Analysis of Local Search Landscapes for k-SAT Instances

A.A. Albrecht, P.C.R. Lane and K. Steinhöfel

Abstract. Stochastic local search is a successful technique in diverse areas of combinatorial optimisation and is predominantly applied to hard problems. When dealing with individual instances of hard problems, gathering information about specific properties of instances in a preprocessing phase is helpful for an appropriate parameter adjustment of local search-based procedures. In the present paper, we address parameter estimations in the context of landscapes induced by k-SAT instances: at first, we utilise a sampling method devised by Garnier and Kallel in 2002 for approximations of the number of local maxima in landscapes generated by individual k-SAT instances and a simple neighbourhood relation. The objective function is given by the number of satisfied clauses. The procedure provides good approximations of the actual number of local maxima, with a deviation typically around 10%. Secondly, we provide a method for obtaining upper bounds for the average number of local maxima in k-SAT instances. The method allows us to obtain the upper bound $2^{n-O(\sqrt{n/k})}$ for the average number of local maxima, if m is in the region of $2^k \cdot n/k$.

Mathematics Subject Classification (2010). Primary 68R99; Secondary 90C27.

Keywords. Combinatorial landscapes, local search, SAT problem, phase transition, Gamma distribution.

1. Introduction

In recent years, much attention has been paid to local search algorithms as one of the basic methods to solve k-SAT problems. A first summary was presented in [14] along with an empirical analysis of run-time distributions for various local search-based methods such as WalkSAT [30]. Improvements on run-time estimations for k-SAT problems as well as for CNFs with unconstrained clause lengths are reported by a number of authors [1, 7, 9, 11, 21, 25, 26], where the results are partly based on randomised local search methods. Significant

progress has been achieved in the analysis of phase transitions since this effect was reported in [18] and [31]. Sophisticated methods from statistical mechanics [19, 17, 20] provided quite accurate estimates for the crucial phase transition parameter, which eventually led to a rigorous proof of a tight bound of $2^k \cdot \log 2 - O(k)$ for the phase transition threshold as presented in [2]; for an overview on statistical mechanics applied to combinatorial optimization we refer the reader to [16].

In the present paper, we attempt to analyse the number of local maxima in a combinatorial landscape induced by a k-CNF and a simple neighbourhood function, with the objective function being the number of satisfied clauses for a given assignment of binary values. In recent years, combinatorial landscape analysis has become a major tool in the design of search-based algorithms, see [23]. For example, instance-specific landscape parameters such as the maximum value of the minimum escape height from local minima can be utilised to obtain relatively tight bounds for the termination of local search when coupled with a confidence parameter, see [3]. The application of this type of run-time bounds to protein folding simulation exhibits a close correspondence between the simulation time (in number of transitions) and estimates of real folding times (in nanoseconds) of protein sequences [5, 32], which is due to the common source of thermodynamics (simulated annealing, minimizing free energy in protein foldings).

In [22] it has been demonstrated how to incorporate the number of local optima into run-time estimates of local search algorithms. For landscapes that can be partitioned into attraction basins, they proved that with probability α all local optima have been covered by local search with random restart after a waiting time of $\nu \cdot \ln(\nu+\gamma) + z_\alpha \cdot \sqrt{(\nu \cdot \pi)^2/6 + 1 - \nu \cdot \ln(\nu+\gamma)}$, where ν is the number of local optima, γ is the Euler-Mascheroni constant, and z_α is an appropriate confidence coefficient. Thus, estimates for ν provide information, e.g., for the selection of the population size in parallelized versions of local search algorithms, such as genetic algorithms or evolutionary algorithms in general.

Related work on combinatorial landscapes is presented in [15] and [34]. Zhang [34] proposes a landscape-based method that performs especially well on overconstrained random MAX-SAT instances. Moreover, Zhang's algorithm finds satisfiable solutions on large k-SAT instances more often than WalkSAT. The paper highlights the importance of how to deal with individual instances rather than with collections of (randomly selected) problem instances. In the context of the present paper, it is interesting to note that Zhang [34] reports a relatively small number of local minima for $n = 100$. Kaski [15] proved that for k-SAT instances with a constant number $m = \text{const}$ of conjunctions the number of local maxima (or minimum number of violated conjunctions) is of order $2^{\Omega(n)}$, with typically barriers of order $\Omega(n)$ between maxima. In the present paper, we aim at variable m that cover the region of phase transition.

We utilise the approach devised in [10] for estimating the number of local maxima for a given problem instance, where sample data are used to approximate a probability distribution associated with the landscape induced by the problem instance. The results are discussed against the information gathered by a complete analysis of the landscape for a limited number of k-SAT problem instances. Given the nature of the problem (i.e., complete search for local maxima and, in particular, optimising parameters of stochastic models for 20 instances per each (n, m)-pair), we were able to analyse only small-scale instances and overall only a limited number of different, randomly generated k-SAT instances. Apart from the experimental analysis based on the Garnier/Kallel-approach, we derive an estimation of the average number of local maxima per k-CNF in terms of parameters of individual problem instances for the given, simple neighbourhood relation. We note that the calculations depend on the type of the neighbourhood, i.e. other neighbourhood relations may produce different values, which will be the subject of future research.

2. Basic Notations

We follow mainly the notations from [2]: for a set V of n Boolean variables let $C_k(V)$ denote the set of all $\binom{n}{k} \cdot 2^k$ different disjunctive k-clauses on V, i.e. repeated literals and tautologies are excluded. A k-CNF is formed by selecting m different clauses C from $C_k(V)$ and taking their conjunction. We note that the selection does not imply - as in [2] - that the k-CNF strictly depends upon all n variables. The set of all such k-CNF consisting of m clauses is denoted by $\mathsf{F}_k(n, m)$. The set of m clauses forming $F \in \mathsf{F}_k(n, m)$ is denoted by $\mathsf{C}(F)$, and $\mathsf{Z}_F(\tilde{\sigma})$ is the number of satisfied clauses $C \in \mathsf{C}(F)$ on the truth assignment $\tilde{\sigma} = (\sigma_1, ..., \sigma_n)$, i.e. $0 \leq \mathsf{Z}_F(\tilde{\sigma}) \leq m$ and F is satisfiable, if there exists $\tilde{\eta}$ such that $\mathsf{Z}_F(\tilde{\eta}) = m$.

In [27], various neighbourhood functions are analysed that employ information about $\mathsf{Z}_F(\tilde{\sigma})$ and elements of $\mathsf{C}(F)$ that maximise changes of the objective function in one way or another. For example, flipping values of truth assignments is determined by unsatisfied clauses only [29, 28]. We consider a simple, unconstrained (i.e., features of clauses w.r.t. $\mathsf{Z}_F(\tilde{\sigma})$ are not taken into account) neighbourhood function where the value of a single variable is flipped, which makes it possible to consider the elements of the unit cube $\{0, 1\}^n$ as elements of the configuration space. Thus, the landscape $\mathsf{L}(F)$ for F is induced by $\mathsf{Z}_F(\tilde{\sigma})$, $\tilde{\sigma} \in \{0, 1\}^n$, and the neighbourhood relation

$$\mathsf{N}(\tilde{\sigma}) = \big\{\tilde{\sigma}' | d(\tilde{\sigma}, \tilde{\sigma}') = 1\big\}, \tag{2.1}$$

where $d(\tilde{\sigma}, \tilde{\sigma}')$ is the Hamming distance.

A path $w(\tilde{\sigma}, \tilde{\sigma}')$ of length $\ell \geq 0$ within $\mathsf{L}(F)$ is a sequence of $\tilde{\sigma}_i \in \mathsf{L}(F)$ such that $\tilde{\sigma}_0 = \tilde{\sigma}$, $\tilde{\sigma}_\ell = \tilde{\sigma}'$, and $\tilde{\sigma}_{i+1} \in \mathsf{N}(\tilde{\sigma}_i)$, $i = 1, ..., \ell$–1. We use $\tilde{\sigma}_i \in w(\tilde{\sigma}, \tilde{\sigma}')$, if $\tilde{\sigma}_i$ belongs to $w(\tilde{\sigma}, \tilde{\sigma}')$. For the simple neighbourhood (2.1), each $\tilde{\sigma}' \in \mathsf{L}(F)$ is reachable from a fixed $\tilde{\sigma}$ through a path of length $\ell \leq n$.

Let $W(\tilde{\sigma})$ denote the set of all paths $w(\tilde{\sigma}, \tilde{\sigma}')$ of length $\ell \leq n$.

Definition 1. If $\forall w(\tilde{\sigma},\tilde{\sigma}')\forall\tilde{\sigma}_i\Big(w(\tilde{\sigma},\tilde{\sigma}')\in W(\tilde{\sigma})\wedge\big(\tilde{\sigma}_i,\tilde{\sigma}_{i+1}\in w(\tilde{\sigma},\tilde{\sigma}')\wedge\big(\mathsf{Z}_F(\tilde{\sigma}_{i+1})>\mathsf{Z}_F(\tilde{\sigma}_i)\big)\rightarrow\exists\tilde{\sigma}_j\big(\tilde{\sigma}_j,\tilde{\sigma}_{j+1}\in w(\tilde{\sigma},\tilde{\sigma}')\wedge\big(j<i\big)\wedge\big(\mathsf{Z}_F(\tilde{\sigma}_{j+1})<\mathsf{Z}_F(\tilde{\sigma}_j)\big)\Big)\Big)$, then $\tilde{\sigma}$ is called a local maximum. The number of local maxima of F is denoted by $\mathsf{N}_{\mathrm{lm}}(F)$.

In order to simplify the combinatorial analysis of local maxima, we consider a potentially larger subset of $\mathsf{L}(F)$ by taking into account only the neighbourhood $\mathsf{N}(\tilde{\sigma})$ rather than the set of paths $W(\tilde{\sigma})$:

Definition 2. If $\forall\tilde{\sigma}'\big(\tilde{\sigma}'\in\mathsf{N}(\tilde{\sigma})\rightarrow(\mathsf{Z}_F(\tilde{\sigma}')\leq\mathsf{Z}_F(\tilde{\sigma})\big)$, then $\tilde{\sigma}$ is called a one-step local maximum. The number of one-step local maxima is denoted by $\mathsf{N}^1_{\mathrm{lm}}(F)$.

Since $\mathsf{N}_{\mathrm{lm}}(F)\leq\mathsf{N}^1_{\mathrm{lm}}(F)$, an upper bound for $\mathsf{N}^1_{\mathrm{lm}}$ is also valid for N_{lm}.

3. The Garnier/Kallel-Approach

In the present paper, we are solely concerned with the landscape analysis called *inverse problem* [10], i.e. M elements of the landscape are selected at random as initial points of a pre-defined local search procedure. Then, for j initial points, where $1\leq j\leq M$, the local search procedure is started and executed until a (local) maximum has been detected. The number of *different* (local) maxima is denoted by β_j. The local search procedure is quasi-deterministic and follows the steepest ascent rule: for the intermediate landscape element $\tilde{\sigma}$, all elements of $\mathsf{N}(\tilde{\sigma})$ are examined and one of the neighbours $\tilde{\sigma}'$ with the highest value of $\mathsf{Z}_F(\tilde{\sigma}')$ among all neighbours is chosen as the successor of $\tilde{\sigma}$ in the search procedure. The search terminates if no improvement of the objective function can be achieved. In [10], and the same applies to [22], a single element $\tilde{\sigma}'\in\mathsf{N}(\tilde{\sigma})$ is assumed at each step that maximises $\mathsf{Z}_F(\tilde{\sigma}')$, which implies a partition of L into attraction basins A_i, where $1\leq i\leq N$ for a total number of N local and global maxima. The set A_i consists of all elements of L that lead to the i^{th} local or global maximum by the steepest ascent local search. The assumption affects the normalised size $\alpha_i=|A_i|/|\mathsf{L}|$ of attraction basins and $\sum_{i=1}^{N}|A_i|/|\mathsf{L}|=1$. Since we employ the Garnier/Kallel-approach in an experimental context, we assume in the following that the impact of random selections among $\tilde{\sigma}'$ that maximise $\mathsf{Z}_F(\tilde{\sigma}')$ within a given neighbourhood is negligible.

