. Each $R_t \in \mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$ determines two partial assignments Λ' and Λ'' of marked and auxiliary variables that correspond to the assignments $\widehat{X}$ and $\widehat{Y}$ after $\mathrm{length}(R_t)$ iterations of line 17 following R_t. Let $\mathcal{A}_{\leq t}$ be the σ-algebra containing all the subsets of $\mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$.

Intuitively, $\mathcal{A}_{\leq t}$ contains all the possible events that may occur in the first t iterations of line 17, which is the only randomised operation in Algorithm 4. When bounding the probability that a clause is failed, we will express this event in terms of events concerning the values that $\widehat{X}$ and $\widehat{Y}$ take on its variables. This motivates Definition 47.

Definition 47 ($D_v(j)$). We define the following events for variable $v \in \mathcal{V}_{\mathrm{a}}$ and a random run $R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)$ of Algorithm 4. Let $D_v(1)$ be the event that $v \in \mathcal{V}_{\mathrm{set}}(R)$ and $\widehat{X}(v) \neq \widehat{Y}(v)$. Let $D_v(2)$ be the event that $v \in \mathcal{V}_{\mathrm{set}}(R)$ and $\widehat{X}(v) = \mathsf{F}$. Let $D_v(3)$ be the event that $v \in \mathcal{V}_{\mathrm{set}}(R)$ and $\widehat{X}(v) = \mathsf{T}$. Let $D_v(4)$ be the event that $v \in \mathcal{V}_{\mathrm{set}}(R)$ and $\widehat{Y}(v) = \mathsf{F}$. Let $D_v(5)$ be the event that $v \in \mathcal{V}_{\mathrm{set}}(R)$ and $\widehat{Y}(v) = \mathsf{T}$.

Finally, in order to study the events $D_v(j)$ for $v \in V$ we will have to reason about the first time that a variable in V is added to $\mathcal{V}_{\mathrm{set}}(R)$, which motivates the following definition.

Definition 48 ($\tau(V)$, $f(V)$). For a set of auxiliary variables V, we let $\tau(V)$ be the random variable that takes the value t if the first time that a variable in V is added to $\mathcal{V}_{\mathrm{set}}(R)$ in Algorithm 4 is the t-th time line 17 is executed, and we denote by $f(V)$ this variable. We set $\tau(V) = \infty$ if $V \cap \mathcal{V}_{\mathrm{set}}(R) = \emptyset$, in which case $f(V)$ is not defined.

We now have all the tools that we need to analyse the coupling process step by step.

Lemma 49. Let $V \subseteq \mathcal{V}_{\mathrm{a}}$ and let $i_v \in \{1, 2, 3, 4, 5\}$ for each $v \in V$. Let $h(1) = 2^{-r_0 k+1}/k$ and $h(i) = \frac{\exp(1/k)}{2}$ for $i \in \{2, 3, 4, 5\}$. Then, we have

$$\Pr_{R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)} \left(\bigcap\nolimits_{v \in V} D_v(i_v) \right) \leq \prod_{v \in V} h(i_v).$$

Proof. We are going to prove, for any positive integer t and $A \in \mathcal{A}_{\leq t}$,

$$\Pr\left(\bigcap\nolimits_{v\in V} D_v(i_v)\middle| A, \tau(V) = t\right) \leq \prod_{v\in V} h(i_v). \tag{29}$$

The lemma will then follow from the arbitrary choice of A and t and the law of total probability.

We carry out the proof of (29) by induction on $M = |V|$. Equation (29) holds when V is empty. Let us assume that (29) holds when $|V| < M$. Let V be a set of auxiliary variables with $M = |V|$ and indexes i_v for all $v \in V$, let t be a positive integer and let $A \in \mathcal{A}_{\leq t}$. To simplify the notation, for each $w \in V$ we define $A_t(w, V) = A \cap [\tau(V) = t] \cap [f(V) = w]$. Then, we have

$$\begin{aligned}\Pr\left(\bigcap\nolimits_{v\in V} D_v(i_v)\middle| A, \tau(V) = t\right) \leq \sum_{w\in V} &\Pr\left(f(V) = w | A, \tau(V) = t\right) \cdot \Pr\left(D_w(i_w)\middle| A_t(w, V)\right) \\ &\cdot \Pr\left(\bigcap\nolimits_{v\in V\setminus\{w\}} D_v(i_v)\middle| A_t(w, V), D_w(i_w)\right).\end{aligned}$$

We note that $\tau(V \setminus \{w\}) > t$ when conditioning on $\tau(V) = t$ and $f(V) = w$. Let $A' = A_t(w, V) \cap D_w(i_w)$. We have

$$\begin{aligned}\Pr\left(\bigcap\nolimits_{v\in V\setminus\{w\}} D_v(i_v)\middle| A'\right) = \sum_{j=t+1}^{\infty} &\Pr\left(\tau(V \setminus \{w\}) = j\middle| A'\right) \\ &\cdot \Pr\left(\bigcap\nolimits_{v\in V\setminus\{w\}} D_v(i_v)\middle| A', \tau(V \setminus \{w\}) = j\right).\end{aligned}$$

By our induction hypothesis for $V \setminus \{w\}$, the condition $\tau(V \setminus \{w\}) = j$ and the event $A' \in \mathcal{A}_{\leq j}$, we find that

$$\Pr\left(\bigcap\nolimits_{v\in V\setminus\{w\}} D_v(i_v)\middle| A'\right) \leq \sum_{j=t+1}^{\infty} \Pr\left(\tau(V \setminus \{w\}) = j\middle| A'\right) \prod_{v\in V\setminus\{w\}} h(i_v) \leq \prod_{v\in V\setminus\{w\}} h(i_v).$$

As a consequence, we obtain

$$\begin{aligned}\Pr\left(\bigcap\nolimits_{v\in V} D_v(i_v)\middle| A, \tau(V) = t\right) \leq \sum_{w\in V} &\Pr\left(f(V) = w | A, \tau(V) = t\right) \cdot \Pr\left(D_w(i_w)\middle| A_t(w, V)\right) \\ &\cdot \prod_{v\in V\setminus\{w\}} h(i_v).\end{aligned}$$

We are going to show that $\Pr(D_w(i_w)|A_t(w, V)) \leq h(i_w)$. Once we have proved this, the proof of (29) is completed by noting that $\sum_{w\in V} \Pr\left(f(V) = w|A, \tau(V) = t\right) = 1$.

Let us now bound $\Pr(D_w(i_w)|A_t(w, V))$. Recall here that $A_t(w, V)$ implies the event $w \in \mathcal{V}_{\text{set}}(R)$. Recall also that $A_t(w, V) \in \mathcal{A}_{\leq t}$, see Definition 46. For each $R_t \in A_t(w, V) \subseteq \mathcal{R}_t(\Phi, \mathcal{M}, u, \Lambda)$, we are going to apply Proposition 10 and the fact that $\widehat{X}(w)$ and $\widehat{Y}(w)$ follow the optimal coupling between two marginal distributions on v of the form $\mu_{\Omega^{\Lambda'}}$ and $\mu_{\Omega^{\Lambda''}}$ for some assignments Λ', Λ'' on some marked and auxiliary variables that are determined by R_t. Here it is important for applying Proposition 10 that the event $A_t(w, V)$ is in $\mathcal{A}_{\leq t}$, so every partial run $R_t \in A_t(w, V)$ only gives information about what has happened in Algorithm 4 before w is added to $\mathcal{V}_{\text{set}}(R)$. Thus, aggregating over all possible runs $R_t \in A_t(w, V)$, we find that

$$\begin{aligned}\max\left\{\Pr\left(\widehat{X}(w) = \mathsf{F}\middle| A_t(w, V)\right), \Pr\left(\widehat{X}(w) = \mathsf{T}\middle| A_t(w, V)\right)\right\} &\leq \frac{1}{2}\exp\left(\frac{1}{k2^{r_0 k}}\right) \\ &\leq \frac{1}{2}\exp\left(\frac{1}{k}\right),\end{aligned} \tag{30}$$

where the probability is over the random run $R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)$. The bound (30) also applies with $\widehat{Y}$ instead of $\widehat{X}$. In particular, we conclude that $\Pr(D_w(j)|A_t(w,V)) \leq \exp(1/k)/2 = h(j)$ for all $j \in \{2,3,4,5\}$. Moreover, using the definition of optimal coupling for two Bernoulli distributions, the probability that $\widehat{X}(w) \neq \widehat{Y}(w)$ can be bounded as

$$\begin{aligned}\Pr\left(\widehat{X}(w) \neq \widehat{Y}(w) \middle| A_t(w,V)\right) &= \left|\Pr\left(\widehat{X}(w) = \mathsf{T} \middle| A_t(w,V)\right) - \Pr\left(\widehat{Y}(w) = \mathsf{T} \middle| A_t(w,V)\right)\right| \\ &\leq \left|\Pr\left(\widehat{X}(w) = \mathsf{T} \middle| A_t(w,V)\right) - 1/2\right| + \left|1/2 - \Pr\left(\widehat{Y}(w) = \mathsf{T} \middle| A_t(w,V)\right)\right| \\ &\leq \exp\left(\frac{1}{k2^{r_0 k}}\right) - 1.\end{aligned}$$

Hence, applying the bound $e^z \leq 1 + 2z$ for $z \in (0,1)$ and the definition of the event $D_{v_j}(1)$, we have $\Pr(D_{v_j}(1)|A_t(w,V)) \leq 2/(k2^{r_0 k}) = h(1)$. This finishes the proof of (29). From the arbitrary choice of A and t and the law of total probability, the statement follows. □

We can now bound the probability that some good clauses are failed with the help of Lemma 49.

Lemma 50. Let Φ, u, Λ be the input of Algorithm 4. Let $c_1, \ldots, c_\ell \in \mathcal{C}_{\text{good}}$ such that the variable u does not appear in any of the clauses in $c_1, \ldots, c_\ell$, and $\text{var}(c_i) \cap \text{var}(c_j) \cap \mathcal{V}_{\text{good}} = \emptyset$ for all $1 \leq i < j \leq \ell$. Then, for $R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)$, we have $\Pr(c_1, \ldots, c_\ell \in \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)) \leq 2^{(-r_0 k + 4)\ell}$.

Proof. Let $c_1, \ldots, c_\ell$ be some good clauses of Φ as in the statement. The hypothesis that u does not appear in any of these clauses is necessary as if $u \in \text{var}(c)$ then $c \in \mathcal{F}_{\text{d}}(R)$ by definition. We consider a random run $R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)$ of Algorithm 4 and let $\widehat{X}, \widehat{Y}$ be the (random) output of Algorithm 4 for the run R. For $j \in \{1, 2, \ldots, \ell\}$, let $F_j(1)$ be the event that there is $v \in \text{var}(c_j) \cap \mathcal{V}_{\text{a}}$ such that $v \in \mathcal{V}_{\text{set}}(R)$ and $\widehat{X}(v) \neq \widehat{Y}(v)$, let $F_j(2)$ be the event that $\text{var}(c_j) \cap \mathcal{V}_{\text{a}} \subseteq \mathcal{V}_{\text{set}}(R)$ and c_j is unsatisfied by $\widehat{X}$, and let $F_j(3)$ be the event that $\text{var}(c_j) \cap \mathcal{V}_{\text{a}} \subseteq \mathcal{V}_{\text{set}}(R)$ and c_j is unsatisfied by $\widehat{Y}$. In light of Proposition 39, we have $[c_1, \ldots, c_\ell \in \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)] = \bigcap_{j=1}^{\ell}(F_j(1) \cup F_j(2) \cup F_j(3))$. We obtain

$$\Pr\left(\bigcap\nolimits_{j=1}^{\ell}(F_j(1) \cup F_j(2) \cup F_j(3))\right) \leq \sum_{(i_1, i_2, \ldots, i_\ell) \in \{1,2,3\}^\ell} \Pr\left(\bigcap\nolimits_{j=1}^{\ell} F_j(i_j)\right). \tag{31}$$

We note that $F_j(1) = \bigcup_{v \in \text{var}(c_j) \cap \mathcal{V}_{\text{a}}} D_v(1)$, see Definition 47. Let $(i_1, i_2, \ldots, i_\ell) \in \{1,2,3\}^\ell$, and let $I_1 = \{j : i_j = 1\}$, $I_2 = \{j : i_j = 2\}$ and $I_3 = \{j : i_j = 3\}$. If the event $\bigcap_{j \in I_1} F_j(i_j)$ holds, then, for each $j \in I_1$ there is a variable $u_j \in \text{var}(c_j) \cap \mathcal{V}_{\text{a}}$ such that $D_{u_j}(1)$ holds. Thus, for the set of tuples $T = \prod_{j \in I_1}(\text{var}(c_j) \cap \mathcal{V}_{\text{a}})$, where $\prod$ here denotes the cartesian product of sets, we have

$$\bigcap_{j \in I_1} F_j(i_j) = \bigcup_{(u_1, u_2, \ldots, u_{|I_1|}) \in T} \bigcap_{j \in I_1} D_{u_j}(1). \tag{32}$$