Garnier and Kallel [10] assume that the normalised sizes α_i of attraction basins can be described by a distribution parametrized by some positive number γ as follows: let $(Z_i)_{i=1,..,N}$ be a sequence of independent random variables whose common distribution has density p_γ defined by

$$p_\gamma=\frac{\gamma^\gamma}{\Gamma(z)}\cdot z^{\gamma-1}\cdot e^{-\gamma\cdot z},\tag{3.1}$$

where $\Gamma(z)=\int_0^\infty e^{-t}\cdot t^{z-1}dt$, i.e. (3.1) represents the Gamma distribution with the parameter setting $[\gamma,\gamma]$, see [10]. Let H^γ denote the assumption

that the $(\alpha_i)_{i=1,...,N}$ can be approximated by $(Z_i/T_N)_{i=1,...,N}$, where $T_N = \sum_{i=1}^{N} Z_i$ with each Z_i having the density function p_γ. Furthermore, let $\beta_{j,\gamma} = \mathbb{E}_\gamma[\beta_j]$ denote the expected value of β_j, $j = 1, ..., M$. Garnier and Kallel [10] prove that

$$\beta_{j,\gamma} \quad = \quad N \cdot \binom{M}{j} \cdot \frac{\Gamma(\gamma+j)}{\Gamma(\gamma)} \cdot \frac{\Gamma(N\cdot\gamma)}{\Gamma((N-1)\cdot\gamma)} \cdot \frac{\Gamma((N-1)\cdot\gamma+M-j)}{\Gamma(N\cdot\gamma+M)}. \quad (3.2)$$

We note that for $N = M/r$, a fixed value of M, and appropriate approximations of the Γ-function, the $\beta_{j,\gamma}$ can be approximated according to (3.2) as functions of (j, γ, r). For fixed r, Garnier and Kallel [10] propose the χ^2 test to approximate γ for H^γ, which consists of calculating

$$T_\gamma = \sum_{j=1}^{M} \frac{(\beta_j - \beta_{j,\gamma})^2}{\beta_{j,\gamma}}, \quad (3.3)$$

where the β_j are given from observation and the $\beta_{j,\gamma}$ are approximated according to (3.2). The goal is then to determine

$$\gamma_0(r) = \text{argmin}\{T_\gamma, \gamma > 0\} \quad (3.4)$$

by appropriate numerical methods. In our computational experiments, we incorporate the approximation of $\gamma_0(r)$ as a sub-routine in calculations where the parameter r varies (is decremented) until $\gamma_0(r)$ changes only marginally for $r = r_{\text{appr}}$, see Section 4. Thus, for a fixed (but sufficiently large) value of M the number of local maxima is finally estimated by

$$N_{\text{appr}} = \frac{M}{r_{\text{appr}}}. \quad (3.5)$$

3.1. Evaluation of random 3-SAT instances

We fixed $k = 3$ and for $n = 18$, 20 we randomly generated twenty instances from $\mathsf{F}_3(n, m)$ for varying ratios m/n around the phase transion threshold $m/n \approx 4.267$. For each of the k-CNF we executed a complete search for local/global maxima in $\{0,1\}^{18}$ and $\{0,1\}^{20}$. We then selected three values for M, the number of random points chosen in $\{0,1\}^{22}$, $\{0,1\}^{23}$ and $\{0,1\}^{24}$ as initial elements for a deterministic steepest ascent search for local maxima. For each of the values M_i, $i = 1, 2, 3$, and the natural order of the M_i points we counted by β_j^i, $j = 1, ..., M_i$ the number of different maxima detected by the first j starting points for steepest ascent search.

3.2. Approximation of H^γ

For the calculation of $\beta_{j,\gamma}$ according to (3.2) we implemented the following procedure, which actually approximates $\beta_{j,\gamma}$, since we employ an approximation of the Γ-function. We recall that in (3.2) the (unknown) N is substituted by M/r, where M is selected as described in Section 3.1 and r is a variable

in our calculations. At first, we represent Eqn. 3.2 by

$$\beta_{j,\gamma} = \frac{M}{r} \cdot \binom{M}{j} \cdot \frac{A_1}{A_2} \cdot \frac{B_1}{B_2} \cdot \frac{C_1}{C_2}, \text{ where} \tag{3.6}$$

$$A_1 = \Gamma(a_1) \text{ for } a_1 = \gamma + j; \tag{3.7}$$

$$A_2 = \Gamma(a_2) \text{ for } a_2 = \gamma; \tag{3.8}$$

$$B_1 = \Gamma(b_1) \text{ for } b_1 = \gamma \cdot \frac{M}{r}; \tag{3.9}$$

$$B_2 = \Gamma(b_2) \text{ for } b_2 = \gamma \cdot \big(\frac{M}{r} - 1\big); \tag{3.10}$$

$$C_1 = \Gamma(c_1) \text{ for } c_1 = M - j + \gamma \cdot \big(\frac{M}{r} - 1\big); \tag{3.11}$$

$$C_2 = \Gamma(c_2) \text{ for } c_2 = M + \gamma \cdot \frac{M}{r}. \tag{3.12}$$

Since in our case some of the values are very large, we use intermediately a representation by the natural logarithm (as a built-in procedure for $\Gamma(x)$), i.e. in the second step we calculate

$$Z = \ln\Big(\binom{M}{j} \cdot \frac{A_1}{A_2} \cdot \frac{B_1}{B_2} \cdot \frac{C_1}{C_2}\Big) \tag{3.13}$$

$$= \ln\binom{M}{j} + \ln A_1 + \ln B_1 + \ln C_1 - \ln A_2 - \ln B_2 - \ln C_2. \tag{3.14}$$

The $\ln\Gamma(x)$ are calculated by a built-in function, and for the binomial coefficient we use the formula

$$\ln\binom{M}{j} = \sum_{s=M-j+1}^{M} \ln s - \sum_{t=1}^{j} \ln t. \tag{3.15}$$

Finally, we set

$$\beta_{j,\gamma} = \frac{M}{r} \cdot e^Z, \tag{3.16}$$

which is used as a sub-routine in the search for optimum settings of (r, γ):

1. For a fixed $r \geq r_0$ we search for γ such that T_γ from (3.3) is minimised, i.e. Eqn. 3.3 and Eqn. 3.16 are repeatedly calculated for $\gamma \geq \gamma_0$ and $\gamma = \gamma + \delta$, until T_γ changes only marginally or increases above the minimum value obtained so far.
2. For r_0 and $r = r + \Delta \leq r_{\max}$, the triplets (r, γ, T_γ) are recorded and finally r_{appr} at the inflection point of the graph of r against T_γ is selected.
3. The output is then determined by $N_{\text{appr}} = M/r_{\text{appr}}$.

In the computational experiments presented in the next section, we locate the inflection point in the graph of r against T_γ to identify a value for r. We search for r for 10 intervals in the range $[0.8 \times r_{\text{true}}, 1.2 \times r_{\text{true}}]$.

4. Numeric Results

4.1. Statistics over k-SAT instances

Table 1 gives data on the k-SAT instances. The columns report the minimum, maximum, mean value and standard deviation of the number of satisfied clauses; these values are averaged across the set of 20 instances. In Table 2, the columns for 'Local maxima' and 'Satisfies' do the same, but for the number of local maxima and number of instantiations which make the expression true, respectively; here, the minimum and maximum are the largest and smallest value in the set of instances, and the mean and standard deviation are computed across the set of instances.

TABLE 1. Summary statistics for varying n and m, with 20 instances for each value of m.

		Number of satisfied clauses			
n	m	min	max	mean	s.d.
18	66	45.3	65.8	57.8	2.7
18	71	50.1	70.9	62.1	2.8
18	76	53.6	75.6	66.5	2.9
18	81	57.6	80.8	70.9	3.0
18	86	61.3	85.3	75.3	3.0
20	74	51.5	73.9	64.8	2.8
20	79	55.4	78.9	69.1	2.9
20	84	59.4	83.9	73.5	3.0
20	89	62.8	88.5	77.9	3.2
20	94	67.1	93.3	82.2	3.3

Tables 1–2 provide evidence that the number of local maxima is on average relatively small and decreases significantly with increasing m for the remaining parameters being fixed. In Section 5 we attempt a theoretical explanation for this behaviour of the number of local maxima. In fact, the experimental observations seem to be counter-intuitive, since with increasing m and fixed n one moves from k-CNF that are satisfiable with high probability to conjunctive normal forms that are not satisfiable with high probability.

In particular, Table 2 displays a sharp decrease in the mean value of local maxima as well as for satisfying assignments for increasing m and fixed n. For each pair (n, m), 20 instances were generated and analysed.

In Table 3, the values of the relative proportion of the average number of global to local maxima is ordered in accordance with increasing values m/n. We recall that the critical threshold is $m/n = 4.23$. The values from the table suggest that there is a strong correlation (0.81, using Spearman's correlation) between increasing values of m/n and decreasing values of the relative proportion of global to local maxima.

Table 2. Local maxima and satisfying assignments.

		Local max				Satisfies		
n	m	min	max	mean	s.d.	max	mean	s.d.
18	66	21	221	64.2	44.9	221	39.1	50.2
18	71	13	78	33.4	17.7	67	19.1	17.1
18	76	10	196	49.4	42.8	17	5.2	5.7
18	81	7	86	37.3	22.1	35	8.1	9.6
18	86	17	73	34.6	15.2	14	1.4	3.1
20	74	29	162	77.1	39.6	131	37.4	36.1
20	79	7	154	57.9	41.3	52	41.6	46.0
20	84	11	244	51.5	49.5	39	8.1	9.3
20	89	14	119	46.6	27.3	49	9.3	13.0
20	94	12	104	48.5	26.8	42	6.0	11.0

Table 3. Relative proportion of global to local maxima.

n	m	Local max	Global max	m/n	Global/Local
18	66	64.2	39.1	3.67	0.61
20	74	77.1	37.4	3.70	0.48
18	71	33.4	19.1	3.94	0.57
20	79	57.9	41.6	3.95	0.72
20	84	51.5	8.1	4.20	0.16
18	76	49.4	5.2	4.22	0.10
20	89	46.6	9.3	4.45	0.20
18	81	37.3	8.1	4.50	0.22
20	94	48.5	6.0	4.70	0.12
18	86	34.6	1.4	4.78	0.04

4.2. Number of satisfied clauses

As can be seen in Tables 4– 5, the value for $\gamma(r)$ is quite small, 0.10. The value of N is within 15% of the true value in all cases, and frequently around 10%. The computed value for r varies considerably by landscape to reflect the changing number of local maxima and hence the shape of the landscape.

For each pair (n, m), 20 instances were generated, and for each of the instances the approximation procedure from Section 3 was executed for three different values of M. The values for N are from Table 2 for the corresponding n.

As displayed in Tables 4–5, the values of $\gamma(r_{\text{appr}})$ are very small, and the range is actually considered to be critical in [10]. Nevertheless, the deviation of approximations N_{appr} from the mean values N is below 15% for all three different values of M_i, and the typical value is in the region of 10%. We note that the small number of local maxima could explain the good performance of local search algorithms on k-SAT instances, see [1].

TABLE 4. n = 18: Results averaged over 20 landscapes.

m	N	M	β_M	$\gamma(r)$	r	N	N/N
66	64.2	2^9	46.95	0.10	10.19	72.14	1.11
66	64.2	2^{10}	51.60	0.10	20.15	73.20	1.12
66	64.2	2^{11}	55.45	0.10	40.29	73.20	1.12
71	33.4	2^9	26.55	0.10	18.35	36.35	1.08
71	33.4	2^{10}	27.70	0.10	36.50	36.57	1.08
71	33.4	2^{11}	28.95	0.10	72.42	37.13	1.10
76	49.4	2^9	34.20	0.10	16.62	55.48	1.10
76	49.4	2^{10}	37.55	0.10	33.15	55.76	1.10
76	49.4	2^{11}	40.30	0.10	66.15	55.83	1.10
81	37.3	2^9	28.70	0.10	20.06	41.08	1.09
81	37.3	2^{10}	30.55	0.10	40.09	41.30	1.10
81	37.3	2^{11}	31.70	0.10	79.63	41.75	1.10
86	34.6	2^9	25.05	0.10	16.22	37.77	1.09
86	34.6	2^{10}	27.00	0.10	32.31	37.91	1.09
86	34.6	2^{11}	28.40	0.10	63.93	38.35	1.10

TABLE 5. n = 20: Results averaged over 20 landscapes.

m	N	M	β_M	$\gamma(r)$	r	N	N/N
74	77.1	2^{10}	61.75	0.10	15.15	87.81	1.13
74	77.1	2^{11}	66.20	0.10	30.08	88.59	1.14
74	77.1	2^{12}	68.85	0.10	60.17	88.59	1.14
79	57.9	2^{10}	50.60	0.10	31.74	66.01	1.13
79	57.9	2^{11}	53.55	0.10	63.48	66.01	1.13
79	57.9	2^{12}	54.45	0.10	126.70	66.33	1.14
84	51.5	2^{10}	40.25	0.10	29.32	58.19	1.11
84	51.5	2^{11}	43.00	0.10	58.37	58.22	1.11
84	51.5	2^{12}	45.55	0.10	116.14	58.59	1.11
89	46.6	2^{10}	37.05	0.10	28.43	51.94	1.10
89	46.6	2^{11}	39.55	0.10	56.09	52.62	1.11
89	46.6	2^{12}	40.15	0.10	111.70	52.83	1.12
94	48.5	2^{10}	39.60	0.10	28.01	54.17	1.10
94	48.5	2^{11}	42.15	0.10	55.65	54.99	1.11
94	48.5	2^{12}	42.65	0.10	111.18	55.16	1.12

5. Local Maxima and k-CNF

For an arbitrary $\tilde{\sigma} \in \{0,1\}^n$ and $F \in \mathsf{F}_k(n,m)$, we set $\mathsf{C_0}(F,\tilde{\sigma}) = \{C | C \in \mathsf{C}(F) \wedge C(\tilde{\sigma}) = 0\}$ and $\mathsf{C_1}(F,\tilde{\sigma}) = \{C | C \in \mathsf{C}(F) \wedge C(\tilde{\sigma}) = 1\}$. Thus, clauses from $\mathsf{C_1}(F,\tilde{\sigma})$ have at least one literal among the k literals that returns 1 on $\tilde{\sigma}$. Since in (2.1) we have $d(\tilde{\sigma},\tilde{\sigma}')=1$, clauses with at least two literals returning 1 on $\tilde{\sigma}$ do not affect the re-calculation of Z_F in neighbourhood transitions out of $\tilde{\sigma}$. We therefore partition $\mathsf{C_1}(F,\tilde{\sigma})$ into $\mathsf{C_1^{(1)}}(F,\tilde{\sigma})$ and $\mathsf{C_1^{(\geq 2)}}(F,\tilde{\sigma})$,

i.e. $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$ contains all $C \in \mathsf{C}(F)$ with exactly one literal that returns 1 on $\tilde{\sigma}$. We note the following simple observation:

Lemma 5.1. *The truth assignment $\tilde{\sigma}$ is a one-step local maximum in $\mathsf{L}(F)$ iff for all $\tilde{\sigma}' \in \mathsf{N}(\tilde{\sigma})$:*

$$|\{C|C(\tilde{\sigma}')=1 \wedge C \in \mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})\}| \;\leq\; |\{C|C(\tilde{\sigma}')=0 \wedge C \in \mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})\}|. \quad (5.1)$$

For a literal x^η we use $x^\eta \in C$ to express that x^η is part of the disjunctive term C. Let $X_{\mathbf{0}}(F) = \{x|\exists C \in \mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma}) \wedge x^{\overline{\sigma}} \in C\}|$ and $p = |X_{\mathbf{0}}(F)|$ be the number of variables that occur in clauses of $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$, where we employ $\sigma^{\overline{\sigma}} \equiv 0$. Furthermore, we set $q = |\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})|$, $r = |\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})|$, and $s = |\mathsf{C}_{\mathbf{1}}^{(\geq 2)}(F,\tilde{\sigma})|$. Thus, we have for $F \in \mathsf{F}_k(n,m)$

$$m \;=\; q + r + s. \quad (5.2)$$

For $X_{\mathbf{1}}(F) = \{x|\exists C \in \mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma}) \wedge x^{\sigma} \in C\}$, $t = |X_{\mathbf{1}}(F)|$, and $h_u = |\{C|C \in \mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma}) \wedge x_{i_u}^{\sigma_{i_u}} \in C\}|$ we have

$$\sum_{u=1}^{t} h_u = r, \quad (5.3)$$

since the corresponding subsets of clauses have to be disjoint (otherwise, a clause from the intersection would belong to $\mathsf{C}_{\mathbf{1}}^{(\geq 2)}(F,\tilde{\sigma})$).

Lemma 5.2. *If $x_{i_u} \in X_{\mathbf{1}} \backslash X_{\mathbf{0}} \neq \emptyset$, then the neighbourhood transition that involves x_{i_u} diminishes $\mathsf{Z}_F(\tilde{\sigma})$ by h_u.*

This follows from the definitions of $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$ and $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$. For $f_u = |\{C|C \in \mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma}) \wedge x_{i_u}^{\overline{\sigma_{i_u}}} \in C\}|$, Lemma 5.1 can now be rewritten as

Lemma 5.3. *The truth assignment $\tilde{\sigma}$ is a one-step local maximum in $\mathsf{L}(F)$ iff $X_{\mathbf{0}} \subseteq X_{\mathbf{1}}$ and for $x_{i_u} \in X_{\mathbf{0}}$:*

$$1 \leq f_u \leq h_u. \quad (5.4)$$

We note that by definition

$$\sum_{u=1}^{p} f_u = q \cdot k, \quad (5.5)$$

and (5.3) and (5.4) imply for a one-step local maximum

$$q \cdot k \leq r. \quad (5.6)$$

The relations are illustrated by a small example for $n > 5$, $k = 3$, and $q = 3$, where only the first five variables are shown: The matrix in Table 6 shows the three 3-clauses $(x_1^{\overline{\sigma_1}} \vee x_4^{\overline{\sigma_4}} \vee x_5^{\overline{\sigma_5}})$, $(x_2^{\overline{\sigma_2}} \vee x_3^{\overline{\sigma_3}} \vee x_4^{\overline{\sigma_4}})$, and $(x_2^{\overline{\sigma_2}} \vee x_4^{\overline{\sigma_4}} \vee x_5^{\overline{\sigma_5}})$, which all return 0 on $(\sigma_1, ..., \sigma_5)$ and belong to the set $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$. For $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$ and, for example, x_2 we need at least two clauses each with $x_2^{\sigma_2}$ and $(k-1) = 2$ literals of type $x_j^{\overline{\sigma_j}}$, $j \neq 2$.

TABLE 6. Matrix with "column sums" k and "row sums" f_u.

x	assignment			f-values
x_1	$\overline{\sigma_1}$	0	0	$f_1 = 1$
x_2	0	$\overline{\sigma_2}$	$\overline{\sigma_2}$	$f_2 = 2$
x_3	0	$\overline{\sigma_3}$	0	$f_3 = 1$
x_4	$\overline{\sigma_4}$	$\overline{\sigma_4}$	$\overline{\sigma_4}$	$f_4 = 3$
x_5	$\overline{\sigma_5}$	0	$\overline{\sigma_5}$	$f_5 = 2$

In Table 7, elements of $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$ are represented by columns no 5 until no 8. In column no 5, both remaining $x_j^{\overline{\sigma_j}}$ are drawn from the first five variables, i.e. the column represents $(x_1^{\sigma_1} \vee x_3^{\overline{\sigma_3}} \vee x_5^{\overline{\sigma_5}})$. We note that the selection of $x_j^{\overline{\sigma_j}}$ is independent of the set of $(k-1)$ variables defining the corresponding clause from $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$. In column no 6, one one of the $x_j^{\overline{\sigma_j}}$ belongs to the first five variables, and none of the $x_j^{\overline{\sigma_j}}$ is chosen from this subset of variables in column no 7 and no 8.

TABLE 7. Matrix representing $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$ and part of $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$.

x	$\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$			$\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$					h-values
x_1	$\overline{\sigma_1}$	0	0	σ_1	0	0	0	$\cdots$	$h_1 = 1 = f_1$
x_2	0	$\overline{\sigma_2}$	$\overline{\sigma_2}$	0	σ_2	σ_2	σ_2	$\cdots$	$h_2 = 3 > f_2$
x_3	0	$\overline{\sigma_3}$	0	$\overline{\sigma_3}$	0	0	0	$\cdots$	$\cdots$
x_4	$\overline{\sigma_4}$	$\overline{\sigma_4}$	$\overline{\sigma_4}$	0	$\overline{\sigma_4}$	0	0	$\cdots$	$\cdots$
x_5	$\overline{\sigma_5}$	0	$\overline{\sigma_5}$	$\overline{\sigma_5}$	0	0	0	$\cdots$	$\cdots$

Let $\mathsf{M}_k^{\tilde{\sigma}}(n,m) \subseteq \mathsf{F}_k(n,m)$ denote the set of k-CNF that have $\tilde{\sigma}$ as a one-step local maximum for the neighbourhood defined by $\mathsf{N}(\tilde{\sigma})$ and the objective function defined by Z_F, where we require $\mathsf{Z}_F(\tilde{\sigma}) < m$ ($\tilde{\sigma}$ is a *local* maximum), i.e. $q \geq 1$ and $\tilde{\sigma}$ is not a satisfying assignment.

We are now going to derive an upper bound for $M_{\tilde{\sigma}} = |\mathsf{M}_k^{\tilde{\sigma}}(n,m)|$. As will be seen later, the ratio $2^n \cdot M_{\tilde{\sigma}}/|\mathsf{F}_k(n,m)|$, when approximated by using an upper bound of $M_{\tilde{\sigma}}$, then provides some information about typical values for the number of one-step local maxima for k-CNF in terms of parameters (k,n,m).

For fixed (q,r,s), we consider the number of potential sets $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$, $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$, and $\mathsf{C}_{\mathbf{1}}^{(2)}(F,\tilde{\sigma})$ under the assumption that the fixed truth assignment $\tilde{\sigma}$ is a one-step local maximum.

We set

$$\mathsf{C}_{\mathbf{0}}(\tilde{\sigma}) \;=\; \bigcup_{F \in \mathsf{M}_k^{\tilde{\sigma}}(n,m)} \mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma}); \tag{5.7}$$

$$\mathsf{C}_{\mathbf{1}}^{(1)}(\tilde{\sigma}) \;=\; \bigcup_{F \in \mathsf{M}_k^{\tilde{\sigma}}(n,m)} \mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma}); \tag{5.8}$$

$$\mathsf{C}_{\mathbf{1}}^{(2)}(\tilde{\sigma}) \quad = \quad \bigcup_{F \in \mathsf{M}_k^{\tilde{\sigma}}(n,m)} \mathsf{C}_{\mathbf{1}}^{(2)}(F, \tilde{\sigma}). \tag{5.9}$$

Note: here we define sets of sets of clauses, and subsets of the sets corresponding to fixed (q, r, s) are indicated by an index.

For a fixed $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$ we have to ensure that each of the p elements of $X_{\mathbf{0}}(F)$ is present in at least one of the clauses from $\mathsf{C}_{\mathbf{0}}$, and we therefore need

$$q \cdot k \geq p \geq k \;\text{ and }\; \binom{p}{k} \geq q. \tag{5.10}$$

Let $A(p, q)$ denote the number of pairwise different sets H of size q consisting of k-selections $S = \{x_{i_1}, ..., x_{i_k}\}$ out of p variables of $X_{\mathbf{0}}(F)$ such that $\forall x\big(x \in X_{\mathbf{0}} \rightarrow \exists S(S \in \mathsf{H} \wedge x \in S)\big)$. Since the q selected clauses might depend on a smaller number $p' < p$ of variables, we have

$$A(p, q) \leq \binom{\binom{p}{k}}{q} \tag{5.11}$$

to upper bound the number of sets $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$ depending on p variables, and for fixed q obviously $|\mathsf{C}_{\mathbf{0}}(\tilde{\sigma})_q| = A(n, q)$.

The selection of $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$ out of $A(p, q) \leq \binom{\binom{p}{k}}{q}$ candidates implies further conditions on $\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})$ and the associated set $X_{\mathbf{1}}$: for f_u clauses from $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$ with $x_{i_u}^{\overline{\sigma_{i_u}}}$ we have $h_u \geq f_u$ clauses from $\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})$ with $x_{i_u}^{\sigma_{i_u}}$, if $\tilde{\sigma}$ is a one-step local maximum. In each of the h_u clauses, the literals different from $x_{i_u}^{\sigma_i}$ are of the same type $x_j^{\overline{\sigma_j}}$ as in $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$, due to the definition of $\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})$.