Now we explain how we bound $\Pr\left(\left(\bigcap_{j \in I_2 \cup I_3} F_j(i_j)\right) \cap \left(\bigcap_{j \in I_1} D_{u_j}(1)\right)\right)$ for a tuple $(u_1, u_2, \ldots, u_{|I_1|}) \in T$. We are going to show that

$$\Pr\left(\left(\bigcap\nolimits_{j \in I_2 \cup I_3} F_j(i_j)\right) \cap \left(\bigcap\nolimits_{j \in I_1} D_{u_j}(1)\right)\right) \leq \left(\frac{\exp(1/k)}{2}\right)^{(k-3) r_0 |I_2 \cup I_3|} \left(\frac{2}{k 2^{r_0 k}}\right)^{|I_1|}. \tag{33}$$

The proof of (33) is not as straightforward as it may seem at first glance due to the dependencies among the events $F_j(i_j)$, $D_{u_j}(1)$. The key idea is re-writing the LHS of (33) as in the statement of

Lemma 49. Indeed we note that for each $j \in I_2$ and for each variable $v \in \mathrm{var}(c_j) \cap \mathcal{V}_\mathrm{a}$, the event $F_j(2)$ implies that there is $i_v \in \{2,3\}$ such that $D_v(i_v)$ holds, concluding $F_j(2) = \bigcap_{v \in \mathrm{var}(c_j)\cap\mathcal{V}_\mathrm{a}} D_v(i_v)$, see Definition 47. Analogously, for each $j \in I_3$ and for each variable $v \in \mathrm{var}(c_j) \cap \mathcal{V}_\mathrm{a}$, we find $i_v \in \{4,5\}$ such that $F_j(3) = \bigcap_{v \in \mathrm{var}(c_j)\cap\mathcal{V}_\mathrm{a}} D_v(i_v)$. Therefore, we have

$$\left(\bigcap\nolimits_{j\in I_2\cup I_3} F_j(i_j)\right) \cap \left(\bigcap\nolimits_{j\in I_1} D_{u_j}(1)\right) = \bigcap_{v\in V_f} D_v(i_v), \tag{34}$$

where V_f contains exactly all the auxiliary variables in the clauses c_j with $j \in I_2 \cup I_3$ and the variables $u_1, u_2, \ldots, u_{|I_1|}$. Recall now that each good clause contains at least $r_0(k-3)$ auxiliary variables, and, thus, the bound given in (33) follows from (34) and Lemma 49. Combining (33), (32) and (31), and counting the number of tuples in T, we conclude that

$$\begin{aligned}
\Pr\left(\bigcap\nolimits_{j=1}^{\ell}(F_j(1)\cup F_j(2)\cup F_j(3))\right) &\le \sum_{(i_1,i_2,\ldots,i_\ell)\in\{1,2,3\}^\ell} k^{|I_1|}\left(\frac{\exp(1/k)}{2}\right)^{(k-3)r_0|I_2\cup I_3|}\left(\frac{2}{k2^{r_0k}}\right)^{|I_1|} \\
&\le \sum_{(i_1,i_2,\ldots,i_\ell)\in\{1,2,3\}^\ell} \left(\frac{e2^{3r_0}}{2^{kr_0}}\right)^{|I_2\cup I_3|}\left(\frac{2}{2^{r_0k}}\right)^{|I_1|} \\
&= \left(\frac{e2^{3r_0}}{2^{kr_0}} + \frac{e2^{3r_0}}{2^{kr_0}} + \frac{2}{2^{r_0k}}\right)^{\ell},
\end{aligned}$$

where we used the multinomial theorem. The result now follows from $2e2^{3r_0} + 2 \le 2^4$. □

Following [24] and motivated by Lemma 50, we introduce the combinatorial structure that we use in our proof of Lemma 14 to bound the expected number of failed clauses.

Definition 51 ($G^{\le k}$, $\mathcal{D}_3(G_\Phi, c, \ell)$). For a graph $G = (V, E)$ and a positive integer k, let $G^{\le k}$ be the graph with vertex set V in which vertices u and v are connected if and only if there is a path from u to v in G of length at most k. Given the graph G_Φ, a clause c and a positive integer ℓ, let $\mathcal{D}_3(G_\Phi, c, \ell)$ be the set of subsets $T \subseteq V(G_\Phi)$ such that the following holds:

1. $|T| = \ell$ and $c \in T$;
2. for any $c_1, c_2 \in T$, $\mathrm{var}(c_1) \cap \mathrm{var}(c_2) \cap \mathcal{V}_\mathrm{good} = \emptyset$;
3. the graph $G_\Phi^{\le 3}[T]$, which is the subgraph of $G_\Phi^{\le 3}$ induced by T, is connected;
4. we have $|T \cap \mathcal{C}_{good}| \ge (1 - 8/k)\ell$.

In [24] the authors consider connected sets in $G_\Phi^{\le 4}$ instead of $G_\Phi^{\le 3}$. Here we manage to perform our union bound on $\mathcal{D}_3(G_\Phi, c, \ell)$ thanks to the fact that the set of failed clauses is connected in our refinement of the coupling process.

Lemma 52 ([24, Corollary 8.19] for $G^{\le 3}$). Let $G = (V, E)$ be a connected graph, let $v \in V$ and let ℓ be a positive integer. Let $n_{G,\ell}(v)$ denote the number of connected induced subgraphs of G with size ℓ containing v. Then, for $\ell' = \min\{3\ell, |V|\}$, we have $n_{G^{\le 3},\ell}(v) \le 2^{\ell'} n_{G,\ell'}(v)$.

Proof. Let T be a connected subgraph of $G^{\le 3}$ with size ℓ containing v. We claim that, for all positive ℓ, we can find a connected subset H of G with size $\ell' = \min\{3\ell, |V|\}$ containing T. The proof is straightforward by induction on ℓ, see [24, Lemma 8.18] for the analogous result on $G^{\le 4}$. We note that there are at most $\binom{\ell'}{\ell-1} \le 2^{\ell'}$ subsets T of H containing v that could be mapped to H by the previous construction. Hence, we conclude that $n_{G^{\le 3},\ell}(v) \le 2^{\ell'} n_{G,\ell'}(v)$ as we wanted. □

Lemma 53 ([24, Lemma 7.9] for $\mathcal{D}_3(G_\Phi, c, \ell)$). Let ℓ be an integer which is at least $\log n$. W.h.p. over the choice of Φ, every clause $c \in \mathcal{C}_{\text{good}}$ has the property that the size of $\mathcal{D}_3(G_\Phi, c, \ell)$ is at most $(18k^2\alpha)^{3\ell}$.

Proof. This follows from bounding the number of connected sets of size ℓ in $G_\Phi^{\leq 3}$ that contain c by combining Lemmas 34 and 52. □

We have now all the tools that we need to bound the expected number of failed clauses in the coupling process given in Algorithm 4 and complete the proof of Lemma 14.

Lemma 14. There is an integer $k_0 \geq 3$ such that for any integer $k \geq k_0$ and any density α with $\alpha \leq 2^{r_0 k/3}/k^3$ the following holds. W.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for any $(r_0 - \delta, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ of Φ, the distribution $\mu_\Omega|_{\mathcal{V}_{\text{m}}}$ is $(2^{-(r_0-\delta)k} \log n)$-spectrally independent.

Proof. Let $u \in \mathcal{V}_{\text{m}}$ and $\Lambda \colon S \to \{\mathsf{F}, \mathsf{T}\}$ with $S \subseteq \mathcal{V}_{\text{m}} \setminus \{u\}$. First of all, we apply Lemma 45 to bound $\sum_{v \in \mathcal{V}_{\text{m}} \setminus (S \cup \{u\})} \left|\mathcal{I}^\Lambda(u \to v)\right|$ by $2^{-r_0 k+1} \mathbb{E}\left[|\mathcal{F}_{\text{u}}(R)|\right]$, where $R \sim \tau_{\mathcal{R}}(\Phi, \mathcal{M}, u, \Lambda)$. In the rest of this proof we show that $\Pr(|\mathcal{F}_{\text{u}}(R)| \geq 2k^4 \log n) \leq O(1/n)$ and, thus, for large enough n, $\mathbb{E}\left[|\mathcal{F}_{\text{u}}(R)|\right] = \sum_{R \in \mathcal{R}(\Phi, \mathcal{M}, u, \Lambda)} \Pr(R) |\mathcal{F}_{\text{u}}(R)| \leq 4k^4 \log n$. Putting all this together, and using the fact that $8k^4 \leq 2^{\delta k}$ for large enough k (here $\delta = 0.00001$) we would obtain the bound $\sum_{v \in \mathcal{V}_{\text{m}} \setminus (S \cup \{u\})} \left|\mathcal{I}^\Lambda(u \to v)\right| \leq 8 \cdot 2^{-r_0 k} k^4 \log n \leq 2^{-(r_0-\delta)k} \log n$ and, thus, the result would follow.

So to finish we just need to show that, w.h.p. over the choice of Φ, $\Pr(|\mathcal{F}_{\text{u}}(R)| \geq 2k^4 \log n) \leq O(1/n)$. Let $L = \lceil 2k^4 \log n \rceil$ and let $\ell = \lceil 0.5 k^4 \log n \rceil$. First, we are going to show that, w.h.p. over the choice of Φ, the following holds:

$$\text{if } Z \subseteq \mathcal{C} \text{ is connected and } |Z| = L, \text{ then } \exists c \in Z \cap \mathcal{C}_{\text{good}} \text{ and } T \in \mathcal{D}_3(G_\Phi, c, \ell) \text{ with } T \subseteq Z. \tag{35}$$

In order to prove (35), we are going to find a large independent set of $Z \cap \mathcal{C}_{\text{good}}$, and we are going to extend it with some clauses in $Z \cap \mathcal{C}_{\text{bad}}$ to obtain $T \in \mathcal{D}_3(G_\Phi, c, \ell)$. We need three results that hold w.h.p. over the choice of Φ: Lemmas 20, 32 and 31. We note that we can apply Lemma 20 for $r = r_0 - \delta$ as our density satisfies $\alpha \leq 2^{r_0 k/3}/k^3 \leq \lceil 2^{(r_0-\delta)k} \rceil / k^3 = \Delta_r / k^3$, where $\delta = 0.00001$. For Z as in (35) we have $|Z| \geq 2k^4 \log n$, so by Lemma 32 with $a = 2k^4$, we find that $|\text{var}(Z)| \geq 2k^4 \log n$ and, thus, in light of Lemma 20, we conclude that $|Z \cap \mathcal{C}_{\text{good}}| \geq (1 - 1/k)|Z|$ and $|Z \cap \mathcal{C}_{\text{bad}}| \leq |Z|/k$. From Lemma 31 with $b = 4k^4$, w.h.p. over the choice of Φ, all connected sets of clauses with size at most $4k^4 \log n$ have tree-excess at most $t := \max\{1, 8k^4 \log(ek^2\alpha)\}$. Thus, we can find $U \subseteq Z \cap \mathcal{C}_{\text{good}}$ such that U is a forest (disjoint union of trees) and $|U| \geq (1 - 1/k)|Z| - t$. In particular, U is bipartite, so there is $I \subseteq U$ such that $\text{var}(c) \cap \text{var}(c') = \emptyset$ for all $c, c' \in I$ and $|I| \geq |U|/2 \geq (1 - 1/k)L/2 - t/2 \geq \frac{1}{2} k^4 \log n$, where the last inequality holds for large enough n. Let I' be an independent set of $Z \cap \mathcal{C}_{\text{good}}$ with the largest possible size. Then we have shown that $|I'| \geq \ell = \lceil \frac{1}{2} k^4 \log n \rceil$.

We claim that the set $T' := I' \cup (Z \cap \mathcal{C}_{\text{bad}})$ is connected in $(G_\Phi[Z])^{\leq 3}$, where $G_\Phi[Z]$ is the subgraph of G_Φ induced by Z. Assume for contradiction that T' is not connected in $(G_\Phi[Z])^{\leq 3}$. In this case, we can write $T' = S_1 \cup S_2$ such that for all $c_1 \in S_1$ and $c_2 \in S_2$, the shortest path between c_1 and c_2 through clauses in Z has length at least 4. Let $(c_1, c_2) \in S_1 \times S_2$ be the pair with the shortest path in Z, and let this path be $c_1 = e_1, e_2, \ldots, e_j = c_2$. Then $j \geq 5$ and $e_2, \ldots, e_{j-1} \in Z \setminus T'$. Moreover, we find that $\text{var}(e_3) \cap \text{var}(c) = \emptyset$ for all $c \in T'$ as otherwise $e_1, e_2, \ldots, e_j$ would not be the shortest path between S_1 and S_2. Moreover, since T' contain all bad clauses in Z, we conclude that e_3 is a good clause. It follows that $I' \cup \{e_3\}$ is an independent set of good clauses of Z, which contradicts the fact that I' has largest possible size among such sets.