There are $\binom{n}{k}$ k-clauses that return the value 0 on $(\sigma_1, ..., \sigma_n)$. Each of the k positions in the $\binom{n}{k}$ k-clauses can be altered from $x_{i_u}^{\overline{\sigma_{i_u}}}$ to $x_{i_u}^{\sigma_{i_u}}$ in order to generate a candidate for $\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})$. Therefore, given $r = k \cdot q + \Delta$ and $\Delta \geq 0$, the number of sets $|\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})_r|$ consisting of r clauses is given by

$$|\mathsf{C}_{\mathbf{1}}^{(1)}(F, \tilde{\sigma})_r| = B(n, r) = \binom{k \cdot \binom{n}{k}}{r}. \tag{5.12}$$

Finally, we consider for $\mathsf{C}_{\mathbf{1}}^{(\geq 2)}(\tilde{\sigma})$ the set of all $\binom{n}{k} \cdot 2^k$ clauses: since $\tilde{\sigma}$ is fixed, among the set of all clauses there are $\binom{n}{k}$ k-clauses that return 0 on $\tilde{\sigma}$ (the clauses of $\mathsf{C}_{\mathbf{0}}(\tilde{\sigma})$ are drawn from this subset); as mentioned before, there are $k \cdot \binom{n}{k}$ k-clauses with exactly one literal of type $x_{i_u}^{\sigma_{i_u}}$ (the clauses of $\mathsf{C}_{\mathbf{1}}^{(1)}(\tilde{\sigma})$ are drawn from this subset). Therefore, the number of different sets $|\mathsf{C}_{\mathbf{1}}^{(\geq 2)}(\tilde{\sigma})_s|$ consisting of sets of s clauses is given by

$$|\mathsf{C}_{\mathbf{1}}^{(\geq 2)}(\tilde{\sigma})_s| = C(n, s) = \binom{(2^k - k - 1) \cdot \binom{n}{k}}{s}. \tag{5.13}$$

Apart from $s = m - q - r$, no further restrictions apply to $C(n, s)$. Therefore, we focus on $A(p, q)$ and $B(n, r)$.

Due to one of the Vandermonde identities, namely $\sum_{c=0}^{h}\binom{a}{c}\cdot\binom{b-a}{h-c}=\binom{b}{h}$, the (partial) summation over products $A(p,q)\cdot B(n,r)\cdot C(n,s)$ results in an upper bound below but close to $\binom{2^k\cdot\binom{n}{k}}{m}$, which can be immediately seen from a simplified version $A(n,q)\cdot B(n,r)\cdot C(n,s)$.

We first identify the range of q and r where the summands in the following simplified upper bound for $M_{\tilde{\sigma}}=|\mathsf{M}_k^{\tilde{\sigma}}(n,m)|$ are relatively small:

$$M_{\tilde{\sigma}} \;<\; \sum_{q=1}^{\lfloor m/(k+1)\rfloor}\;\sum_{r=k\cdot q}^{m-q} A(n,q)\cdot B(n,r)\cdot C(n,m-q-r). \tag{5.14}$$

The upper bound will be improved in successive steps.

Since we are interested in the average number of k-CNF having $\widetilde{\sigma}$ as a one-step local maximum, we introduce the inverse value of $\binom{2^k\cdot\binom{n}{k}}{m}$ and set

$$P(q,r) \;=\; \frac{A(n,q)\cdot B(n,r)\cdot C(n,m-q-r)}{\binom{2^k\cdot\binom{n}{k}}{m}};\; D=\binom{n}{k}. \tag{5.15}$$

In the sum

$$\sum_{r=k\cdot q}^{m-q} P(q,r) \;=\; \sum_{r=k\cdot q}^{m-q}\frac{A(n,q)\cdot B(n,r)\cdot C(n,m-q-r)}{\binom{2^k\cdot\binom{n}{k}}{m}} \tag{5.16}$$

we analyse a single summand, which is given by

$$\begin{aligned}
&\frac{\binom{D}{q}\cdot\binom{k\cdot D}{r}\cdot\binom{(2^k-k-1)\cdot D}{m-q-r}}{\binom{2^k\cdot D}{m}}\\
=\;& \frac{D\cdots(D-q+1)\cdot k\cdot D\cdots(k\cdot D-r+1)}{q!\cdot r!}\times\\
&\times\frac{(2^k-k-1)\cdot D\cdots((2^k-k-1)\cdot D-m+q+r+1)\cdot m!}{(m-q-r)!\cdot 2^k\cdot D\cdots(2^k\cdot D-m+1)}\\
=\;& \binom{m}{q}\cdot\binom{m-q}{r}\cdot D\cdots(D-q+1)\cdot k\cdot D\cdots(k\cdot D-r+1)\times\\
&\times\frac{(2^k-k-1)\cdot D\cdots((2^k-k-1)\cdot D-m+q+r+1)}{2^k\cdot D\cdots(2^k\cdot D-m+1)}.
\end{aligned} \tag{5.17}$$

The factors D, $k\cdot D$, $(2^k-k-1)\cdot D$, and $2^k\cdot D$ are taken out and (5.17) turns to

$$\begin{aligned}
&\frac{\binom{D}{q}\cdot\binom{k\cdot D}{r}\cdot\binom{(2^k-k-1)\cdot D}{m-q-r}}{\binom{2^k\cdot D}{m}}\\
=\;& \frac{\binom{m}{q}\cdot\binom{m-q}{r}}{2^{k\cdot q}}\cdot\Big(\frac{k}{2^k}\Big)^r\cdot\Big(\frac{2^k-k-1}{2^k}\Big)^{m-q-r}\cdot\mathsf{Z},
\end{aligned} \tag{5.18}$$

where

$$\mathsf{Z} \;=\; \frac{\prod\limits_{u=1}^{q-1}\left(1-\frac{u}{D}\right)\cdot\prod\limits_{u=1}^{r-1}\left(1-\frac{u}{k\cdot D}\right)\cdot\prod\limits_{u=1}^{m-q-r-1}\left(1-\frac{u}{(2^k-k-1)\cdot D}\right)}{\prod\limits_{u=1}^{m-1}\left(1-\frac{u}{2^k\cdot D}\right)}. \quad (5.19)$$

The products have the same structure and we apply $(1+z)^z < e$ as well as $(1+1/z)^{z+1} > e$. This way we obtain:

$$\mathsf{Z} \;<\; \frac{e^{\sum\limits_{u=1}^{m-1}\frac{u}{2^k\cdot D-u}}}{e^{\sum\limits_{u=1}^{q-1}\frac{u}{D}}\cdot e^{\sum\limits_{u=1}^{r-1}\frac{u}{k\cdot D}}\cdot e^{\sum\limits_{u=1}^{m-q-r-1}\frac{u}{(2^k-k-1)\cdot D}}}. \quad (5.20)$$

We introduce the condition

$$m \le \sqrt{2^{k+1}\cdot\binom{n}{k}}, \quad (5.21)$$

which roughly means that m is upper bounded by $\sim n^{k/2}$. From (5.21) we have

$$\begin{aligned}\mathsf{Z} \;&<\; \frac{e^{\sum\limits_{u=1}^{m-1}\frac{u}{(2^k-1)\cdot D}}}{e^{\sum\limits_{u=1}^{q-1}\frac{u}{D}}\cdot e^{\sum\limits_{u=1}^{r-1}\frac{u}{k\cdot D}}\cdot e^{\sum\limits_{u=1}^{m-q-r-1}\frac{u}{(2^k-k-1)\cdot D}}}\\ &=\; \frac{e^{\frac{m\cdot(m-1)}{2\cdot(2^k-1)\cdot D}}}{e^{\frac{q\cdot(q-1)}{2\cdot D}}\cdot e^{\frac{r\cdot(r-1)}{2\cdot k\cdot D}}\cdot e^{\frac{(m-q-r)\cdot(m-q-r-1)}{2\cdot(2^k-k-1)\cdot D}}}\\ &<\; e, \end{aligned} \quad (5.22)$$

and (5.18) turns to

$$\begin{aligned}\frac{\binom{D}{q}\cdot\binom{k\cdot D}{r}\cdot\binom{(2^k-k-1)\cdot D}{m-q-r}}{\binom{2^k\cdot D}{m}} \;&<\; \frac{e\cdot\binom{m}{q}}{2^{k\cdot q}}\cdot\binom{m-q}{r}\cdot\left(\frac{k}{2^k}\right)^r\times\\ &\quad\times\left(\frac{2^k-k-1}{2^k}\right)^{m-q-r}. \end{aligned} \quad (5.23)$$

Therefore, by taking into account the full binomial sum, (5.16) and (5.23) lead to:

$$\begin{aligned}\frac{\sum\limits_{r=k\cdot q}^{m-q}P(q,r)}{e\cdot\binom{m}{q}\cdot 2^{-k\cdot q}} \;&<\; \left(\frac{k}{2^k}+\frac{2^k-k-1}{2^k}\right)^{m-q}-\\ &\quad-\sum_{r=0}^{k\cdot q-1}\binom{m-q}{r}\cdot\left(\frac{k}{2^k}\right)^r\cdot\left(1-\frac{k+1}{2^k}\right)^{m-q-r}, \end{aligned}$$

and we finally have

$$\frac{\sum_{r=k\cdot q}^{m-q} P(q,r)}{e\cdot\binom{m}{q}\cdot 2^{-k\cdot q}} < \left(1-\frac{1}{2^k}\right)^{m-q} - \\ -\sum_{r=0}^{k\cdot q-1}\binom{m-q}{r}\cdot\left(\frac{k}{2^k}\right)^r\cdot\left(1-\frac{k+1}{2^k}\right)^{m-q-r}. \quad (5.24)$$

Let $E(m-q)$ be defined as

$$E(m-q) = \sum_{r=0}^{k\cdot q-1}\binom{m-q}{r}\cdot\left(\frac{k}{2^k}\right)^r\cdot\left(1-\frac{k+1}{2^k}\right)^{m-q-r}, \quad (5.25)$$

and (5.24) can then be written as

$$\frac{\sum_{r=k\cdot q}^{m-q} P(q,r)}{e\cdot\binom{m}{q}\cdot 2^{-k\cdot q}} < \left(1-\frac{1}{2^k}\right)^{m-q} - E(m-q). \quad (5.26)$$

Based on (5.16) and (5.23) until (5.26), we analyse the upper bound

$$M_{\tilde{\sigma}} < e\cdot\binom{2^k\cdot\binom{n}{k}}{m}\cdot\sum_{q=1}^{\lfloor m/(k+1)\rfloor}\frac{\binom{m}{q}}{2^{k\cdot q}}\cdot\left\{\left(1-\frac{1}{2^k}\right)^{m-q} - E^{m-q}\right\} \\ < e\cdot\binom{2^k\cdot\binom{n}{k}}{m}\cdot\sum_{q=1}^{\lfloor m/(k+1)\rfloor}\binom{m}{q}\cdot\left(\frac{1}{2^k}\right)^q\cdot\left(1-\frac{1}{2^k}\right)^{m-q}. \quad (5.27)$$

If the factor e is discarded and the summand for $q=0$ is added, the sum on the RHS of (5.27) can be treated as a Poisson process and Chernoff bounds [13] can be applied: Let X_i denote independent random variables and $\Pr[X_i = 1] = \frac{1}{2^k}$ and $\Pr[X_i = 0] = 1-\frac{1}{2^k}$. For $X = \sum_{i=1}^{r} X_i$ we then have

$$\Pr[X = q] = \binom{m}{q}\cdot\left(\frac{1}{2^k}\right)^v\cdot\left(1-\frac{1}{2^k}\right)^{m-q}, \quad (5.28)$$

which represents exactly summands on the RHS of (5.27). Thus, depending on the relative position of q to the expected value $\mu = E[X] = m/2^k$, one can apply Chernoff bounds for the lower and upper tail, respectively:

$$\Pr[X < (1-\delta)\cdot\mu] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} < e^{-\frac{\delta^2}{2}\cdot\mu} \quad (5.29)$$

$$\Pr[X > (1+\delta)\cdot\mu] < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} < e^{-\frac{\delta^2}{3}\cdot\mu}. \quad (5.30)$$

For $q_1 = m/2^k - h_1 = (1-\delta_1)\cdot m/2^k$ and $q_2 = m/2^k + h_2 = (1+\delta_2)\cdot m/2^k$ we therefore obtain:

$$\Pr[X < (1-\delta_1)\cdot\frac{m}{2^k}] \;<\; e^{-\frac{h_1^2}{2}\cdot\frac{2^k}{m}} \tag{5.31}$$

$$\Pr[X > (1+\delta_2)\cdot\frac{m}{2^k}] \;<\; e^{-\frac{h_2^2}{3}\cdot\frac{2^k}{m}}. \tag{5.32}$$

Thus, if $h_1, h_2 >> \sqrt{m/2^k}$, we obtain exponentially small upper bounds. In order to fix the values, we set

$$h_1 \;=\; \lfloor\frac{1}{2}\cdot\frac{m}{2^k}\rfloor \text{ and } h_2 \;=\; \lceil\frac{1}{2}\cdot\frac{m}{2^k}\rceil; \tag{5.33}$$

$$q_1 \;=\; \lfloor\frac{m}{2^{k+1}}\rfloor \text{ and } q_2 \;=\; \lceil 3\cdot\frac{m}{2^{k+1}}\rceil, \tag{5.34}$$

and we finally have

$$\Pr[X < (1-\delta_1)\cdot\frac{m}{2^k}] \;<\; e^{-\frac{1}{8}\cdot\frac{m}{2^k}} \tag{5.35}$$

$$\Pr[X > (1+\delta_2)\cdot\frac{m}{2^k}] \;<\; e^{-\frac{1}{12}\cdot\frac{m}{2^k}}. \tag{5.36}$$

In the same way we can analyse a modified upper bound of (5.16) and (5.24), where we incorporate (5.23):

$$\begin{aligned}\frac{\sum_{r=k\cdot q}^{m-q} P(q,r)}{e\cdot\binom{m}{q}\cdot 2^{-k\cdot q}} &< \sum_{r=1}^{m-q}\binom{m-q}{r}\cdot\Big(\frac{k}{2^k}\Big)^r\cdot\Big(1-\frac{k+1}{2^k}\Big)^{m-q-r}\\ &< \sum_{r=1}^{m-q}\binom{m-q}{r}\cdot\Big(\frac{k}{2^k}\Big)^r\cdot\Big(1-\frac{k}{2^k}\Big)^{m-q-r}.\end{aligned} \tag{5.37}$$

Again, the RHS is treated as a Poisson process with $\mu = E[X] = (m-q)\cdot k/2^k$. For $r_1 = (m-q)\cdot k/2^k - h_1' = (1-\delta_1')\cdot(m-q)\cdot k/2^k$ and $q_2 = (m-q)\cdot k/2^k + h_2' = (1+\delta_2')\cdot(m-q)\cdot k/2^k$ we obtain

$$\Pr[X < (1-\delta_1')\cdot\frac{k}{2^k}\cdot(m-q)] \;<\; e^{-\frac{1}{8}\cdot\frac{k}{2^k}\cdot(m-q)} \tag{5.38}$$

$$\Pr[X > (1+\delta_2')\cdot\frac{k}{2^k}\cdot(m-q)] \;<\; e^{-\frac{1}{12}\cdot\frac{k}{2^k}\cdot(m-q)}, \tag{5.39}$$

where we use the setting

$$h_1' \;=\; \lfloor\frac{1}{2}\cdot\frac{k}{2^k}\cdot(m-q)\rfloor \text{ and } h_2' \;=\; \lceil\frac{1}{2}\cdot\frac{k}{2^k}\cdot(m-q)\rceil; \tag{5.40}$$

$$r_1(q) \;=\; \lfloor\frac{k}{2^{k+1}}\cdot(m-q)\rfloor \text{ and } r_2(q) \;=\; \lceil 3\cdot\frac{k}{2^{k+1}}\cdot(m-q)\rceil, \tag{5.41}$$

Finally, (5.27) can be simplified to

$$\begin{aligned} M_{\tilde{\sigma}} &< e\cdot\binom{2^k\cdot\binom{n}{k}}{m}\cdot\Big\{\frac{1}{e^{\frac{1}{8}\cdot\frac{m}{2^k}}}+\frac{1}{e^{\frac{1}{12}\cdot\frac{m}{2^k}}}+\sum_{q=q_1}^{q_2}\Big(\frac{1}{e^{\frac{1}{8}\cdot\frac{k}{2^k}\cdot(m-q)}}+\frac{1}{e^{\frac{1}{12}\cdot\frac{k}{2^k}\cdot(m-q)}}\Big)+ \\ &\quad +\sum_{q=q_1}^{q_2}\sum_{r=r_1(q)}^{r_2(q)}P(q,r)\Big\} \\ &< e\cdot\binom{2^k\cdot\binom{n}{k}}{m}\cdot\Big\{\frac{2}{e^{\frac{1}{12}\cdot\frac{m}{2^k}}}+\frac{20\cdot 2^k}{k}\cdot\frac{1}{e^{\frac{1}{12}\cdot\frac{k}{2^k}\cdot(m-q_2)}}+\sum_{q=q_1}^{q_2}\sum_{r=r_1(q)}^{r_2(q)}P(q,r)\Big\} \\ &< e\cdot\binom{2^k\cdot\binom{n}{k}}{m}\cdot\Big\{\frac{21\cdot 2^k}{k}\cdot\frac{1}{e^{\frac{1}{12}\cdot\frac{m}{2^k}}}+\sum_{q=q_1}^{q_2}\sum_{r=r_1(q)}^{r_2(q)}P(q,r)\Big\}. \end{aligned} \tag{5.42}$$

We now focus on improved upper bounds with respect to $A(n,q)\cdot B(n,r)\cdot C(n,s)$ for the range of values $q_1\le q\le q_1$ and $r_1(q)\le r\le r_2(q)$. Since any subset of clauses counted by $C(n,s)$ can be combined with clauses counted by $A(p,q)$ and $B(n,r)$, an improvement significantly below $A(n,q)\cdot B(n,r)\cdot C(n,s)$ has to come from a detailed analysis of the combination of clauses counted by $A(p,q)$ and $B(n,r)$. A natural way would be to look at binary matrices derived from representations as presented in Table 6 and Table 7. By definition, each of the $B(n,r)$ selections $\mathsf{S}_r=\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$ of r k-clauses is characterized by the subset $X_{\mathbf{1}}(F)\subseteq\{x_1,...,x_n\}$ of variables that appear as literals x^σ in disjunctive clauses of S_r, whereas $x^{\overline{\sigma}}$ is element of at least one clause in some $\mathsf{S}_q=\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$. We note that a single S_r can be combined with several S_q: the clauses in S_r can be "ordered" and counted with respect to each of the x^σ, and for the t variables from $X_{\mathbf{1}}$ we then have $\sum_{u=1}^t h_u=r$ for $h_u\ge 1$, $u=1,...,t$, see (5.3); out of the values h_u, i.e. the vector $[h_1,...,h_t]$, $p\le t$ values $f_v\le h_v$ can be chosen such that $\sum_{v=1}^p f_v=k\cdot q$ and $f_v\ge 1$, $v=1,...,p$; each of the vectors $[f_1,...,f_p]$ then defines the row numbers and the "row sum values" for a binary matrix (if there is a binary matrix for this vector). The set of sets of k-clauses S_q is determined by the number of solutions as binary matrices with a row sum vector $[f_1,...,f_p]$ and column sums equal to k, subject to permutations of columns.

Thus, a potential way would be to take an r-selection of clauses counted by $B(n,r)$ and to multiply $B(n,r)$ by the number of binary matrices that produce a fixed row sum $[f_1,...,f_p]$, $\sum_{v=1}^p f_v=kq$, with all column sums equal to k, i.e. here one would "count from $\mathsf{C}_{\mathbf{1}}^{(1)}(F,\tilde{\sigma})$ to $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$." Unfortunately, there are no tight upper bounds that improve on $A(n,q)$. The study of such matrices has a long history, but asymptotic upper bounds refer to enumerated rows and columns under restrictions on maximal row and column sum values (sparse matrices and almost square matrices), see the fundamental papers [6, 8] that utilise sophisticated combinatorial methods.

Therefore, we choose a second way, where we consider all $A(n,q)$ selections of sets $\mathsf{C}_{\mathbf{0}}(F,\tilde{\sigma})$ individually, with each selection producing a row sum

vector $[f_1, ..., f_p]$, i.e. "counting from $\mathsf{C}_{\mathbf{0}}(F, \tilde{\sigma})$ to $\mathsf{C}^{(1)}_{\mathbf{1}}(F, \tilde{\sigma})$." The next step is to identify the number of $\mathsf{C}^{(1)}_{\mathbf{1}}(F, \tilde{\sigma})$ of size r that can be combined with $[f_1, ..., f_p]$.

Given $\widetilde{f} = [f_1, ..., f_p]$, $\sum_{v=1}^{p} f_v = k \cdot q$ and $\widetilde{f}$ induced by a fixed element S_q (which is a set of q k-clauses) of $\mathsf{C}_{\mathbf{0}}(\tilde{\sigma})_q$, we denote by $X_{\mathbf{0}}(\mathsf{S}_q)$ the set $\{x_{i_1}, ..., x_{i_p}\}$ of variables that appear as literals in S_q (similar to $X_{\mathbf{0}}(F)$ defined before, but here for selections out of $\mathsf{C}_{\mathbf{0}}(\tilde{\sigma})_q$).

Furthermore, let $E(q, r, \widetilde{f})$ denote the number of all elements $\mathsf{S}_r \in \mathsf{C}^{(1)}_{\mathbf{1}}(\tilde{\sigma})_r$ such that for f_v occurrences of $x^{\overline{\sigma}}_{i_v}$, $x_{i_v} \in X_{\mathbf{0}}(\mathsf{S}_q)$, the number of occurrences h_v of $x^{\sigma}_{i_v}$ in S_r satisfies $h_v \geq f_v$, $v = 1, ..., p \leq n$.

By using these notations, we then have

$$E(q, r, \widetilde{f}) \quad \leq \quad \sum_{a=0}^{r-k\cdot q} \binom{p \cdot \binom{n-1}{k-1}}{a + k\cdot q} \cdot \binom{(n-p) \cdot \binom{n-1}{k-1}}{r-a}. \tag{5.43}$$

Indeed, the first factors ensures that at least $k \cdot q$ elements are drawn for variables from $X_{\mathbf{0}}(\mathsf{S}_q)$ (all $r \geq k \cdot q$ elements are not excluded), and the second factor makes sure that the upper bound is not further overestimated by restricting the selections to variables outside $X_{\mathbf{0}}(\mathsf{S}_q)$. Here, we employ that for a fixed $x^{\sigma}_{i_v}$, there are $\binom{n-1}{k-1}$ different ways to select $x^{\overline{\sigma}}_{i_u}$, $u \neq v$, for k-clauses of S_r. We note that $\binom{n-1}{k-1} = \frac{k}{n} \cdot \binom{n}{k}$ and therefore $p \cdot \binom{n-1}{k-1} + (n-p) \cdot \binom{n-1}{k-1} = k \cdot \binom{n}{k}$, as has been used in (5.12).

Obviously, (5.43) overestimates $E(q, r, \widetilde{f})$, and, based on the Vandermonde identity, for small q

$$\sum_{a=0}^{r-k\cdot q} \binom{p \cdot \binom{n-1}{k-1}}{a + k\cdot q} \cdot \binom{(n-p) \cdot \binom{n-1}{k-1}}{r-a} \quad \sim \quad \binom{k \cdot \binom{n}{k}}{r} = B(n, r), \tag{5.44}$$

which leads back to $A(n, q) \cdot B(n, r) \cdot C(n, s)$. The problem is produced by $\binom{p \cdot \binom{n-1}{k-1}}{a+k\cdot q}$, since it relates to all partitions into p summands generating $k \cdot q$, and not to a particular $\widetilde{f}$. Therefore, we finally focus on an improvement of this particular upper bound.

Let $G(q, p, a, \widetilde{f})$ denote the number of all elements $\mathsf{S}_b \in \mathsf{C}^{(1)}_{\mathbf{1}}(\tilde{\sigma})_b$ such that for f_v occurrences of $x^{\overline{\sigma}}_{i_v}$, $x_{i_v} \in X_{\mathbf{0}}(\mathsf{S}_q)$ and $v = 1, ..., p \leq n$:

1. the number of occurrences $g_v + f_v$ of $x^{\sigma}_{i_v}$ in S_b satisfies $g_v \geq 0$;
2. $b = \sum_{v=1}^{p} (g_v + f_v) = a + k \cdot q$;
3. only $x^{\sigma}_{i_v}$ with $x_{i_v} \in X_{\mathbf{0}}(\mathsf{S}_q)$ are literals in clauses of sets S_b consisting of b k-clauses.