Finally, as $|T'| \geq \ell$, we can find a good clause c and a subset T of T' with size ℓ such that $c \in T$, T is connected in $G_\Phi^{\leq 3}$ and $|T \cap \mathcal{C}_{\text{bad}}| \leq |Z \cap \mathcal{C}_{\text{bad}}| \leq L/k \leq 8\ell/k$. We conclude that $T \in \mathcal{D}_3(G_\Phi, c, \ell)$. This finishes the proof of (35).

In the rest of the proof we use (35) to bound $\Pr(|\mathcal{F}_{\text{u}}(R)| \geq L)$. Recall that the set of failed clauses $\mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)$ is connected (Proposition 39). If $|\mathcal{F}_{\text{u}}(R)| \geq L$, then there is $Z \subseteq \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)$ with $|Z| = L$ such that Z is connected in G_Φ, and, thus, we can find c and T as in (35). We have shown that the event $|\mathcal{F}_{\text{u}}(R)| \geq L$ is contained in the event that there is a good clause c and $T \in \mathcal{D}_3(\Phi, c, \ell)$ such that $T \subseteq \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)$. As a consequence, we have

$$\begin{aligned}\Pr\left[|\mathcal{F}_{\text{u}}(R)| \geq L\right] &\leq \sum_{c \in \mathcal{C}} \sum_{T \in \mathcal{D}_3(\Phi, c, \ell)} \Pr\left[T \subseteq \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)\right] \\ &\leq \sum_{c \in \mathcal{C}} \sum_{T \in \mathcal{D}_3(\Phi, c, \ell)} \Pr\left[T \cap \mathcal{C}_{\text{good}} \subseteq \mathcal{F}_{\text{d}}(R) \cup \mathcal{F}_{\text{u}}(R)\right].\end{aligned}$$

We note that for any $T \in \mathcal{D}_3(\Phi, c, \ell)$ there is at most one good clause c' that contains the marked variable u. Thus, by definition of $\mathcal{D}_3(\Phi, c, \ell)$, there are at least $(1 - 8/k)\ell - 1$ good clauses in T that do not contain the variable u. Hence, we can apply Lemma 53 on the size of $\mathcal{D}_3(\Phi, c, \ell)$ and Lemma 50 on the probability of good clauses (that do not share good variables) failing to further obtain

$$\Pr\left[|\mathcal{F}_{\text{u}}(R)| \geq L\right] \leq m \left(18k^2\alpha\right)^{3\ell} 2^{-(r_0 k - 4)[(1-8/k)\ell - 1]}.$$

In what follows it is essential that $\alpha \leq 2^{r_0 k/3}/k^3$, and this is the only proof in this paper where we need this bound on the density – other proofs only require the less restrictive bounds $\alpha \leq 2^{(r_0-\delta)k}/k^3$ or $\alpha \leq 2^{(r_0-3\delta)k}/k^3$. Thus, we conclude that

$$\Pr\left[|\mathcal{F}_{\text{u}}(R)| \geq L\right] \leq m \left(18\frac{2^{r_0 k/3}}{k}\right)^{3\ell} 2^{-(r_0 k - 4)(1-8/k)\ell}\, 2^{r_0 k - 4} = m \left(\frac{18^3}{k^3} 2^{8r_0 + 4(1-8/k)}\right)^{\ell} 2^{r_0 k - 4}.$$

Finally, for large enough k we find that $\Pr\left[|\mathcal{F}_{\text{u}}(R)| \geq L\right] \leq me^{-\ell} 2^{r_0 k} \leq mn^{-0.5k^4} 2^{r_0 k} = O(1/n)$ as we wanted. □

8.3 Mixing time of the ρ-uniform-block Glauber dynamics

Finally, we combine the results in this section with Lemma 13 to complete the proof of Lemma 15.

Remark 54. The distribution $\mu_\Omega|_{\mathcal{V}_{\text{m}}}$ on assignments of the marked variables (Definition 11) is b-marginally bounded for $b = 1 - (1/2)\exp(1/k)$ by Proposition 10 (or, equivalently, Lemmas 26 and 28). Since $\exp(1/k) \leq 1 + 2/k$, we have $b \geq 1/2 - 1/k \geq 1/e$ for $k \geq 8$.

Lemma 15. There is a function $k_0(\theta) = \Theta(\log(1/\theta))$ such that, for any $\theta \in (0, 1)$, for any integer $k \geq k_0(\theta)$ and any density α with $\alpha \leq 2^{0.039k}$ the following holds. W.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$, for any $(r_0 - \delta, r_0, r_0, 2r_0)$-marking $(\mathcal{V}_{\text{m}}, \mathcal{V}_{\text{a}}, \mathcal{V}_{\text{c}})$ of Φ and for $\rho = \lceil 2^{-k-1}|\mathcal{V}_{\text{m}}| \rceil$, the ρ-uniform-block Glauber dynamics for updating the marked variables has mixing time $T_{\text{mix}}(\rho, \varepsilon/2) \leq T := \lceil 2^{2k+3} n^\theta \log \frac{2n}{\varepsilon^2} \rceil$.

Proof. In view of Lemma 14, as $\alpha \leq 2^{0.039k} \leq 2^{r_0 k/3}/k^3$ for large enough k, w.h.p. over the choice of Φ, the distribution $\mu_\Omega|_{\mathcal{V}_{\text{m}}}$ is η-spectrally independent for $\eta = 2^{-(r_0-\delta)k} \log n$. Moreover, this distribution is b-marginally bounded for $b = 1/e$ when $k \geq 8$. We are going to apply Lemma 13 with $V = \mathcal{V}_{\text{m}}$, $\mu = \mu_\Omega|_{\mathcal{V}_{\text{m}}}$, $M = |\mathcal{V}_{\text{m}}|$ and $\kappa = 2^{-k-1}$. First, we check that the hypothesis

$M \geq \frac{2}{\kappa}(4\eta/b^2+1)$ of Lemma 13 holds. By Corollary 30 with $r = r_0 - \delta$ and $V = \mathcal{V}_{\mathrm{m}}$, we have $M \geq (r_0-\delta)(k\alpha/\Delta_r)n = \Omega(n)$, so $M \geq \frac{2}{\kappa}(4\eta/b^2+1)$ holds for large enough n as $\frac{2}{\kappa}(4\eta/b^2+1) = O(\log n)$. Hence, we can apply Lemma 13 to obtain

$$T_{\mathrm{mix}}(\rho, \varepsilon/2) \leq \left\lceil C_\rho \frac{M}{\rho}\left(\log\log\frac{1}{\mu_{\min}} + \log\frac{2}{\varepsilon^2}\right)\right\rceil,$$

where $\rho = \lceil \kappa M \rceil$ and $C_\rho = (2/\kappa)^{4\eta/b^2+1}$. We have

$$C_\rho = \exp\left((\log 2)(k+2)\left(\frac{4\eta}{b^2}+1\right)\right) \leq 2^{k+2}\exp\left(\frac{(\log 2)(\log n)(k+2)4e^2}{2^{(r_0-\delta)k}}\right),$$

so there exists a function $k_0(\theta) = \Theta(\log(1/\theta))$ such that when $k \geq k_0(\theta)$, we have $C_\rho \leq 2^{k+2}n^\theta$. In light of Remark 54, we have $\mu_{\min} \geq b^M$, so $\log\log(1/\mu_{\min}) \leq \log(M\log(1/b)) = \log M$ as $b = 1/e$. Thus, we conclude that

$$T_{\mathrm{mix}}(\rho, \varepsilon/2) \leq \left\lceil 2^{2k+3}n^\theta\left(\log M + \log\frac{2}{\varepsilon^2}\right)\right\rceil \leq \left\lceil 2^{2k+3}n^\theta \log\frac{2n}{\varepsilon^2}\right\rceil. \qquad \square$$

9 Proof of Theorem 1

In this section we complete the proof of Theorem 1. The proofs in this section do not present any challenging steps. In fact, they amount to combining the main technical results that have already been proved in this work. We start by showing that the calls to the method Sample in Algorithm 1 are unlikely to ever return error, that is, the connected components of G_{Φ^Λ} have size at most $2k^4(1+\xi)\log(n)$ almost every time the method is called. As pointed out in our proof outline, this is a straightforward consequence of Lemma 17 and the fact that the probability distribution of the output of the Glauber dynamics is $(1/k)$-uniform (Corollary 29).

Lemma 55. Let $\theta \in (0,1)$. There is an integer $k_0 \geq 3$ such that, for any integers $k \geq k_0$, $\xi \geq 1$ and any density $\alpha \leq 2^{(r_0-3\delta)k}/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. In the execution of Algorithm 1 with input Φ, with probability at least $1 - n^{-3\xi}$ over the random choices made by Algorithm 1, every time that the algorithm calls the method Sample(Φ^Λ, S), the connected components of G_{Φ^Λ} have size at most $2k^4(1+\xi)\log(n)$.

Proof. Let $\varepsilon = n^{-\xi}$ and let $T = \lceil 2^{2k+3}n^\theta \log\frac{2n}{\varepsilon^2}\rceil$ be the mixing time established in Lemma 15. Note that $\alpha \leq 2^{(r_0-3\delta)k}/k^3 \leq 2^{(r_0-\delta)k}/k^3$, so we an indeed compute the marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ in Algorithm 1 with the help of Lemma 26. We need $\alpha \leq 2^{(r_0-3\delta)k}/k^3$ so that we can apply Lemma 17 with $r = r_0 - \delta$. Algorithm 1 calls the method Sample exactly $T+1$ times in total: T times in line 7 (when simulating the ρ-uniform-block Glauber dynamics) and one time in line 10 to extend the assignment X_T of marked variables to all variables.

Let $t \in \{0, 1, \ldots, T\}$ and let π_t be the probability distribution of X_t, where X_t is the state of the ρ-block-uniform Glauber dynamics on the marked variables described in Algorithm 1 after t steps. Recall that $\rho = \lceil 2^{-k-1}|\mathcal{V}_{\mathrm{m}}|\rceil$ and that X_0 is chosen uniformly at random. First, we focus on the case $t < T$. We are going to apply Lemma 17 with $r = r_0 - \delta$, $a = 2k^4$, $b = 2a(1+\xi)$, $V = \mathcal{V}_{\mathrm{m}}$, $\mu = \pi_t$ and this choice of ρ. The set $\mathcal{V}_{\mathrm{m}}$ is r_0-distributed by the definition of $(r_0-\delta, r_0, r_0, 2r_0)$-marking (Definition 8). Moreover, π_t is $(1/k)$-uniform by Corollary 29, and we have $\rho \leq |\mathcal{V}_{\mathrm{m}}|/2^k$. Hence, we can indeed apply Lemma 17. Consider the following experiment described in Lemma 17 for $L = \lceil a(1+\xi)\log n\rceil$, which satisfies $a\log n \leq L \leq b\log n$. First, draw $S \subseteq \mathcal{V}_{\mathrm{m}}$ from the uniform

distribution τ over subsets of $\mathcal{V}_{\mathrm{m}}$ with size ρ. Then, sample an assignment Λ_{t+1} from $\pi_t|_{\mathcal{V}_{\mathrm{m}}\setminus S}$, the marginal of π_t on $\mathcal{V}_{\mathrm{m}} \setminus S$. Denote by $\mathcal{F}$ the event that that there is a connected set of clauses Y of Φ with $|Y| \geq L$ such that all clauses in Y are unsatisfied by Λ_{t+1}. Then we have

$$\Pr_{S\sim\tau}\left(\Pr_{\Lambda_{t+1}\sim\pi_t|_{\mathcal{V}_{\mathrm{m}}\setminus S}}(\mathcal{F}) \leq 2^{-\delta kL}\right) \geq 1 - 2^{-\delta kL}. \tag{36}$$

Alternatively, this experiment is the same as first sampling an assignment X_t of all variables in $\mathcal{V}_{\mathrm{m}}$ from π_t, and then restricting the assignment to a random set $S \sim \tau$, obtaining Λ_{t+1}. Note that this exact experiment occurs before calling the method Sample in the t-th step of the ρ-uniform-block Glauber dynamics in Algorithm 1. Thus, in light of (36), the probability that in step $t+1$ of the execution of Algorithm 1 the graph $G_{\Phi^{\Lambda_{t+1}}}$ has a connected component with size at least L is at most $2^{-\delta kL} + 2^{-\delta kL}$, where the first $2^{-\delta kL}$ comes from the probability of choosing a wrong set $S \sim \tau$ in (36) and the second $2^{-\delta kL}$ comes from the bound on the probability of the event $\mathcal{F}$ once we have chosen S. We have shown that with probability at least $1 - 2{\cdot}2^{-\delta kL}$, all the connected components of the graph $G_{\Phi^{\Lambda_t}}$ appearing in step $t+1$ of the execution of Algorithm 1 have size less than L. We have $2{\cdot}2^{-\delta kL} \leq 2{\cdot}n^{-\delta ka(1+\xi)\log 2} \leq n^{-5\xi}$ for large enough k, so the probability that Sample returns error at step $t+1$ is at most $n^{-5\xi}$. The case $t = T$ is analogous, the only difference here is that we call Sample on Φ^{X_T}, where $X_T \sim \pi_T$ is an assignment of all marked variables, so we apply Lemma 17 with $\rho = 0$ instead of $\rho = \lceil 2^{-k-1}|\mathcal{V}_{\mathrm{m}}|\rceil$.