For $G(q, p, a, \widetilde{f})$ we then have

$$G(q, p, a, \widetilde{f}) \quad = \quad \sum_{g_1 + \cdots + g_v = a} \prod_{v=1}^{p} \binom{\frac{k}{n} \cdot \binom{n}{k}}{g_v + f_v}. \tag{5.45}$$

Since we are interested in the improvement over $\binom{p\cdot\binom{n-1}{k-1}}{a+k\cdot q}$, we analyse the following ratio:

$$Q(q,p,a) \;=\; \max_{\widetilde{f}} \frac{G(q,p,a,\widetilde{f})}{\binom{p\cdot\binom{n-1}{k-1}}{a+k\cdot q}} \;\leq\; \frac{\sum\limits_{g_1+\cdots+g_v=a}\prod\limits_{v=1}^{p}\binom{\frac{k}{n}\cdot\binom{n}{k}}{g_v+\lceil\frac{k\cdot q}{p}\rceil}}{\binom{p\cdot\binom{n-1}{k-1}}{a+k\cdot q}}, \quad (5.46)$$

where we employ that $x>y$ implies $\binom{A}{x}\cdot\binom{A}{y}\leq\binom{A}{x-1}\cdot\binom{A}{y+1}$. We note that

$$\begin{aligned}
\prod_{v=1}^{p}\binom{\frac{k}{n}\cdot\binom{n}{k}}{g_v+f_v} &= \prod_{v=1}^{p}\frac{1}{(g_v+f_v)!}\cdot\prod_{v=1}^{p}\frac{k}{n}\cdot D\cdot\big(\frac{k}{n}\cdot D-1\big)\cdots\big(\frac{k}{n}\cdot D-g_v-f_v+1\big)\\
&= \prod_{v=1}^{p}\frac{1}{(g_v+f_v)!}\cdot\Big(\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot\prod_{v=1}^{p}\prod_{u=1}^{g_v+f_v-1}\Big(1-\frac{u}{\frac{k}{n}\cdot D}\Big)\\
&< \prod_{v=1}^{p}\frac{1}{(g_v+f_v)!}\cdot\Big(\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot\prod_{v=1}^{p}e^{-\sum_{u=1}^{g_v+f_v-1}\frac{u}{\frac{k}{n}\cdot D}}\\
&= \prod_{v=1}^{p}\cdot D\Big)^{a+k\cdot q}\frac{1}{(g_v+f_v)!}\cdot\prod_{v=1}^{p}e^{-\frac{(g_v+f_v)\cdot(g_v+f_v-1)}{\frac{2\cdot k}{n}\cdot D}}\\
&< \Big(\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot\prod_{v=1}^{p}\frac{1}{(g_v+f_v)!}. \qquad (5.47)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\left\{\binom{p\cdot\binom{n-1}{k-1}}{a+k\cdot q}\right\}^{-1} &= \left\{\frac{1}{(a+k\cdot q)!}\cdot\frac{p\cdot k}{n}\cdot D\cdots\big(\frac{p\cdot k}{n}\cdot D-a-k\cdot q+1\big)\right\}^{-1}\\
&= \left\{\frac{1}{(a+k\cdot q)!}\cdot\Big(p\cdot\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot\prod_{u=1}^{a+k\cdot q-1}\Big(1-\frac{u}{\frac{p\cdot k}{n}\cdot D}\Big)\right\}^{-1}\\
&< \left\{\frac{1}{(a+k\cdot q)!}\cdot\Big(p\cdot\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot e^{-\sum_{u=1}^{a+k\cdot q-1}\frac{u}{\frac{p\cdot k}{n}\cdot D-u}}\right\}^{-1}\\
&\leq \left\{\frac{1}{(a+k\cdot q)!}\cdot\Big(p\cdot\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot e^{-\sum_{u=1}^{a+k\cdot q-1}\frac{u}{\frac{p\cdot k}{n}\cdot D-m}}\right\}^{-1}\\
&\leq \left\{\frac{1}{(a+k\cdot q)!}\cdot\Big(p\cdot\frac{k}{n}\cdot D\Big)^{a+k\cdot q}\cdot e^{-\frac{(m)\cdot(m-1)}{2\cdot(\frac{p\cdot k}{n}\cdot D-m)}}\right\}^{-1}\\
&< e\cdot p^{-(a+k\cdot q)}\cdot\Big(\frac{k}{n}\cdot D\Big)^{-(a+k\cdot q)}\cdot(a+k\cdot q)!, \qquad (5.48)
\end{aligned}$$

where we utilise (5.21) again. We now have for (5.46):

$$Q(q,p,a) \;<\; e\cdot\frac{(a+k\cdot q)!}{p^{a+k\cdot q}}\cdot\sum_{g_1+\cdots+g_v=a}\prod_{v=1}^{p}\frac{1}{(g_v+f_p)!}, \qquad (5.49)$$

where for simplicity of notations we keep f_p for $f_p = \lceil \frac{k \cdot q}{p} \rceil$. The upper bound is further analysed later, where we apply a technique that is used several times for identifying regions of fast convergence.

Let $\widetilde{Q}(q,p,a)$ denote the value of the RHS of (5.49). From (5.45) and (5.46) we obtain for (5.43):

$$\begin{aligned}\sum_{\tilde{f}} E(q,r,\widetilde{f}) \quad &< \quad \sum_{\tilde{f}} \Big\{ \sum_{a=0}^{r-k\cdot q} \widetilde{Q}(q,p,a) \times \\ &\qquad \times \binom{p_{\tilde{f}} \cdot \binom{n-1}{k-1}}{a+k\cdot q} \cdot \binom{(n-p_{\tilde{f}}) \cdot \binom{n-1}{k-1}}{r-a} \Big\}. \end{aligned} \tag{5.50}$$

We recall that $\widetilde{f}$ is induced by one of the $A(n,q)$ selections of $\mathsf{C_0}(\tilde{\sigma})_q$. For $p \geq p_0$ (the value of p_0 is specified later) we set

$$\widetilde{Q}(q,r) \quad = \quad \begin{cases} 1, & \text{if } k\cdot q \leq a+k\cdot q < r_1(q); \\ \max \widetilde{Q}(q,p_{\tilde{f}},a) & \text{if } r_1(q) \leq a+k\cdot q \leq r_2(q); \\ 1, & \text{if } r_2(q) < k\cdot q \leq a+k\cdot q. \end{cases} \tag{5.51}$$

where r_1 and r_2 are from (5.40) and (5.41). We then have from (5.50):

$$\begin{aligned}\sum_{\tilde{f}} E(q,r,\widetilde{f}) \quad &< \quad \widetilde{Q}(q,r) \cdot \sum_{\tilde{f}} \sum_{a=0}^{r-k\cdot q} \binom{p_{\tilde{f}} \cdot \binom{n-1}{k-1}}{a+k\cdot q} \cdot \binom{(n-p_{\tilde{f}}) \cdot \binom{n-1}{k-1}}{r-a} \\ &\leq \quad \widetilde{Q}(q,r) \cdot \sum_{\tilde{f}} B(n,r) \cdot 1 \\ &\leq \quad \widetilde{Q}(q,r) \cdot A(n,q) \cdot B(n,r) \\ &= \quad A(n,q) \cdot \Big(\widetilde{Q}(q,r) \cdot B(n,r) \Big), \end{aligned} \tag{5.52}$$

where we emphasise the fact that through the ratio (5.46) we aim at a smaller value of $\Big(\widetilde{Q}(q,r) \cdot B(n,r) \Big)$ compared to $B(n,r)$. Therefore, we obtain

Lemma 5.4. *For fixed (q,r,s), the number of feasible pairs $[\mathsf{S}_q, \mathsf{S}_r]$ of sets of clauses from $\mathsf{C_0}(\tilde{\sigma})_q$ and $\mathsf{C}^{(1)}_{\mathbf{1}}(\tilde{\sigma})_r$, respectively, is upper bounded by $A(n,q) \cdot \Big(\widetilde{Q}(q,r) \cdot B(n,r) \Big)$.*

Based on (5.15), (5.42), and Lemma 5.4, we obtain for $M_{\tilde{\sigma}} = |\mathsf{M}^{\tilde{\sigma}}_k(n,m)|$ the upper bound

$$\begin{aligned} M_{\tilde{\sigma}} \quad &< \quad e \cdot \binom{2^k \cdot \binom{n}{k}}{m} \cdot \Big\{ \frac{21 \cdot 2^k}{k} \cdot \frac{1}{e^{\frac{1}{12} \cdot \frac{m}{2^k}}} + \\ &\qquad + \sum_{q=q_1}^{q_2} \sum_{r=r_1(q)}^{r_2(q)} A(n,q) \cdot \widetilde{Q}(q,r) \cdot B(n,r) \cdot C(n,m-q-r) \Big\}. \end{aligned} \tag{5.53}$$

However, the value of $\widetilde{Q}(q,p,a)$ from (5.51) that determines $\widetilde{Q}(q,r)$ is defined for $k \leq p \leq n$. Furthermore, in (5.52), the value of $\widetilde{Q}(q,r)$ is multiplied

by $A(n,q)$, based on the counting argument that a single selection of $\mathsf{S}_q \in \mathsf{C_0}(\tilde{\sigma})_q$ that produces $\widetilde{f}$ defined by p variables can be combined with at most $\widetilde{Q}(q,r) \cdot B(n,r)$ r-selections of type $\mathsf{S}_r \in \mathsf{C_1^{(1)}}(\tilde{\sigma})_r$. But, how many of the potentially $A(n,q)$ selections of S_q can produce $\widetilde{f}$ depending only on $p < n$ variables? Therefore, we are now going to specify p_0 from (5.51).

Let $H(p,q)$ denote the number of pairwise different S_q that induce $\widetilde{f}$ depending exactly on $p < n$ variables. We now proceed in two steps: (i) we try to identify p_0 such that $\sum_{p=k}^{p_0} H(p,q)$ is small compared to $A(n,q)$; (ii) the value of $\widetilde{Q}(q,p,a)$ from (5.51) is further analysed later only for $p_0 < p \leq n$.

We recall that $k \leq p \leq k \cdot q$ according to (5.10) and $q \leq \binom{p}{k}$. We set $X = \{x_1, ..., x_n\}$ and $X_P = \{x_{i_1}, ..., x_{i_p}\}$. Let $H(X_p, q)$ denote the number of S_q that depend on exactly the p variables of X_p:

$$\Pr_X(d_1 \wedge \cdots \wedge d_q) = \big\{x | x \in X \wedge \exists d_j (x^{\overline{\sigma_h}} \in d_j)\big\}; \quad (5.54)$$

$$H(X_p, q) = |\big\{d_1 \wedge \cdots \wedge d_q | \Pr_X(d_1 \wedge \cdots \wedge d_q) = X_p\big\}|. \quad (5.55)$$

Here, we use the informal notation $x^{\overline{\sigma}} \in d$ for being part of a disjunctive clause of a CNF $d_1 \wedge \cdots \wedge d_q$, as mentioned before. Since we consider a fixed $\widetilde{\sigma}$ and only literals of type $x^{\overline{\sigma}}$ constitute clauses in (5.54) and (5.55), we have

Lemma 5.5.

$$H(p,q) = H(X_p, q) = H(X'_p, q) \text{ for each pair } X_p \text{ and } X'_p.$$

Lemma 5.6. *For fixed $q \leq \binom{p}{k}$ and $k \leq p < k \cdot q$, $H(p,q)$ does not decrease for increasing p.*

Proof. We assume that for some p there is $H(p,q) > H(p{+}1,q)$. Given a CNF $F_p = d_1 \wedge \cdots \wedge d_q$ with $\Pr_X(d_1 \wedge \cdots \wedge d_q) = X_p$, and w.l.o.g. $X_p = \{x_1, ..., x_p\}$ and $F_p = d_1^{(p)} \wedge \cdots \wedge d_t^{(p)} \wedge d_{t+1} \wedge \cdots \wedge d_q$, where $d_j^{(p)}$ indicates that $x_p^{\overline{\sigma_p}}$ is part of the clause. We consider two cases:

1. $t \geq 2$: If $x_p^{\overline{\sigma_p}}$ is substituted by $x_{p+1}^{\overline{\sigma_{p+1}}}$ in each clause of subsets $T \subset \{d_1^{(p)}, ..., d_t^{(p)}\}$ of size $|T| = 1, ..., t-1$, then the corresponding CNFs F_{p+1} are pairwise different, depend on $p{+}1$ variables, and are therefore counted by $H(p{+}1,q)$. Thus, F_p induces by this procedure $2^t{-}2$ different F_{p+1}.
2. $t = 1$:
   (a) If $d_1^{(p)}$ and at least one clause out of $d_2, ..., d_q$ have at least one literal $x_h^{\overline{\sigma_h}}$ in common, then substituting $x_h^{\overline{\sigma_h}}$ in $d_1^{(p)}$ by $x_{p+1}^{\overline{\sigma_{p+1}}}$ creates a CNF F'_{p+1} that has not been generated in the first case, since the new clause $d_1^{(p+1)}$ depends on x_p as well as on x_{p+1}, and F'_{p+1} is counted by $H(p{+}1,q)$.
   (b) $d_1^{(p)}$ and $d_2, ..., d_q$ do not have literals in common. If $d_2, ..., d_q$ together depend on $k \cdot (q{-}1)$ variables, then F_p depends on $k \cdot (q{-}1){+}k = k \cdot q$ variables, which contradicts $p < k \cdot q$. Therefore, there exists x_h with the smallest index h among the variables with

the smallest number $u \geq 2$ of occurrences as literals in $d_2, ..., d_q$. Substituting $x_u^{\overline{\sigma_u}}$ by $x_p^{\overline{\sigma_p}}$ in one clause (w.l.o.g. in d_2) leads to $d_1^{(p)} \wedge d_2' \wedge d_3 \cdots \wedge d_q$ counted by $H(p,q)$, where

(i) $d_1^{(p)}$ and d_2' are indeed different, since the variables from $d_1^{(p)}$ do not occur in $d_2, ..., d_q$;

(ii) $d_1^{(p)} \wedge d_2' \wedge d_3 \cdots \wedge d_q$ has been considered in the first case and generated $2^2 - 2 = 2$ different F_{p+1}, namely $F_{p+1}^{(1)} = d_1^{(p+1)} \wedge d_2' \wedge d_3 \cdots \wedge d_q$ and $F_{p+1}^{(2)} = d_1^{(p)} \wedge d_2^{(p+1)} \wedge d_3 \cdots \wedge d_q$. On clauses with k literals one can impose an order, e.g. by an order of literals according to ascending indices, and then by a position representation to the basis $(p+2)$. Therefore, one can identify predecessor and successor with respect to $d_1^{(p+1)}$ and $d_2^{(p+1)}$ according to the imposed order, which can be extended to $F_{p+1}^{(1)}$ and $F_{p+1}^{(2)}$, since only $d_1^{(p+1)}$ and $d_2^{(p+1)}$ depend on x_{p+1}, and the remaining clauses a are the same. We then decide to count the CNF that appears first in the imposed order in Step 1, whereas the second CNF is counted in Step 2.

Thus, we obtain a contradiction to the assumption $H(p,q) > H(p+1,q)$. □

Since $H(p,q)$ does not decrease for fixed q and increasing p, we have $H(p,q) \leq H(k \cdot q, q)$. For $H(k \cdot q, q)$, the q k-clauses do not have literals in common, and therefore

$$H(p,q) \leq H(k \cdot q, q) = \frac{\binom{k \cdot q}{k} \cdot \binom{k \cdot q - k}{k} \cdots \binom{k \cdot q - (q-1) \cdot k}{k}}{q!} = \frac{(k \cdot q)!}{(k!)^q \cdot q!}. \tag{5.56}$$

There are $\binom{n}{p} \leq \binom{n}{k \cdot q}$ different p-selections of variables, and we note that $H(k \cdot q, q)$ from (5.56) monotonically increases for increasing q, i.e. increasing $k \cdot q$. Thus, we assume $p \leq k \cdot q = p_0 = n - 1$:

$$\sum_{p=k}^{k \cdot q} H(p,q) \quad \leq \quad k \cdot (q-1) \cdot \binom{n}{k \cdot q} \cdot \frac{(k \cdot q)!}{(k!)^q \cdot q!}. \tag{5.57}$$

For the crucial ratio from Step (i) on p. 21 we obtain

$$\begin{aligned} \frac{\binom{n}{k \cdot q}}{A(n,q)} \cdot \frac{(k \cdot q)!}{(k!)^q \cdot q!} &= \frac{\binom{n}{k \cdot q}}{\binom{\binom{n}{k}}{q}} \cdot \frac{(k \cdot q)!}{(k!)^q \cdot q!} \\ &= \frac{1}{(k!)^q} \cdot \frac{n! \cdot \left(\binom{n}{k} - q\right)!}{\binom{n}{k}! \cdot (n - k \cdot q)!} \\ &= \frac{1}{(k!)^q} \cdot \frac{n \cdot (n-1) \cdots (n - k \cdot q + 1)}{\binom{n}{k} \cdot \left(\binom{n}{k} - 1\right) \cdots \left(\binom{n}{k} - q + 1\right)} \\ &= \frac{1}{(k!)^q} \cdot \frac{n \cdot (n-1) \cdots (n - k \cdot q + 1)}{\left\{\binom{n}{k}\right\}^q \cdot \prod_{u=1}^{q-1} \left(1 - \frac{u}{\binom{n}{k}}\right)}. \end{aligned} \tag{5.58}$$

As in (5.22), we have from (5.21):

$$\begin{aligned}\frac{\binom{n}{k\cdot q}}{A(n,q)}\cdot\frac{(k\cdot q)!}{(k!)^q\cdot q!} &< e\cdot\frac{n\cdot(n-1)\cdots(n-k\cdot q+1)}{(k!)^q\cdot\left\{\binom{n}{k}\right\}^q}\\ &= e\cdot\prod_{t=1}^{q-1}\frac{(n-t\cdot k)\cdots(n-t\cdot k-k+1)}{n\cdot(n-1)\cdots(n-k+1)}\\ &= e\cdot\prod_{t=1}^{q-1}\prod_{u=0}^{k-1}\Big(1-\frac{t\cdot k}{n-u}\Big).\end{aligned}\tag{5.59}$$

In case that q is large enough, the product is split into two parts:

$$\frac{\binom{n}{k\cdot q}}{A(n,q)}\cdot\frac{(k\cdot q)!}{(k!)^q\cdot q!} < e\cdot\left(\frac{\chi(q)}{2^{\frac{n-k+1}{2\cdot k}}}+\prod_{t=1}^{q'-1}\prod_{u=0}^{k-1}\Big(1-\frac{t\cdot k}{n-u}\Big)\right),\tag{5.60}$$

where $q'=\min\{q,\frac{n}{2\cdot k}\}$ and $\chi(q)=1$ for $q'<q$, $\chi(q)=0$ otherwise. We then obtain

$$\begin{aligned}\frac{\binom{n}{k\cdot q}}{A(n,q)}\cdot\frac{(k\cdot q)!}{(k!)^q\cdot q!} &< e\cdot\left(\frac{\chi(q)}{2^{\frac{n-k+1}{2\cdot k}}}+\prod_{t=1}^{q'-1}e^{-t\cdot k\cdot\sum_{u=0}^{k-1}\frac{1}{n-u}}\right)\\ &< e\cdot\left(\frac{\chi(q)}{2^{\frac{n-k+1}{2\cdot k}}}+\prod_{t=1}^{q'-1}e^{-t\cdot k\cdot\ln\frac{n}{n-k+1}}\right)\\ &< e\cdot\left(\frac{\chi(q)}{2^{\frac{n-k+1}{2\cdot k}}}+e^{-(k-1)\cdot(q'-1)\cdot\frac{k\cdot q'}{2\cdot n}}\right).\end{aligned}$$

We now have that $n/2\le k\cdot q\le n-1$ implies

$$\frac{\sum_{p=k}^{k\cdot q}H(p,q)}{A(n,q)} < e\cdot k\cdot(q-1)\cdot 2^{-\frac{n-k+1}{2\cdot k}}.\tag{5.61}$$

In case of $k\cdot q<n/2$, we introduce the following lower bound

$$\sqrt{2\cdot n\cdot(\ln n+\varphi(n))}<k\cdot q \quad\Rightarrow\quad \frac{\sum_{p=k}^{k\cdot q}H(p,q)}{A(n,q)}<e^{-\varphi(n)}.\tag{5.62}$$

Eqn. (5.62) provides a lower bound on m: We recall that $r\ge k\cdot q$ and $m=q+r+s$, and therefore

$$m\ge\sqrt{2\cdot n\cdot(\ln n+\varphi(n))}\cdot\Big(1+\frac{1}{k}\Big),\tag{5.63}$$

where $\varphi(n)\to\infty$ with $n\to\infty$. We note that $k\cdot q\le n-1$, but above $O\big(\sqrt{n\cdot\varphi(n)}\big)$, implies a diminishing factor of $e^{-\varphi(n)}$.

The upper bound (5.62) implies that for $k\le p\le k\cdot q\le n-1$ and $\varphi(n)<\ln 2\cdot n/(2\cdot k)-O\big(\ln n\big)$ the upper bound in (5.52) can be substituted in the following way:

Lemma 5.7. *For fixed (q,r,s) and $m \geq O\big(\sqrt{n \cdot \varphi(n)}\big)$, the number of feasible pairs $[\mathsf{S}_q, \mathsf{S}_r]$ of sets of clauses, where the clauses in S_q depend on at most $p \leq p_0 = n-1$ variables, is upper bounded by $\Big(e^{-\varphi(n)} A(n,q)\Big) \cdot \Big(\widetilde{Q}(q,r) B(n,r)\Big)$, where $1 < \varphi(n) < n \cdot \ln 2/(2 \cdot k) - O\big(\ln n\big)$.*

We note that, based on (5.61), the analysis of (5.49) in accordance with Step (ii) from page 21 can be restricted to $k \cdot q \geq n$. Therefore, we re-define (5.51):

$$Q(q,r) \;=\; \begin{cases} 0, & \text{if } k \cdot q \leq n-1; \\ \widetilde{Q}(q,r) & \text{if } k \cdot q \geq n. \end{cases} \tag{5.64}$$

Therefore, $Q(q,r) = 0$ for $q < n/k$, which means that in (5.53) the summation over q may at least partially lead to summands equal to zero, if relatively small values of m imply values for q_1 and q_2, as defined in (5.34), below n/k.

Thus, from (5.53), (5.64), and Lemma 5.7 we obtain

Lemma 5.8. *If m satisfies (5.63) for $1 < \varphi(n) < n \cdot \ln 2/(2 \cdot k) - O\big(\ln n\big)$, then*

$$\begin{aligned} M_{\tilde{\sigma}} \;<\; & e \cdot \binom{2^k \cdot \binom{n}{k}}{m} \cdot \Bigg\{ \frac{21 \cdot 2^k}{k} \cdot \frac{1}{e^{\frac{1}{12} \cdot \frac{m}{2^k}}} + \frac{1}{2^{\varphi(n)}} + \\ & + \sum_{q=q_1}^{q_2} \sum_{r=r_1(q)}^{r_2(q)} A(n,q) \cdot Q(q,r) \cdot B(n,r) \cdot C(n, m-q-r) \Bigg\}. \end{aligned} \tag{5.65}$$

In accordance with (5.64) and (5.65), we have to find a tight upper bound for $Q(q,r)$, where r is within the range defined in (5.41) and q within the range defined in (5.34), with the additional condition $q \geq n/k$. Therefore, we finally return to (5.49), which means executing Step (ii) from page 21 for $\widetilde{p}$ with $p_0 = (n-1) < \widetilde{p} = n$.

The upper bound (5.49) is analysed by induction and independent of the particular values of p, i.e. we include value $p = 2$, which is below the lower bound $p \geq k \geq 3$, cf. (5.10). Since the case $p = 2$ is analysed in the same way as the general case, we immediately switch to the inductive step. Based on (5.46) and (5.49), our aim is to prove

$$S(a,p) \;=\; \sum_{g_1 + \cdots + g_v = a} \prod_{v=1}^{p} \frac{1}{(g_v + f_p)!} \;\leq\; \frac{p^{a+k \cdot q}}{(a+k \cdot q)!} \cdot \prod_{v=2}^{p} \rho_v, \tag{5.66}$$

where $\rho_v < 1$ and $f_p = \lceil \frac{k \cdot q}{p} \rceil \geq 1$, since (5.61) implies $p \leq k \cdot q$ and (5.64) together with Lemma 5.8 imply $k \cdot q \geq n$. By definition we have

$$S(a,p) \;=\; \sum_{t=0}^{a} \frac{P(a-t, p-1)}{(t+f_p)!}. \tag{5.67}$$

If we assume that (5.66) is true, we need to show

$$\sum_{t=0}^{a} \frac{1}{(t+f_p)!} \cdot \frac{(p-1)^{a-t+\frac{p-1}{p} \cdot k \cdot q}}{(a-t+\frac{p-1}{p} \cdot k \cdot q)!} \;\leq\; \rho_p \cdot \frac{p^{a+k \cdot q}}{(a+k \cdot q)!}. \tag{5.68}$$

We consider a single summand S_t when the LHS is divided by the RHS, except for ρ_v:

$$\begin{aligned} S_t &= \frac{(a+k\cdot q)!}{(a-t+\frac{p-1}{p}\cdot k\cdot q)!\cdot(t+f_p)!}\cdot\Big(\frac{p-1}{p}\Big)^{a-t+\frac{p-1}{p}\cdot k\cdot q}\cdot\Big(\frac{1}{p}\Big)^{t+f_p} \\ &= \binom{a+k\cdot q}{t+f_p}\cdot\Big(\frac{p-1}{p}\Big)^{a+k\cdot q-t-f_p}\cdot\Big(\frac{1}{p}\Big)^{t+f_p}. \end{aligned}$$