Finally, we carry out a union bound over $t \in \{0, 1, \ldots, T\}$, so the probability that any of the calls to Sample returns error is at most $(T+1)n^{-5\xi} \leq n^{-3\xi}$ for large enough n as $T = O(n^\theta \log n) = O(n \log n)$. □

Once we have established Lemmas 15, 19, and 55, the proof of Theorem 1 follows as below.

Theorem 1. For any real $\theta \in (0, 1)$, there is $k_0 \geq 3$ with $k_0 = O(\log(1/\theta))$ such that, for any integers $k \geq k_0$ and $\xi \geq 1$, and for any positive real $\alpha \leq 2^{0.039k}$, the following holds.

There is an efficient algorithm to sample from the satisfying assignments of a random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$ within $n^{-\xi}$ total variation distance of the uniform distribution. The algorithm runs in time $O(n^{1+\theta})$, and succeeds w.h.p. over the choice of Φ.

Proof. Let $k_0(\theta) = \Theta(\log(1/\theta))$ be large enough so that, for all integers $k \geq k_0(\theta)$, $\xi \geq 1$ and all densities $\alpha \leq 2^{0.039\cdot k}$, the conclusions of Lemmas 26, 15, 19, and 55 hold w.h.p. over the choice of the random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. These lemmas are enough to analyse Algorithm 1 and tackle this proof. We analyse the distribution μ_{alg} of the output of Algorithm 1. This distribution outputs either a satisfying assignment of the input formula Φ or error. Let $\varepsilon = n^{-\xi}$. Let $\mathcal{E}$ be the event that running Algorithm 1 outputs error. This happens with probability at most $\varepsilon/4$ when computing the marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ in line 2 of the algorithm, and in lines 7 and 10 if the method Sample$(\hat{\Phi}, S)$ returns error, which occurs when $G_{\hat{\Phi}}$ has a connected component with size more than $2k^4(1+\xi)\log(n)$. In view of Lemma 55, the probability that Algorithm 1 outputs error due to the failure of the method Sample is at most $n^{-3\xi} = \varepsilon^3$. We conclude that the probability that the algorithm outputs error is bounded above by $\varepsilon/4 + \varepsilon^3 \leq \varepsilon/2$. Let $\mu_{Glauber}$ be the distribution that Algorithm 1 would output if there were no errors (that is, the distribution assuming that the method Sample always outputs from the appropriate distribution). Then $d_{\mathrm{TV}}(\mu_{alg}, \mu_{Glauber})$ is the probability that an error occurs, which is at most $\varepsilon/2$. Let $\pi_{Glauber}$ be the distribution output by the ρ-uniform-block Glauber dynamics on $\mathcal{V}_{\mathrm{m}}$ after T steps. By Lemma 15 on the mixing time of the Glauber dynamics, we have $d_{\mathrm{TV}}(\pi_{Glauber}, \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}) \leq \varepsilon/2$. As $\mu_{Glauber}$ comes from sampling an assignment X_T from $\pi_{Glauber}$ and then completing X_T to all $\mathcal{V}$ by sampling from $\mu_\Omega(\cdot|X_T)$, we have $d_{\mathrm{TV}}(\mu_{Glauber}, \mu_\Omega) \leq d_{\mathrm{TV}}(\pi_{Glauber}, \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}) \leq \varepsilon/2$. We find that $d_{\mathrm{TV}}(\mu_{alg}, \mu_\Omega) \leq d_{\mathrm{TV}}(\mu_{alg}, \mu_{Glauber}) +$

$d_{\mathrm{TV}}(\mu_{Glauber}, \mu_\Omega) \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$ as we wanted. The running time of Algorithm 1 is now easily obtained by adding up the running times of the following subroutines. The good clauses and good variables are computed in time $O(n + km) = O(n)$, see Proposition 7. The marking $(\mathcal{V}_{\mathrm{m}}, \mathcal{V}_{\mathrm{a}}, \mathcal{V}_{\mathrm{c}})$ is computed with probability at least $1 - \varepsilon/4$ in time $O(n\Delta_r k^2 \log(4/\varepsilon)) = O(n \log n)$, see Lemma 26. Recall that there are $T + 1 = O(n^\theta \log(n/\varepsilon^2)) = O(n^\theta \log n)$ calls to the method Sample(Φ', S), and each call takes time $O(|S| \log n) = O(n \log n)$ by Lemma 19. We conclude that the running time of Algorithm 1 is $O(n^{1+\theta} \log(n)^2)$. The result now follows by choosing $k_1 = k_0(\theta/2)$, so the running time for $k \geq k_1$ is $O(n^{1+\theta/2} \log(n)^2) = O(n^{1+\theta})$. □

We have now proved that it is possible to (approximately) sample uniformly at random from the satisfying assignments of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. At this point, standard techniques can be applied to obtain a randomised approximation scheme for counting the satisfying assignments of Φ. These techniques are based on the self-reducibility of k-SAT [33]. The following remark shows how to obtain a randomised approximation scheme that runs in time $O(n^\theta (n/\varepsilon)^2)$ following [21, Chapter 7], where the authors base their counting algorithm on the simulated annealing method [47, 30, 35].

Remark 56 (Approximate counting for random k-SAT formulae). Let $k_0(\theta)$ be the integer depending on $\theta \in (0, 1)$ obtained in Theorem 1. Let $k_1 = k_0(\theta/2)$, let $k \geq k_1$ be an integer, let ξ be a positive integer and let $\alpha \leq 2^{0.039k}$ be a density. We apply Theorem 1 to obtain an algorithm to sample from the satisfying assignments of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$ within $n^{-4\xi}$ total variation distance from the uniform distribution. This algorithm runs in time $O(n^{1+\theta/2})$ and succeeds w.h.p. over the choice of Φ.

Let $\varepsilon \in (0, 1)$ with $\varepsilon \geq n^{-\xi}$. A modified version of the approximate counting algorithm of [21, Section 7], using $O(\varepsilon^{-2} n \log(n/\varepsilon))$ samples from the sampling algorithm above, approximates the number of satisfying assignments of the k-CNF formula Φ with multiplicative error ε, thus, achieving running time $O(n^{\theta/2}(n/\varepsilon)^2 \log(n/\varepsilon)) = O(n^\theta (n/\varepsilon)^2)$. Here we describe these minor modifications.

Let Ω_{bad} be the set of assignments $X \colon \mathcal{V} \to \{\mathsf{F}, \mathsf{T}\}$ that satisfy the bad clauses of Φ. For $X \in \Omega_{\mathrm{bad}}$, we define $F(X)$ to be the set of good clauses that are not satisfied by X. For $\kappa \in \mathbb{R}$, we define $w_\kappa(X) = \exp(-\kappa |F(X)|)$ and we define the partition function $Z(\kappa) = \sum_{X \in \Omega_{\mathrm{bad}}} w_\kappa(X)$. The simulated annealing algorithm of [21, Section 7] uses $Z(\kappa)$ (with Ω^* from Definition 9 in place of Ω_{bad}) to approximate the number of satisfying assignments of Φ. We note that $Z(0) = |\Omega_{\mathrm{bad}}|$, which can be computed in linear time in n using the exact counting algorithm given in Proposition 36. Here one has to use the fact that the connected components of $G_{\Phi'}$ for the formula $\Phi' = (\mathcal{V}, \mathcal{C}_{\mathrm{bad}})$ have size at most $2k^4 \log n$, see Lemma 69 from Appendix A and Lemma 32, and the fact that these connected component have tree-excess upper bounded as a function of k (Lemma 31). Once one has performed these modifications, the algorithm given in [21, Section 7] applies without any difficulties.

10 Proof of Theorems 3 and 5

In this section we exploit Lemma 17 to prove Theorems 3 and 5 on the connectivity and looseness of the solution space of random k-CNF formulae. We recall that the density threshold in Theorems 3 and 5 is $\alpha \leq 2^{0.227k}$, significantly larger than our algorithmic threshold in Theorem 1, which is $\alpha \leq 2^{0.039k}$. In order to conclude connectivity for densities up to $2^{0.227k}$, we let $r_1 = 0.227092$ and consider the threshold $\Delta_r = \lceil 2^{rk} \rceil$ for $r = r_1 - \delta$ in the definition of high-degree variables instead of $\Delta_{r_0 - \delta} = \lceil 2^{(r_0 - \delta)k} \rceil$. In all this section we set $r = r_1 - \delta$, so we omit r in the notation and we write $\mathcal{V}_{\mathrm{good}}$ instead of $\mathcal{V}_{\mathrm{good}}(r)$. We work with an $(r, r_1, 0, r_1)$-marking $(\mathcal{V}_{\mathrm{m}}, \emptyset, \mathcal{V}_{\mathrm{c}})$ (the set of auxiliary variables is empty), which we can find w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$ as in Lemma 27.

Let us briefly recall some of the properties of this marking. First of all, by definition, the set $\mathcal{V}_{\mathrm{m}}$ is r_1-distributed and is a subset of $\mathcal{V}_{good}$. Moreover, the distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(1/k)$-uniform by Lemma 28. In light of Lemma 17 for $r = r_1 - \delta$, these properties allow us to show that, when sampling $\Lambda \sim \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$, the connected components of Φ^Λ are logarithmic in size with probability $1 - o(1)$ over the choice $\Lambda \sim \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$. In fact, this is also the case when $\Lambda \sim \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}\setminus\{v\}}$ for any variable v.

Corollary 57. There is an integer $k_0 \geq 3$ such that, for any integer $k \geq k_0$, any density $\alpha \leq \alpha_1 := 2^{(r_1-3\delta)k}$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$.

Let V be a set of good variables of Φ that is r_1-distributed, let μ be a $(1/k)$-uniform distribution over the assignments $V \to \{\mathsf{F}, \mathsf{T}\}$ and let $v \in V$. Then, with probability at least $1 - n^{-k}$ over the choice $\Lambda \sim \mu|_{V\setminus\{v\}}$, the connected components of Φ^Λ have size at most $2k^4 \log n$.

Proof. The result is an application of Lemma 17 with $r = r_1 - \delta$, $b = 4k^4$, $\rho = 1$ and $L = \lceil 2k^4 \log n \rceil$. We need large enough k_0 such that $2^{-\delta k L} \leq 2^{-\delta 2k^5 \log n} \leq n^{-k}$ for all $k \geq k_0$. For these parameters, in the setting of Lemma 17, the distribution τ is the uniform distribution over the variables in V. The experiment in the statement of Lemma 17 consists in sampling $v \sim \tau$ and then sampling $\Lambda \sim \mu|_{V\setminus\{v\}}$. Let $\mathcal{F}_v$ be the event, concerning the choice $\Lambda \sim \mu|_{V\setminus\{v\}}$, that there is a connected set of clauses Y of Φ with $|Y| \geq \lceil 2k^4 \log n \rceil$ such that all clauses in Y are unsatisfied by Λ. Then by Lemma 17 we have $\Pr_{v\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L}\right) \geq 1 - 2^{-\delta k L}$. From $2^{-\delta k L} \leq n^{-k}$, we obtain the bound $\Pr_{v\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L}\right) \geq 1 - n^{-k}$. Since τ is the uniform distribution over the variables in V, for $v \sim \tau$, either the event that $\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L}$ has probability 1 or it has probability at most $1 - 1/|V| \leq 1 - 1/n$. The latter option is not possible due to $\Pr_{v\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L}\right) \geq 1 - n^{-k}$ and $k \geq 3$. Thus, we conclude that $\Pr_{v\sim\tau}\left(\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L}\right) = 1$, so for any $v \in V$ we have $\Pr_{\Lambda\sim\mu|_{V\setminus\{v\}}}(\mathcal{F}_v) \leq 2^{-\delta k L} \leq n^{-k}$. That is, for any $v \in V$, with probability at least $1 - n^{-k}$ over the choice of $\Lambda \sim \mu|_{V\setminus\{v\}}$ the connected components of Φ^Λ have size at most $L - 1 = \lceil 2k^4 \log n \rceil - 1 < 2k^4 \log n$ as we wanted to prove. □

Our connectivity and looseness results will follow from Corollary 57. In Section 10.1 we prove Theorem 3 and in Section 10.2 we prove Theorem 5.

10.1 Proof of Theorem 3

We consider Algorithm 5 that receives two satisfying assignments of a k-CNF formula Φ as the input and constructs a path between them. Before introducing this algorithm, recall that the graph H_Φ is the dependency graph of the variables of Φ introduced in Definition 18.

Algorithm 5 Finding a $(\mathrm{poly}(k)\log n)$-path between two satisfying assignments

Input: a k-CNF formula $\Phi = (\mathcal{V}, \mathcal{C})$ with n variables, an $(r, r_1, 0, r_1)$-marking $(\mathcal{V}_\mathrm{m}, \emptyset, \mathcal{V}_\mathrm{c})$ of Φ, and two satisfying assignments σ, σ'.

1: Let $v_1, v_2, \ldots, v_\ell$ be the variables in $\mathcal{V}_\mathrm{m}$.
2: $\zeta_0 \leftarrow \sigma$.
/* Stage 1: Update the marked variables */
3: for $i \in [\ell]$ do
4: Find $\zeta_i \in \Omega$ with marked variables specified by $\zeta_i(v_j) = \begin{cases} \sigma'(v_j), & j \le i; \\ \sigma(v_j), & j > i; \end{cases}$ such that $\|\zeta_i - \zeta_{i-1}\|_1$ is minimised.
5: end for
6: $\xi_0 = \zeta_\ell$
/* Stage 2: Update the rest of variables */
7: Let $\tau' = \sigma'|_{\mathcal{V}_\mathrm{m}}$ and suppose that $H_{\Phi^{\tau'}}$ has connected components $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_t$.
8: for $i \in [t]$ do
9: Let $\xi_i \in \Omega$ be defined as $\xi_i(v) = \begin{cases} \sigma'(v), & v \in \left(\mathcal{V} \setminus \bigcup_{j=1}^{t} \mathcal{E}_j\right) \cup \left(\bigcup_{j=1}^{i} \mathcal{E}_j\right); \\ \zeta_\ell(v), & v \in \bigcup_{j=i+1}^{t} \mathcal{E}_j. \end{cases}$
10: end for
11: return The path $\sigma = \zeta_0 \leftrightarrow \cdots \leftrightarrow \zeta_\ell = \xi_0 \leftrightarrow \cdots \leftrightarrow \xi_r = \sigma'$.

To prove Theorem 3, it suffices to show that the output of Algorithm 5 is with high probability a D-path in the solution space for $D = 2k^5 \log n$ for the inputs $\sigma \sim \mu_\Omega$ and $\sigma' \sim \mu_\Omega$. We will not actually require $\sigma \sim \mu_\Omega$ and $\sigma' \sim \mu_\Omega$ in the proof; instead we will just use the fact that the restrictions of σ and σ' on $\mathcal{V}_\mathrm{m}$ follow a $(1/k)$-uniform distribution as guaranteed by Lemma 28, see the proof of Lemma 59 for details.

We need the following two lemmas to establish Theorem 3. The first lemma (Lemma 58) shows that all the truth assignments ζ_i, ξ_i in the algorithm exist and satisfy the formula (i.e. the algorithm is well-defined), implying our constructed path is indeed a valid path comprising only satisfying assignments. The second lemma (Lemma 59) shows that w.h.p., two adjacent assignments differ by at most $2k^5 \log n$ variables. This result is an application of Corollary 57.

Lemma 58. For any k-CNF formula Φ with n variables, any $(r, r_1, 0, r_1)$-marking $(\mathcal{V}_\mathrm{m}, \emptyset, \mathcal{V}_\mathrm{c})$ of Φ, and any two satisfying assignments σ, σ', Algorithm 5 on these inputs is well-defined in the following sense:

1. It is always possible to implement Line 4 such that $\zeta_i \in \Omega$.
2. We have $\xi_i \in \Omega$ for each $i \in [t]$.

Proof. To prove item 1, we are going to show that for any partial assignment $X \colon \mathcal{V}_\mathrm{m} \to \{\mathsf{F}, \mathsf{T}\}$, we have $\Pr_{\mu_\Omega}(X) > 0$ and, thus, can extend X to some satisfying assignment. If this claim holds, then we can indeed compute the satisfying assignments $\zeta_1, \zeta_2, \ldots, \zeta_\ell$ in Algorithm 5. Recall that the distribution $\mu_\Omega|_{\mathcal{V}_\mathrm{m}}$ is $(1/k)$-uniform, see Lemma 28. From the definition of $(1/k)$-uniform distribution, we find that an analogous statement to Proposition 10 holds for our $(r, r_1, 0, r_1)$-marking (here $r = r_1 - \delta$): for any $v \in \mathcal{V}_{\mathrm{good}}(r)$, any $V \subseteq \mathcal{V}_\mathrm{m}$ with $v \notin V$, and any $\Lambda \colon V \to \{\mathsf{F}, \mathsf{T}\}$, we have

$$\max\left\{\Pr_{\mu_{\Omega^\Lambda}}\left(v \mapsto \mathsf{F} \middle| \Lambda\right), \Pr_{\mu_\Omega}\left(v \mapsto \mathsf{T} \middle| \Lambda\right)\right\} \le \frac{1}{2}\exp\left(\frac{1}{k}\right).$$

Thus, by induction on the size of a set $S \subseteq \mathcal{V}_{\mathrm{m}}$, we conclude that any assignment $\Lambda\colon S \to \{\mathsf{F},\mathsf{T}\}$ has $\Pr_{\mu_\Omega}(\Lambda) > 0$, finishing the proof of item 1.

Next consider item 2. Let $\tau' = \sigma'|_{\mathcal{V}_{\mathrm{m}}}$ as in Algorithm 5. All clauses that do not appear in $G_{\Phi^{\tau'}}$ are satisfied by the partial assignment τ'. Now consider two satisfying assignments Λ, Λ' such that $\Lambda(\mathcal{V}_{\mathrm{m}}) = \Lambda'(\mathcal{V}_{\mathrm{m}}) = \tau'$. Let $G_{\Phi^{\tau'}}$ have connected components $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{t'}$. In particular, $\Lambda|_{\mathrm{var}(\mathcal{C}_i)}$ and $\Lambda'|_{\mathrm{var}(\mathcal{C}_i)}$ each satisfy all clauses in $\mathcal{C}_i$. Each clause in $\Phi^{\tau'}$ is in exactly one connected component $\mathcal{C}_i$. Consequently, any assignment X such that $X|_{\mathcal{V}_{\mathrm{m}}} = \tau'$ and $X|_{\mathrm{var}(\mathcal{C}_i)} \in \{\Lambda|_{\mathrm{var}(\mathcal{C}_i)}, \Lambda'|_{\mathrm{var}(\mathcal{C}_i)}\}$ for all $i \in [t']$ is a satisfying assignment (any variables that do not appear in $\mathcal{V}_{\mathrm{m}} \cup \left(\bigcup_{i=1}^{t'} \mathrm{var}(\mathcal{C}_i)\right)$ can be chosen arbitrarily). We note that there are two types of connected components of H_{Φ^τ}. The first type are those corresponding to $\mathrm{var}(\mathcal{C}_i)$ for some $i \in [t']$. The second type are those connected components with variables in $\mathcal{V} \setminus \left(\mathcal{V}_{\mathrm{m}} \cup \left(\bigcup_{i=1}^{t'} \mathrm{var}(\mathcal{C}_i)\right)\right)$. These connected components are singleton and consist of one variable v that does not appear in Φ^τ or, equivalently, every clause of Φ containing v is satisfied by τ. As a consequence, taking $\Lambda = \zeta_\ell$, $\Lambda' = \sigma'$ and $X = \xi_i$ in the argument above, we conclude that $\xi_0, \xi_1, \ldots, \xi_t$ are satisfying assignments by construction in Algorithm 5 and item 2 holds. □

Lemma 59. There is an integer $k_0 \geq 3$ such that, for any integer $k \geq k_0$, any density $\alpha \leq 2^{(r_1 - 3\delta)k}$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. In Algorithm 5 with inputs the formula Φ, an $(r, r_1, 0, r_1)$-marking of Φ and the two satisfying assignments σ and σ', with probability at least $1 - 1/n$ over the choices $\sigma \sim \mu_\Omega, \sigma' \sim \mu_\Omega$, we have

1. $\|\zeta_i - \zeta_{i-1}\|_1 \leq 2k^5 \log n$ for all $i \in [\ell]$;
2. $\|\xi_i - \xi_{i-1}\|_1 \leq 2k^5 \log n$ for all $i \in [t]$.

Proof. Let Φ and $(\mathcal{V}_{\mathrm{m}}, \emptyset, \mathcal{V}_{\mathrm{a}})$ be the first two inputs of Algorithm 5, and let $v_1, v_2, \ldots, v_\ell$ be the variables in $\mathcal{V}_{\mathrm{m}}$ in the order considered in Algorithm 5. Let $\sigma \sim \mu_\Omega$ and $\sigma' \sim \mu_\Omega$. Let $\sigma = \zeta_0 \leftrightarrow \cdots \leftrightarrow \zeta_\ell = \xi_0 \leftrightarrow \cdots \leftrightarrow \xi_r = \sigma'$ be the path between σ and σ' output by Algorithm 5. In light of Lemma 58, the assignments $\zeta_0, \zeta_1, \ldots, \zeta_\ell, \xi_1, \ldots, \xi_r$ are satisfying assignments of Φ. We also note that the set of marked variables $\mathcal{V}_{\mathrm{m}}$ is r_1-distributed and does not contain bad variables by Definition 8. We are going to apply Corollary 57 with $V = \mathcal{V}_{\mathrm{m}}$ several times in this proof. In view of Lemma 28, the distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(1/k)$-uniform, and this will be relevant when applying Corollary 57. We prove that Item 1 holds with probability at least $1 - 1/(2n)$ and that Item 2 holds with probability $1 - 1/(2n)$, so the result follows from a union bound.

Item 1. Let $i \in [\ell]$ and let τ_i be the restriction of ζ_i to $\mathcal{V}_{\mathrm{m}}$. By construction, τ_i agrees with σ' on $v_1, v_2, \ldots, v_i$ and it agrees with σ on $v_{i+1}, v_{i+2}, \ldots, v_\ell$. Let Λ_i denote the restriction of τ_i on $\mathcal{V}_{\mathrm{m}} \setminus \{v_i\}$, which agrees with ζ_i and ζ_{i-1} on $\mathcal{V}_{\mathrm{m}} \setminus \{v_i\}$. Recall that, by definition, ζ_i is the satisfying assignment that extends τ_i that minimises $\|\zeta_i - \zeta_{i-1}\|_1$, see Algorithm 5. We consider the connected components of $G_{\Phi^{\Lambda_i}}$, which can be seen as CNF formulae with variables in $\mathcal{V}_{\mathrm{c}} \cup \{v_i\}$ due to the fact that all marked variables other than v_i are set by Λ_i. Each one of these connected components are satisfied as CNF formulae by the assignments ζ_i and ζ_{i-1}. We conclude that ζ_i and ζ_{i-1} agree on the variables of all these connected components except for those variables in the connected component of the clauses containing v_i, where ζ_i and ζ_{i-1} may disagree. Let us denote this connected component by C_{v_i}, which is empty when all the clauses containing v_i are satisfied by Λ_i. We have $\|\zeta_i - \zeta_{i-1}\|_1 \leq k|C_{v_i}|$, where the factor k comes from the fact that each clause has at most k variables. We now bound the size of C_{v_i}. Since the restrictions of σ and σ' to $\mathcal{V}_{\mathrm{m}}$ follow $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$, which is $(1/k)$-uniform, we find, by Definition 12, that τ_i also follows an $(1/k)$-uniform distribution over the assignments $\mathcal{V}_{\mathrm{m}} \to \{\mathsf{F},\mathsf{T}\}$. Let us denote this distribution by μ_i. Then $\Lambda_i \sim \mu_i|_{\mathcal{V}_{\mathrm{m}} \setminus \{v_i\}}$

and, by Corollary 57 with $V = \mathcal{V}_{\mathrm{m}}$, $\Lambda = \Lambda_i$ and $\mu = \mu_i$, with probability at least $1 - n^{-k}$ over the choice $\Lambda_i \sim \mu_i|_{\mathcal{V}_{\mathrm{m}} \setminus \{v_i\}}$, the connected component $C_{v_i} \subset G_{\Phi^{\Lambda_i}}$ containing v_i has at most $2k^4 \log n$ clauses. Thus, with probability at least $1 - n^{-k}$, we have $\|\zeta_i - \zeta_{i-1}\|_1 \le k|C_{v_i}| \le 2k^5 \log n$. By a union bound over $i \in [\ell]$ and the fact that $k \ge 3$ and $\ell \le n$, we conclude that, with probability at least $1 - 1/n^2$, we have $\|\zeta_i - \zeta_{i-1}\|_1 \le 2k^5 \log n$ for all $i \in [\ell]$.

Item 2. Let $\tau' = \sigma'|_{\mathcal{V}_{\mathrm{m}}}$ as in Algorithm 5. By construction, $\xi_0 = \zeta_\ell$ and $\xi_t = \sigma'$ agree with τ' on $\mathcal{V}_{\mathrm{m}}$. Since $\sigma' \sim \mu_\Omega$, we have $\tau' \sim \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$, which is $(1/k)$-uniform by Lemma 28. In view of Corollary 57 for $V = \mathcal{V}_{\mathrm{m}}$, $\Lambda = \tau'$ and $\mu = \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$, with probability at least $1 - n^{-k}$, all of the connected components of $G_{\Phi^{\tau'}}$, have size at most $2k^4 \log n$. Thus, all the connected components of $H_{\Phi^{\tau'}}$ have size at most $2k^5 \log n$. By construction, see Line 9 in Algorithm 5, the assignments ξ_{i-1} and ξ_i agree on the variables in all the connected components of $H_{\Phi^{\tau'}}$ except for the variables in the i-th connected component, where they may disagree. Thus, they disagree on at most $2k^5 \log n$ variables. This gives the desired result. □

We can now complete the proof of Theorem 3.

Theorem 3. There is $k_0 \ge 3$ and a polynomial $p(k)$ with non-negative integer coefficients such that, for any integer $k \ge k_0$, and for any positive real $\alpha \le 2^{0.227k}$, the following claim holds with high probability over the choice of a random k-CNF formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Two satisfying assignments chosen uniformly at random are $p(k)\log(n)$-connected with probability at least $1 - 1/n$.

Proof. Since $\alpha \le 2^{0.227k} \le 2^{(r_1 - 3\delta)k}/k^3 \le 2^{(r_1-\delta)k}/k^3$ for large enough k, w.h.p. over the choice of Φ, there is an $(r, r_1, 0, r_1)$-marking $(\mathcal{V}_{\mathrm{m}}, \emptyset, \mathcal{V}_{\mathrm{c}})$ of Φ, see Lemma 27. We run Algorithm 5 with inputs Φ, and the associated marking $(\mathcal{V}_{\mathrm{m}}, \emptyset, \mathcal{V}_{\mathrm{c}})$. W.h.p. over the choice of Φ, Lemma 59 holds. Therefore, with probability at least $1 - 1/n$ over the choice of two random satisfying assignments $\sigma \sim \mu_\Omega$ and $\sigma' \sim \mu_\Omega$, the output path of Algorithm 5 is well-defined by Lemma 58 and satisfies that $\|\zeta_i - \zeta_{i-1}\|_1 \le 2k^5 \log n$ for all $i \in [\ell]$ and $\|\xi_i - \xi_{i-1}\|_1 \le 2k^5 \log n$ for all $i \in [t]$ by Lemma 59. Hence, it is a D-path in the solution space Ω for $D = 2k^5 \log n$ as we wanted. □

10.2 Proof of Theorem 5

We next show $O(\log n)$-looseness for all variables with high probability over (Φ, σ) for random k-CNF instances Φ and uniformly random satisfying assignment $\sigma \in \Omega$. Consequently, in an algorithmic regime where $\alpha \ll 2^{ck}$ for some $c < 1$, we resolve a conjecture of [1]. Our proof exploits Corollary 57 on the size of the connected components of Φ^Λ. It is important in our arguments that every variable in the formula is flippable.

Definition 60. Let $\Phi = \Phi(k, n, m)$ be a random k-CNF. A variable $v \in \mathcal{V}$ is flippable if there exists a pair of satisfying assignments (X, Y) to Φ, in one of which $X(v) = \mathsf{F}$ and in the other $Y(v) = \mathsf{T}$.

Lemma 61. For $\alpha < 2^{k-2}$, with high probability over the choice of $\Phi = \Phi(k, m, n)$, all variables in Φ are flippable.

Proof. Observe that we can define an NAE-SAT problem based on Φ. By definition, any NAE-satisfying assignment ensures that every clause contains at least one satisfied literal and at least one unsatisfied literal. By Theorem 2 in [4], with high probability Φ is NAE-satisfiable. Consequently, we can find some assignment σ that NAE-satisfies Φ with high probability, and then the opposite assignment $\overline{\sigma}$ also NAE-satisfies Φ by the symmetry of NAE-SAT solutions. In particular, both σ and $\overline{\sigma}$ are solutions to the original SAT formula Φ. Observe that for every variable $v \in V$ we have $X(v) = \mathsf{T}$ and $X(v) = \mathsf{F}$ in exactly one of $\sigma, \overline{\sigma}$ and thus, with high probability, every variable in Φ is flippable. □

Lemma 62. For any variable $v \in \mathcal{V}$ and any partial assignment $X \colon \mathcal{V}_{\mathrm{m}} \setminus \{v\} \to \{\mathsf{F}, \mathsf{T}\}$, we have

$$\Pr_{\mu_\Omega}(v \mapsto \mathsf{F}|X) > 0 \quad \text{and} \quad \Pr_{\mu_\Omega}(v \mapsto \mathsf{T}|X) > 0.$$

Proof. We prove $\Pr_{\mu_\Omega}(v \to \mathsf{F}|X) > 0$; the proof of $\Pr_{\mu_\Omega}(v \to \mathsf{T}|X) > 0$ is analogous. We distinguish two cases.

The first case is when v is a good variable. Lemma 28 gives $\Pr_{\mu_\Omega}(v \mapsto \mathsf{F}|X, \Lambda_{\mathrm{bad}}) \geq 1 - \exp(1/k)/2 > 0$ for any satisfying assignment of the bad clauses Λ_{bad}. Thus, we have $\Pr_{\mu_\Omega}(v \mapsto \mathsf{F}|X) > 0$.

The second case is when v is a bad variable. By Lemma 61 there exists a satisfying assignment σ with $\sigma(v) = \mathsf{F}$. Let $\Lambda_{\mathrm{bad}} = \sigma|_{\mathcal{V}_{\mathrm{bad}}}$ be the assignment on bad variables and so in particular $\Pr_{\mu_\Omega}(\Lambda_{\mathrm{bad}}) > 0$. Then by Lemma 28 we have $\Pr_{\mu_\Omega}(X|\Lambda_{\mathrm{bad}}) \geq (1 - \exp(1/k)/2)^{|\mathcal{V}_{\mathrm{m}}|} > 0$. This implies that $\Pr_{\mu_\Omega}(X, \Lambda_{\mathrm{bad}}) > 0$ and in particular $\Pr_{\mu_\Omega}(v \mapsto \mathsf{F}, X) > 0$, so $\Pr_{\mu_\Omega}(v \mapsto \mathsf{F}|X) > 0$. □

We can now prove Theorem 5 with the help of Corollary 57.

Theorem 5. There is $k_0 \geq 3$ such that, for any integer $k \geq k_0$, and for any positive real $\alpha \leq 2^{0.227k}$, the random k-CNF formula $\Phi(k, n, \lfloor \alpha n \rfloor)$ is $\mathrm{poly}(k)\log(n)$-loose.

Proof. Note that $2^{0.227k} \leq 2^{(r_1 - 3\delta)k} \leq 2^{(r_1-\delta)k}/k^3$ for large enough k. Thus, w.h.p. over the choice of Φ, there is an $(r, r_1, \emptyset, r_1)$-marking $(\mathcal{V}_{\mathrm{m}}, \emptyset, \mathcal{V}_{\mathrm{c}})$ of Φ, see Lemma 26. The distribution $\mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$ is $(1/k)$-uniform by Lemma 28. Hence, Corollary 57 holds for $V = \mathcal{V}_{\mathrm{m}}$ and $\mu = \mu_\Omega|_{\mathcal{V}_{\mathrm{m}}}$. Let v be a variable of Φ. Let $\sigma \sim \mu_\Omega$ and let Λ be the restriction of σ to $\mathcal{V}_{\mathrm{m}} \setminus \{v\}$. Then, with probability at least $1 - n^{-k}$, the connected components of G_{Φ^Λ} have size at most $2k^4 \log n$. Let $\mathcal{C}_j^\Lambda$ be the connected component containing the variable v, which is empty if all clauses containing v are satisfied. Let ω be the negation of $\sigma(v)$. By Lemma 62, we have $\Pr_{\mu_\Omega}(v \mapsto \omega|\Lambda) > 0$. Therefore, there is an assignment Y of the variables in $\mathrm{var}(\mathcal{C}_j^\Lambda)$ that satisfies the clauses in $\mathcal{C}_j^\Lambda$ and has $Y(v) = \omega$. We construct the assignment σ' that has $\sigma'(v) = \omega$, agrees with Y in $\mathrm{var}(\mathcal{C}_j^\Lambda)$ and agrees with σ in the rest of the variables of Φ. In particular, this assignment agrees with Λ and satisfies each one of the connected components of Φ^Λ. Thus, σ' is a satisfying assignment of Φ. Moreover, w.h.p. σ' differs with σ in at most $2k^5 \log n$ variables (the variables in $\mathrm{var}(\mathcal{C}_j^\Lambda)$). We have shown that, w.h.p. over the choice of Φ, with probability at least $1 - n^{-k}$ a random assignment $\sigma \sim \mu_\Omega$ is $(2k^5 \log n)$-loose, so the statement follows. □

References

[1] Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In 49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, pages 793–802. IEEE Computer Society, 2008. doi:10.1109/FOCS.2008.11.

[2] Dimitris Achlioptas, Amin Coja-Oghlan, Max Hahn-Klimroth, Joon Lee, Noëla Müller, Manuel Penschuck, and Guangyan Zhou. The number of satisfying assignments of random 2-SAT formulas. Random Structures Algorithms, 58(4):609–647, 2021. doi:10.1002/rsa.20993.

[3] Dimitris Achlioptas, Amin Coja-Oghlan, and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. Random Structures Algorithms, 38(3):251–268, 2011. doi:10.1002/rsa.20323.

[4] Dimitris Achlioptas and Cristopher Moore. The asymptotic order of the random k-SAT threshold. In The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings., pages 779–788. IEEE, 2002. doi:10.1109/SFCS.2002.1182003.

[5] Vedat Levi Alev and Lap Chi Lau. Improved analysis of higher order random walks and applications. In STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pages 1198–1211. ACM, New York, 2020. doi:10.1145/3357713.3384317.

[6] Konrad Anand and Mark Jerrum. Perfect sampling in infinite spin systems via strong spatial mixing. SIAM J. Comput., 51(4):1280–1295, 2022. doi:10.1137/21M1437433.

[7] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science, pages 1319–1330. 2020. doi:10.1109/FOCS46700.2020.00125.

[8] Ivona Bezáková, Andreas Galanis, Leslie Ann Goldberg, and Daniel Štefankovič. Fast sampling via spectral independence beyond bounded-degree graphs. In 49th International Colloquium on Automata, Languages, and Programming (ICALP 2022), volume 229, pages 21:1–21:16, 2022. doi:10.4230/LIPIcs.ICALP.2022.21.

[9] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On mixing of Markov chains: Coupling, spectral independence, and entropy factorization. In Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 3670–3692, 2022. doi:10.1137/1.9781611977073.145.

[10] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Rapid mixing of Glauber dynamics up to uniqueness via contraction. In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science, pages 1307–1318. 2020. doi:10.1109/FOCS46700.2020.00124.

[11] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of Glauber dynamics: entropy factorization via high-dimensional expansion. In STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 1537–1550. ACM, New York, 2021. doi:10.1145/3406325.3451035.

[12] Amin Coja-Oghlan. A better algorithm for random k-SAT. SIAM J. Comput., 39(7):2823–2864, 2010. doi:10.1137/09076516X.

[13] Amin Coja-Oghlan and Alan Frieze. Analyzing Walksat on random formulas. SIAM J. Comput., 43(4):1456–1485, 2014. doi:10.1137/12090191X.

[14] Amin Coja-Oghlan, Noela Müller, and Jean B. Ravelomanana. Belief propagation on the random k-SAT model. Ann. Appl. Probab., 32(5):3718–3796, 2022. doi:10.1214/21-aap1772.

[15] Amin Coja-Oghlan and Angelica Y. Pachon-Pinzon. The decimation process in random k-SAT. SIAM J. Discrete Math., 26(4):1471–1509, 2012. doi:10.1137/110842867.

[16] Amin Coja-Oghlan and Konstantinos Panagiotou. The asymptotic k-SAT threshold. Adv. Math., 288:985–1068, 2016. doi:10.1016/j.aim.2015.11.007.

[17] Amin Coja-Oghlan and Daniel Reichman. Sharp thresholds and the partition function. In Journal of Physics: Conference Series, volume 473, page 012015. IOP Publishing, 2013. doi:10.1088/1742-6596/473/1/012015.

[18] Amin Coja-Oghlan and Nick Wormald. The number of satisfying assignments of random regular k-SAT formulas. Combin. Probab. Comput., 27(4):496–530, 2018. doi:10.1017/S0963548318000263.

[19] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. Ann. of Math. (2), 196(1):1–388, 2022. doi:10.4007/annals.2022.196.1.1.

[20] Paul Erdős and László Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In Infinite and finite sets (Colloq., Keszthely, 1973; dedicated to P. Erdős on his 60th birthday), Vol. II, pages 609–627. Colloq. Math. Soc. János Bolyai, 1975.

[21] Weiming Feng, Heng Guo, Yitong Yin, and Chihao Zhang. Fast sampling and counting k-SAT solutions in the local lemma regime. J. ACM, 68(6):Art. 40, 42, 2021. doi:10.1145/3469832.

[22] Weiming Feng, Kun He, and Yitong Yin. Sampling constraint satisfaction solutions in the local lemma regime. In STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 1565–1578. ACM, New York, 2021. doi:10.1145/3406325.3451101.

[23] Alan Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of k-SAT. J. Algorithms, 20(2):312–355, 1996. doi:10.1006/jagm.1996.0016.

[24] Andreas Galanis, Leslie Ann Goldberg, Heng Guo, and Kuan Yang. Counting solutions to random CNF formulas. SIAM J. Comput., 50(6):1701–1738, 2021. doi:10.1137/20M1351527.

[25] Bernhard Haeupler, Barna Saha, and Aravind Srinivasan. New constructive aspects of the Lovász local lemma. J. ACM, 58(6):Art. 28, 28, 2011. doi:10.1145/2049697.2049702.

[26] Kun He, Chunyang Wang, and Yitong Yin. Sampling Lovász local lemma for general constraint satisfaction solutions in near-linear time. In 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022, pages 147–158. IEEE Computer Soc., Los Alamitos, CA, 2022. doi:10.1109/FOCS54457.2022.00021.

[27] Kun He, Chunyang Wang, and Yitong Yin. Deterministic counting Lovász local lemma beyond linear programming. In Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 3388–3425. SIAM, Philadelphia, PA, 2023. doi:10.1137/1.9781611977554.ch130.

[28] Kun He, Kewen Wu, and Kuan Yang. Improved bounds for sampling solutions of random CNF formulas. In Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 3330–3361. SIAM, Philadelphia, PA, 2023. doi:10.1137/1.9781611977554.ch128.

[29] Jonathan Hermon, Allan Sly, and Yumeng Zhang. Rapid mixing of hypergraph independent sets. Random Structures Algorithms, 54(4):730–767, 2019. doi:10.1002/rsa.20830.

[30] Mark Huber. Approximation algorithms for the normalizing constant of Gibbs distributions. Ann. Appl. Probab., 25(2):974–985, 2015. doi:10.1214/14-AAP1015.

[31] Vishesh Jain, Huy Tuan Pham, and Thuy-Duong Vuong. On the sampling Lovász local lemma for atomic constraint satisfaction problems. arXiv preprint, 2021. URL: https://arXiv.org/abs/2102.08342.

[32] Vishesh Jain, Huy Tuan Pham, and Thuy Duong Vuong. Towards the sampling Lovász Local Lemma. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021, pages 173–183. IEEE Computer Soc., Los Alamitos, CA, 2022. doi:10.1109/FOCS52979.2021.00025.

[33] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. Theoret. Comput. Sci., 43(2-3):169–188, 1986. doi:10.1016/0304-3975(86)90174-X.

[34] Tali Kaufman and Izhar Oppenheim. High order random walks: beyond spectral gap. Combinatorica, 40(2):245–281, 2020. doi:10.1007/s00493-019-3847-0.

[35] Vladimir Kolmogorov. A faster approximation algorithm for the Gibbs partition function. In Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 228–249. PMLR, 06–09 Jul 2018. URL: https://proceedings.mlr.press/v75/kolmogorov18a.html.

[36] Florent Krzakała, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. Proceedings of the National Academy of Sciences, 104(25):10318–10323, 2007. doi:10.1073/pnas.0703685104.

[37] Marc Mézard, Thierry Mora, and Riccardo Zecchina. Clustering of solutions in the random satisfiability problem. Physical Review Letters, 94(19):197205, 2005. doi:10.1103/PhysRevLett.94.197205.

[38] Michael Mitzenmacher and Eli Upfal. Probability and computing. Cambridge University Press, Cambridge, 2005. Randomized algorithms and probabilistic analysis. doi:10.1017/CBO9780511813603.

[39] Ankur Moitra. Approximate counting, the Lovász local lemma, and inference in graphical models. J. ACM, 66(2):Art. 10, 25, 2019. doi:10.1145/3268930.

[40] Andrea Montanari and Devavrat Shah. Counting good truth assignments of random k-SAT formulae. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1255–1264. ACM, New York, 2007. URL: https://dl.acm.org/doi/abs/10.5555/1283383.1283518.

[41] Thierry Mora, Marc Mézard, and Riccardo Zecchina. Pairs of sat assignments and clustering in random boolean formulae. arXiv preprint, 2007. URL: https://arxiv.org/abs/cond-mat/0506053.

[42] Robin A. Moser and Gábor Tardos. A constructive proof of the general Lovász local lemma. J. ACM, 57(2):Art. 11, 15, 2010. doi:10.1145/1667053.1667060.

[43] Wolfgang Mulzer. Five proofs of Chernoff's bound with applications. Bull. Eur. Assoc. Theor. Comput. Sci. EATCS, 1(124):59–76, 2018. URL: http://bulletin.eatcs.org/index.php/beatcs/article/view/525.

[44] Danny Nam, Allan Sly, and Youngtak Sohn. One-step replica symmetry breaking of random regular NAE-SAT. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021, pages 310–318. IEEE Computer Soc., Los Alamitos, CA, 2022. doi:10.1109/FOCS52979.2021.00039.

[45] Allan Sly, Nike Sun, and Yumeng Zhang. The number of solutions for random regular NAE-SAT. Probab. Theory Related Fields, 182(1-2):1–109, 2022. doi:10.1007/s00440-021-01029-5.

[46] Joel Spencer. Asymptotic lower bounds for Ramsey functions. Discrete Math., 20(1):69–76, 1977/78. doi:10.1016/0012-365X(77)90044-9.

[47] Daniel Štefankovič, Santosh Vempala, and Eric Vigoda. Adaptive simulated annealing: a near-optimal connection between sampling and counting. J. ACM, 56(3):Art. 18, 36, 2009. doi:10.1145/1516512.1516520.

[48] Lenka Zdeborová. Statistical physics of hard optimization problems. arXiv preprint, 2008. URL: https://arxiv.org/abs/0806.4112.

Appendix A Proof of Lemma 20

In this section we prove Lemma 20. Recall that this result is [24, Lemma 8.16] with a less restrictive bound on the density of the formula and a more restrictive definition of good variables/clauses, see Section 4 for details. Moreover, the obtained upper bound on the number of bad clauses in our version of [24, Lemma 8.16] is tighter. The original proof of Lemma 20 given in [24, Section 8] is split into a sequence of results on random formulae. Here we restate some of these results — only those whose statement needs to change as a consequence of our definition of good variables/clauses and the tighter upper bound. We also explain how these changes affect the proofs if any modifications are necessary.

Let us fix some notation first. The results stated in this section only hold for large enough k unless we say otherwise. We note that in [24] the density α is at most $2^{k/300}/k^3$ and $\Delta = 2^{k/300}$, where Δ is the threshold in the definition of high-degree variables, and the proofs are carried out for these particular values. It turns out that, following the proofs in [24, Section 8], the only properties of α and Δ needed in order to proof Lemma 20 are that, for $r \in (0, 1/(2\log 2))$, we have $\Delta_r = \lceil 2^{rk} \rceil$ and α is bounded above by Δ_r/k^3 (note the subscript r here to indicate that Δ_r depends on r). First, we need some definitions. For any set of variables $S \subseteq \mathcal{V}$ of Φ, we denote by $\mathrm{HD}(S, r)$ the set of high-degree variables in S (recall that a variable is of high-degree if the degree of v is at least Δ_r).

Lemma 63 ([24, Lemma 8.1]). Let $r \in (0, 1)$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. The size of $\mathcal{V}_0(r) := \mathrm{HD}(\mathcal{V}, r)$ is at most $(\alpha/\Delta_r)n/2^{k^{10}}$.

Proof. The proof is the same to that of [24, Lemma 8.1], apart from one change that we highlight here. The degrees of the variables in Φ have the same distribution as a balls-and-bins experiment with km balls and n bins. Let $D_1, D_2, \ldots, D_n$ be independent variables following the Poisson distribution $\mathrm{Poi}(\mu)$ with parameter $\mu = k\alpha$. The degrees of the variables of Φ have the same distribution as $\{D_1, D_2, \ldots, D_n\}$ conditioned on the event $\mathcal{E}$ that $D_1 + D_2 + \cdots + D_n = m$, see for instance [38, Chapter 5.4]. Let $U = \{i \in [n] : D_i \geq \Delta_r\}$. We want to show that $\Pr(|U| > (\alpha/\Delta_r)n/2^{k^{10}} | \mathcal{E}) = o(1)$. In [24, Lemma 8.1] the authors show that $\Pr(|U| > n/2^{k^{10}} | \mathcal{E}) = o(1)$. Their bound is not tight, but it is enough for their purposes. In fact, one can change k^{10} by any polynomial and the result would still hold for large enough k. Here we obtain the extra factor α/Δ_r by slightly modifying the application of the tail bound $\Pr(\mathrm{Poi}(\mu) \geq x) \leq e^{-\mu}(e\mu)^x/x^x$. For $x = \Delta_r$, instead of using the bound $e^{-\mu}(e\mu)^x/x^x \leq e^{-\Delta_r} \leq 2^{-k^{10}-1}$, which holds for large enough k as $\mu/x \leq k^{-2}$ and Δ_r is exponential in k, we use the bound $e^{-\mu}(e\mu)^x/x^x \leq (e\mu/x)e^{-x+1} \leq (\alpha/\Delta_r)2^{-k^{10}-1}$. The rest of the proof is analogous; we have $\mathbb{E}[|U|] \geq n(\alpha/\Delta_r)2^{-k^{10}-1}$, so by a Chernoff bound we find that $\Pr(|U| \geq (\alpha/\Delta_r)n/2^{k^{10}}) \leq \exp(-\Omega(n))$. From the connection between a balls-and-bins experiment and the Poisson distribution, see [38, Theorem 5.7], we conclude that $\Pr(|U| \geq (\alpha/\Delta_r)n/2^{k^{10}} | \mathcal{E}) \leq \exp(-\Omega(n))$ as we wanted. □

Corollary 64 ([24, Corollary 8.4]). There is a positive integer k_0 such that for any integer $k \geq k_0$ and any density α with $\alpha \leq 2^k/(ek^3)$ the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For every set of variables Y such that $2 \leq |Y| \leq n/2^k$, the number of clauses that contain at least 3 variables from Y is at most $|Y|$.

Proof. This is a consequence of [24, Lemma 35] with $b = 3$ and $t = 2/(b-1) = 1$, whose proof only requires $\alpha \leq 2^k/(ek^3)$. □

Recall that the graph H_Φ is the dependency graph of the variables of Φ, see Definition 18.

Lemma 65 ([24, Lemma 8.8]). Let $r \in (0,1)$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r / k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Every connected set U of variables in H_Φ with size at least $2k^4 \log n$ satisfies that $|\mathrm{HD}(U, r)| \leq \frac{1}{2k^3}|U|$.

Proof. The proof is that of [24, Lemma 8.8], with the difference that $\delta_0 = 1/(2k^3)$ instead of $\delta_0 = 1/21600$, as the exact value of δ_0 does not play a role in the proof as long as, for $\theta_0 = \Delta_r - 2(k+1)$, we have $\delta_0 \theta_0 \log \frac{\theta_0}{k^2 \alpha} \geq 3 \log k + \log \alpha$, which holds for large enough k when $\delta_0 = \mathrm{poly}(k)$. Moreover, the only restriction on α is that of Corollary 64, and the fact that $\alpha \leq \Delta_r / k^3$. □

Lemma 66 ([13, Lemma 2.4] and [24, Lemma 8.10]). Let $k \geq 3$ be an integer and let α be a positive real number with $\alpha \leq e^{k/2}/(2e^2 k^2)$. For any $\varepsilon \in [1/n, 1)$ (depending on n) such that $\varepsilon < e^{-3k}$ for all n, the following holds w.h.p. over the choice of the random formula $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let Z be a set of clauses with size at most εn and let $c_1, \ldots, c_l \in \mathcal{C} \setminus Z$ be distinct clauses. For $s \in \{1, 2, \ldots, \ell\}$, let $N_s := \mathrm{var}(Z) \cup \bigcup_{j=1}^{s-1} \mathrm{var}(c_j)$. If $|\mathrm{var}(c_s) \cap N_s| \geq 3$ for all $s \in \{1, 2, \ldots, \ell\}$, then $\ell \leq \varepsilon n$.

Proof. The proof is almost identical to the proof of [13, Lemma 2.4]. There are four differences. First, here, as it is also the case in [24, Lemma 44], ε can depend on n. This will arise later in this proof. Second, the proof of [13, Lemma 2.4] is carried out for the condition $|\mathrm{var}(c_s) \cap N_s| \geq \lambda$, where λ is an integer with $\lambda > 4$. Here we set $\lambda = 3$ and impose stricter hypotheses on α and ε to compensate for a smaller λ. Their (more relaxed) hypotheses on α and ε are $\alpha \leq 2^k \log 2$, $\varepsilon \leq k^{-3}$ and $\varepsilon^\lambda \leq (2e)^{-4k}/e$. Third, we substitute the last inequality of [13, Equation 4], which is

$$\left[\left(\frac{em/n}{\varepsilon} \right)^2 \exp(2k)(2k\varepsilon)^\lambda \right]^{\varepsilon n} \leq \left[(2e)^{2k} \varepsilon^{\lambda/2} \right]^{\varepsilon n},$$

by the inequality

$$\begin{aligned} \left[\left(\frac{em/n}{\varepsilon} \right)^2 \exp(2k)(2k\varepsilon)^\lambda \right]^{\varepsilon n} &\leq \left[(em/n)^2 \exp(2k)(2k)^3 \varepsilon \right]^{\varepsilon n} \\ &\leq \left[\exp(3k-1)\varepsilon \right]^{\varepsilon n}, \end{aligned} \tag{37}$$

where we used $\lambda = 3$ and $m/n \leq \alpha \leq e^{k/2}/(2e^2k^2)$. Now, as it is done in [24, Lemma 8.10], we distinguish two cases depending on ε. If $\varepsilon \geq 10(\log n)/n$, then using this in conjunction with $\varepsilon < e^{-3k}$, the right hand size of (37) is bounded by $e^{-\varepsilon n} \leq 1/n^{10} = o(1/n)$. If $1/n \leq \varepsilon < 10(\log n)/n$, then, for large enough n, the right hand size of (37) is bounded above by $\exp(3k-1)\varepsilon = o(1)$. The last difference between the proofs is that our argument works for all $k \geq 3$, whereas the bound [13, Equation 4] only holds for large k. □

The remaining results in this section do not need any changes in their original proofs, other than that every time Corollary 8.4, Lemma 8.8 and Lemmas 8.10-8.16 are invoked in [24, Section 8], we use the version given in this appendix instead. We note that the statements of these results are slightly different to their [24, Section 8] versions, and these changes are again due to the fact that we use $\lambda = 3$ instead of $\lambda = k/10$ in the definition of good variables/clauses.

Corollary 67 ([24, Corollary 8.11]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r / k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Let Z be a set of clauses with size at most $2n/2^{k^{10}}$ and let $c_1, \ldots, c_l \in \mathcal{C} \setminus Z$ be distinct clauses. For $s \in \{1, 2, \ldots, \ell\}$, let $N_s := \mathrm{var}(Z) \cup \bigcup_{j=1}^{s-1} \mathrm{var}(c_j)$. If $|\mathrm{var}(c_s) \cap N_s| \geq 3$ for all $s \in \{1, 2, \ldots, \ell\}$, then $\ell \leq |Z|$.

Proof. The proof given in [24, Corollary 8.11] also applies here. We note that the density α is at most $e^{k/2}/(2e^2k^2)$ so we can indeed apply Lemma 66 when the proof given in [24, Corollary 8.11] invokes [24, Lemma 8.10]. □

Lemma 68 ([24, Lemma 8.13]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For any bad component S of variables, we have $|S| \leq 2k|\mathrm{HD}(S, r)|$.

Proof. The proof given in [24, Lemma 8.13] applies using our versions of [24, Lemma 8.1, Corollary 8.4 and Corollary 8.11]. □

Lemma 69 ([24, Lemma 8.14]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. Every bad component S has size at most $2k^4 \log n$.

Proof. The proof given in [24, Lemma 8.14] applies using our versions of [24, Lemma 8.8 and Lemma 8.13]. □

Lemma 70 ([24, Lemma 8.15]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For every connected set of S variables with size at least $2k^4 \log n$, we have $|S \cap \mathcal{V}_{\mathrm{bad}}| \leq |S|/k^2$.

Proof. The proof is analogous to that given in [24, Lemma 8.15]. The only differences are that we apply Lemma 65 instead of [24, Lemma 8.8], we apply Lemma 68 instead of [24, Lemma 8.13], and we have $\delta_0 = 1/(2k^3)$ instead of $\delta_0 = 1/21600$. □

Lemma 20 ([24, Lemma 8.16]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. For every connected set of clauses Y in G_Φ such that $|\mathrm{var}(Y)| \geq 2k^4 \log n$, we have $|Y \cap \mathcal{C}_{\mathrm{bad}}(r)| \leq |Y|/k$.

Proof. The same proof applies using our versions of [24, Corollary 8.4 and Lemma 8.15]. □

Lemma 21 ([24, Lemma 8.12]). Let $r \in (0, 1/(2\log 2)]$. There is a positive integer k_0 such that for any integer $k \geq k_0$, $\Delta_r = \lceil 2^{rk} \rceil$, and any density α with $\alpha \leq \Delta_r/k^3$, the following holds w.h.p. over the choice of $\Phi = \Phi(k, n, \lfloor \alpha n \rfloor)$. We have $|\mathcal{C}_{\mathrm{bad}}(r)| \leq 2(\alpha/\Delta_r)n/2^{k^{10}}$ and $|\mathcal{V}_{\mathrm{bad}}(r)| \leq 2(k+1)(\alpha/\Delta_r)n/2^{k^{10}}$.

Proof. We consider the set of high-degree variables $\mathcal{V}_0(r) = \mathrm{HD}(\mathcal{V}, r)$, which w.h.p. over the choice of Φ has $|\mathcal{V}_0(r)| \leq (\alpha/\Delta_r)n/2^{k^{10}}$ by Lemma 63. In view of Corollary 64 with $Y = \mathcal{V}_0(r)$, we have $|\mathcal{C}_0(r)| \leq |\mathcal{V}_0(r)| \leq n/2^{k^{10}}$, where $\mathcal{C}_0(r)$ is the set of clauses with at least 3 variables in $\mathcal{V}_0(r)$, see Algorithm 2. From Corollary 67 and the construction of $\mathcal{C}_{\mathrm{bad}}(r)$ in Algorithm 2, we find that $|\mathcal{C}_{\mathrm{bad}}(r)| \leq 2|\mathcal{C}_0(r)| \leq 2|\mathcal{V}_0(r)| \leq 2(\alpha/\Delta_r)n/2^{k^{10}}$. By construction of $\mathcal{V}_{\mathrm{bad}}(r)$, see Algorithm 2, we conclude that $|\mathcal{V}_{\mathrm{bad}}(r)| \leq |\mathcal{V}_0(r)| + k|\mathcal{C}_{\mathrm{bad}}(r)| \leq 2(k+1)(\alpha/\Delta_r)n/2^{k^{10}}$. □

Appendix B Proof of Lemma 13

In this section we collect the results from [11] that one needs to combine to obtain Lemma 13 on the mixing time of the ρ-uniform-block Glauber dynamics.

Definition 71. Let μ be a distribution supported on $\Omega \subseteq [q]^V$. Let $f\colon \Omega \to \mathbb{R}_{\geq 0}$. We denote the entropy of f by $\mathrm{Ent}_\mu(f)$, that is, $\mathrm{Ent}_\mu(f) = \mathbb{E}_\mu(f \log f)) - \mathbb{E}_\mu(f)\log(\mathbb{E}_\mu(f))$ when $\mathbb{E}_\mu(f) > 0$, and $\mathrm{Ent}_\mu(f) = 0$ when $\mathbb{E}_\mu(f) = 0$. For $S \subseteq V$, we denote $\mathrm{Ent}_\mu^S(f) = \mathbb{E}_{\tau\sim\mu|_{V\setminus S}} \mathrm{Ent}_\mu(f|\,\tau)$, where $\mathrm{Ent}_\mu(f|\,\tau)$ is the entropy of f conditioning to the event that the assignment drawn from μ agrees with τ in $V \setminus S$.

Let $\rho \in \{1, 2, \ldots, n\}$. We say that μ satisfies the ρ-uniform block factorisation of entropy (with constant C_ρ) if for all $f\colon \Omega \to \mathbb{R}_{\geq 0}$ we have

$$\frac{\rho}{n}\,\mathrm{Ent}_\mu(f) \leq C_\rho \frac{1}{\binom{n}{\rho}} \sum_{S\in\binom{V}{\rho}} \mathrm{Ent}_\mu^S(f).$$

One of the main results of [11] is showing that μ satisfies the ρ-uniform block factorisation of entropy when the distribution μ is η-spectrally independent and b-marginally bounded. In the proof of [8, Corollary 19] the authors observe that the proof of Lemma 72 also holds when η depends on n and, in particular, in the case $\eta = \varepsilon \log n$.

Lemma 72 ([11, Lemma 2.4]). The following holds for any reals $b, \eta > 0$, any $\kappa \in (0, 1)$ and any integer n with $n \geq \frac{2}{\kappa}(4\eta/b^2 + 1)$.

Let $q \geq 2$ be an integer, let V be a set of size n and let μ be a distribution over $[q]^V$. If μ is b-marginally bounded and η-spectrally independent, then μ satisfies the $\lceil \kappa n \rceil$-uniform block factorisation of entropy with constant $C = (2/\kappa)^{4\eta/b^2+1}$.

It turns out that one can bound the mixing time of the ρ-uniform-block Glauber dynamics when the target distribution μ satisfies the ρ-uniform block factorisation of entropy.

Lemma 73 (See, e.g., [11, Lemma 2.6 and Fact 3.5(4)] or [8, Lemma 17]). Let $q \geq 2$, $\rho \geq 1$ be integers and V be a set of size $n \geq \rho + 1$. Let μ be a distribution supported on $\Omega \subseteq [q]^V$ that satisfies the ρ-uniform-block factorisation of entropy with multiplier C_ρ. Then, for any $\varepsilon > 0$, the mixing time of the ρ-uniform-block Glauber dynamics on μ satisfies, for $\mu_{\min} = \min_{\Lambda\in\Omega} \mu(\Lambda)$,

$$T_{\mathrm{mix}}(\varepsilon) \leq \left\lceil C_\rho \frac{n}{\rho} \left(\log\log \frac{1}{\mu_{\min}} + \log \frac{1}{2\varepsilon^2} \right) \right\rceil.$$

Proof of Lemma 13. The proof of Lemma 13 follows directly from combining Lemmas 72 and 73. □

Appendix C Notation and definitions reference

Here we gather the notation and definitions that are used globally in our work. If some notation is not here, then it is only used in one section of our work (and it is defined in that section).

C.1 Table of notation

Notation	Description	Reference
$\Phi(k, n, m)$	A random k-CNF formula with n variables and m clauses.	Section 1
α	The density of the formula Φ, so $\alpha = m/n$.	Section 1
$\mathcal{V}$	The set of variables of Φ.	Section 1
$\mathcal{C}$	The set of clauses of Φ.	Section 1
w.h.p.	Stands for "with high probability".	Section 1
d_{TV}	The total variation distance between two distributions.	Section 1
ξ	Our sampling algorithm has error at most $n^{-\xi}$.	Theorem 1
Δ_r	The high-degree threshold, set to $\lceil 2^{(r_0-\delta)k} \rceil$.	Definition 6
r_0, r_1, δ	$r_0 = 0.117841$, $r_1 = 0.227092$ and $\delta = 0.00001$.	Definition 8
$\mathrm{var}(c)$	The set of variables in a clause c.	Section 2.1
$\mathrm{var}(S)$	The set of variables $\bigcup_{c \in S} \mathrm{var}(c)$.	Section 2.1
$\mathcal{C}_{\mathrm{good}}(r)$, $\mathcal{C}_{\mathrm{bad}}(r)$	Good and bad clauses, a partition of $\mathcal{C}$.	Section 4
$\mathcal{V}_{\mathrm{good}}(r)$, $\mathcal{V}_{\mathrm{bad}}(r)$	Good and bad variables, a partition of $\mathcal{V}$.	Section 4
$\mathcal{V}_{\mathrm{m}}$, $\mathcal{V}_{\mathrm{a}}$, $\mathcal{V}_{\mathrm{c}}$	The sets of marked, auxiliary and control variables.	Definition 8
Ω^*	The set of all assignments $\mathcal{V} \to \{\mathsf{F}, \mathsf{T}\}$	Definition 9
Ω	The set of satisfying assignments of Φ.	Definition 9
μ_A	The uniform distribution over $A \subseteq \Omega*$.	Definition 9
Φ^Λ	The formula Φ simplified under Λ.	Definition 9
$\mathcal{V}^\Lambda$, $\mathcal{C}^\Lambda$	The variables and clauses of Φ^Λ	Definition 9
Ω^Λ	The set of satisfying assignments of Φ^Λ.	Definition 9
$\mu\|_V$	The marginal distribution of μ on V.	Definition 11
$T_{\mathrm{mix}}(\rho, \varepsilon)$	The mixing time of the ρ-uniform-block Glauber dynamics.	Section 2.2.1
$\mathcal{I}^\Lambda(u \to v)$	The influence of u on v (under Λ).	Section 2.2.1, (1)
G_Φ	The dependency graph of $\mathcal{C}$.	Definition 16
H_Φ	The dependency graph of $\mathcal{V}$.	Definition 18
$\Phi_{\mathrm{good}}(r)$	The subformula of Φ with all good variables and good clauses.	Definition 24
$\Phi_{\mathrm{bad}}(r)$	The subformula of Φ with all bad variables and bad clauses.	Definition 24

C.2 Table of definitions

Name	Reference
high-degree	Definition 6, page 5
r-distributed	Definition 8, page 6
$(r, r_{\mathrm{m}}, r_{\mathrm{a}}, r_{\mathrm{c}})$-marking	Definition 8, page 6
ε-uniform	Definition 12, page 7
b-marginally bounded	Section 2.2.1, page 8
η-spectrally independent	Section 2.2.1, page 8