For the sum we need $\sum_{t=0}^{a} S_t \leq \rho_p$ and obtain:

$$\begin{aligned} \sum_{t=0}^{a} S_t &= \sum_{t=0}^{a}\binom{a+k\cdot q}{t+f_p}\cdot\Big(\frac{p-1}{p}\Big)^{a+k\cdot q-t-f_p}\cdot\Big(\frac{1}{p}\Big)^{t+f_p} \\ &= \sum_{h=0}^{r}\binom{r}{h}\cdot\Big(\frac{p-1}{p}\Big)^{r-h}\cdot\Big(\frac{1}{p}\Big)^{h} - \sum_{u=0}^{f_p-1}\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} - \\ &\quad - \sum_{v=r-\frac{p-1}{p}\cdot k\cdot q+1}^{r}\binom{r}{v}\cdot\Big(\frac{p-1}{p}\Big)^{r-v}\cdot\Big(\frac{1}{p}\Big)^{v} \\ &= \Big(\frac{p-1}{p}+\frac{1}{p}\Big)^{r} - \sum_{u=0}^{f_p-1}\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} - \\ &\quad - \sum_{v=r-\frac{p-1}{p}\cdot k\cdot q+1}^{r}\binom{r}{v}\cdot\Big(\frac{p-1}{p}\Big)^{r-v}\cdot\Big(\frac{1}{p}\Big)^{v} \\ &= 1 - \sum_{u=0}^{f_p-1}\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} - \\ &\quad - \sum_{v=r-\frac{p-1}{p}\cdot k\cdot q+1}^{r}\binom{r}{v}\cdot\Big(\frac{p-1}{p}\Big)^{r-v}\cdot\Big(\frac{1}{p}\Big)^{v}. \end{aligned} \tag{5.69}$$

For an upper bound of $\sum_{t=0}^{a} S_t$ we have to find lower bounds for the sums on the RHS, i.e. Chernoff bounds do not apply. We note that

$$\begin{aligned} \frac{r}{p} &> \frac{k\cdot q}{p} - 1 \\ r+p &> k\cdot q \\ a+k\cdot q+p &> k\cdot q \\ a+p &> 0. \end{aligned} \tag{5.70}$$

Therefore, we focus on the first sum

$$\sum_{t=0}^{a} S_t < 1 - \sum_{u=0}^{f_p-1}\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u}, \tag{5.71}$$

where we assume $f_p >> 1$. In order to simplify calculations, we consider $u = f_p$ instead of $f_p - 1$, and we analyse only a single summands. We note

that the values of the corresponding summands differ by the factor

$$\frac{(p-1)\cdot k\cdot q}{p\cdot(r+1)-k\cdot q}, \tag{5.72}$$

which is later taken into account. We utilise in the following the tight form of Stirling's formula

$$n! \;=\; \sqrt{2\cdot\pi}\cdot n^{n+\frac{1}{2}}\cdot e^{-n+r(n)}, \;\; \text{where } \frac{1}{12\cdot n+1} < r(n) < \frac{1}{12\cdot n}, \tag{5.73}$$

as presented in [24]. For $u = f_p = k\cdot q/p > 1$ we now proceed with the lower bound

$$\begin{aligned}
\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} &> \frac{1}{e^{\varepsilon}}\cdot\sqrt{\frac{2\cdot\pi\cdot r}{2\cdot\pi\cdot u\cdot 2\cdot\pi\cdot(r-u)}}\cdot\frac{r^r}{u^u\cdot(r-u)^{r-u}}\times\\
&\quad\times\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u}\\
&= \frac{1}{e^{\varepsilon}}\cdot\sqrt{\frac{r}{2\cdot\pi\cdot u\cdot(r-u)}}\cdot\Big(\frac{r}{u}\Big)^{u}\cdot\Big(1+\frac{u}{r-u}\Big)^{r-u}\times\\
&\quad\times\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u}\\
&= \frac{1}{e^{\varepsilon}}\cdot\sqrt{\frac{r}{2\cdot\pi\cdot u\cdot(r-u)}}\cdot\Big(\frac{r}{p\cdot u}\Big)^{u}\cdot\Big(1-\frac{r-p\cdot u}{p\cdot(r-u)}\Big)^{r-u}\\
&> \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot\Big(\frac{r}{p\cdot u}\Big)^{u}\cdot\Big(1-\frac{r-p\cdot u}{p\cdot(r-u)}\Big)^{r-u}\\
&= \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot\Big(\frac{r}{p\cdot u}\Big)^{u}\cdot\Big(1+\frac{r-p\cdot u}{(p-1)\cdot r}\Big)^{-(r-u)}.
\end{aligned}$$

We have $r-p\cdot u < (p-1)\cdot r$, and therefore we continue with

$$\begin{aligned}
\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} &> \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot\Big(\frac{a+k\cdot q}{k\cdot q}\Big)^{u}\cdot e^{-\frac{a}{p-1}\cdot\left(1-\frac{u}{r}\right)}\\
&= \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot\Big(1+\frac{a}{k\cdot q}\Big)^{u}\cdot e^{-\frac{a}{p-1}\cdot\left(1-\frac{u}{r}\right)}.
\end{aligned}$$

We distinguish between $a < k\cdot q$ and $a \geq k\cdot q$. For $a < k\cdot q$ we obtain

$$\begin{aligned}
\binom{r}{u}\cdot\Big(\frac{p-1}{p}\Big)^{r-u}\cdot\Big(\frac{1}{p}\Big)^{u} &> \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot e^{\frac{a}{k\cdot q+a}\cdot u}\cdot e^{-\frac{a}{p-1}\cdot\left(1-\frac{u}{r}\right)}\\
&= \frac{1}{\sqrt{2\cdot\pi\cdot u}}\cdot e^{a\cdot\frac{u}{r}}\cdot e^{-\frac{a}{p-1}\cdot\left(1-\frac{u}{r}\right)}\\
&= \sqrt{\frac{p}{2\cdot\pi\cdot k\cdot q}}\cdot e^{-\frac{a}{p-1}\cdot\left(1-p\cdot\frac{u}{r}\right)}\\
&= \sqrt{\frac{p}{2\cdot\pi\cdot k\cdot q}}\cdot e^{-\frac{a}{p-1}\cdot\frac{a}{r}}\\
&> \sqrt{\frac{p}{2\cdot\pi\cdot k\cdot q}}\cdot e^{-\frac{a}{p-1}},
\end{aligned} \tag{5.74}$$

which, apart from the the $\sqrt{\cdots}$-factor, comes close to an upper bound derived by the Chernoff-method. For $a \geq k \cdot q$ we have

$$\begin{aligned}\binom{r}{u} \cdot \left(\frac{p-1}{p}\right)^{r-u} \cdot \left(\frac{1}{p}\right)^{u} &> \frac{1}{\sqrt{2 \cdot \pi \cdot u}} \cdot \left(1 + \frac{a}{k \cdot q}\right)^{u} \cdot e^{-\frac{a}{p-1} \cdot \left(1 - \frac{u}{r}\right)} \\ &> \frac{1}{\sqrt{2 \cdot \pi \cdot u}} \cdot e^{u \cdot \ln 2} \cdot e^{-\frac{a}{p-1} \cdot \left(1 - \frac{u}{r}\right)} \\ &= \frac{1}{\sqrt{2 \cdot \pi \cdot u}} \cdot e^{-\frac{a}{p-1} \cdot \left(1 - \frac{(p-1) \cdot \ln 2}{a} - \frac{u}{r}\right)}.\end{aligned}$$

Since $\ln 2 < 1$ and $a \geq k \cdot q > p-1 \geq 1$ in the present case, we obtain together with $e^3 > 2^4$

$$\begin{aligned}\binom{r}{u} \cdot \left(\frac{p-1}{p}\right)^{r-u} \cdot \left(\frac{1}{p}\right)^{u} &> \frac{1}{\sqrt{2 \cdot \pi \cdot u}} \cdot e^{-\frac{a}{p-1} \cdot \left(1 - \ln 2 - \frac{u}{r}\right)} \\ &> \frac{1}{\sqrt{2 \cdot \pi \cdot u}} \cdot e^{-\frac{a}{p-1} \cdot \left(1 - \ln 2 - \frac{1}{4}\right)} \\ &= \sqrt{\frac{p}{2 \cdot \pi \cdot k \cdot q}} \cdot e^{-\frac{a}{p-1} \cdot \left(\frac{3}{4} - \ln 2\right)}, \qquad (5.75)\end{aligned}$$

which is larger than the lower bound from (5.74). Taking into account (5.72), we employ

$$\frac{(p-1) \cdot k \cdot q}{p \cdot (r+1) - k \cdot q} \cdot \sqrt{\frac{p}{2 \cdot \pi \cdot k \cdot q}} > \frac{\sqrt{p \cdot k \cdot q}}{2 \cdot \pi \cdot r}. \qquad (5.76)$$

We note that we did not use $k \cdot q \geq n$ so far, and for (5.71) not to degenerate, we need $k \cdot q \geq p+1$. For ρ_p from (5.66) and (5.68) we now set in accordance with (5.71), (5.74), and (5.76):

$$\rho_p = \begin{cases} 1 - \frac{\sqrt{p \cdot k \cdot q}}{2 \cdot \pi \cdot r} \cdot e^{-\frac{a}{p-1}}, & \text{if } 2 \leq p \leq n-1; \\ 1, & \text{if } p = n. \end{cases} \qquad (5.77)$$

For the product of ρ_p we obtain

$$\prod_{p=2}^{n-1} \rho_p = \prod_{p=2}^{n-1} \left(1 - \alpha_p \cdot e^{-\frac{a}{p-1}}\right) < e^{-\sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}}}. \qquad (5.78)$$

We proceed with

$$\begin{aligned}\sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}} &= \frac{\alpha_{n-1}}{e^{\frac{a}{n-2}}} \cdot \left(1 + \sum_{u=1}^{n-3} \frac{\alpha_{n-1-u}}{\alpha_{n-1}} \cdot e^{-\frac{u \cdot a}{(n-2) \cdot (n-2-u)}}\right) \\ &> \frac{\alpha_{n-1}}{e^{\frac{a}{n-2}}} \cdot \left(1 + \sum_{u=1}^{n-3} \sqrt{\frac{n-1-u}{n-1}} \cdot e^{-\frac{u \cdot a}{(n-2) \cdot (n-2-u)}}\right).\end{aligned}$$

For $u \leq (3/4) \cdot (n-1)$ and $n \geq 24$ we have $1/e^{(u \cdot a)/((n-2) \cdot (n-2-u))} > 1/e^{(4 \cdot a)/n}$ and $\sqrt{\frac{n-1-u}{n-1}} \geq 1/2$. Therefore, if $n \geq 24$, then

$$\sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}} > \frac{\alpha_{n-1}}{e^{\frac{a}{n-2}}} \cdot \left(1 + \frac{3}{4} \cdot (n-1) \cdot \frac{1}{2} \cdot e^{-\frac{4 \cdot a}{n}}\right). \qquad (5.79)$$

We now incorporate (5.34) and (5.41). Since $r = a + k \cdot q$, we have from $r_1(q) \leq r \leq r_2(q)$ and $q_1 \leq q \leq q_2$:

$$a \;\leq\; \frac{2^{k+2}-3}{2^{k+1}} \cdot \frac{k}{2^{k+1}} \cdot m \;<\; \frac{k}{2^k} \cdot m; \tag{5.80}$$

$$a + k \cdot q \;\leq\; 3 \cdot \frac{k}{2^{k+1}} \cdot m. \tag{5.81}$$

Therefore, (5.79) turns to

$$\sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}} \;>\; \frac{2^{\frac{k+1}{2}}}{6 \cdot \pi \cdot k} \cdot \frac{1}{\sqrt{m}} \cdot \frac{1}{e^{\frac{k}{2^k} \cdot \frac{m}{n}}} \cdot \Big(1 + \frac{3}{8} \cdot (n-1) \cdot e^{-\frac{4 \cdot k}{2^k} \cdot \frac{m}{n}}\Big). \tag{5.82}$$

We now consider two cases:

1. $m \leq \frac{2^k}{k} \cdot n \cdot \alpha$ for $0 < \alpha = \text{const}$: from (5.82) we obtain

$$\begin{aligned} \sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}} \;&>\; \frac{\sqrt{2}}{6 \cdot \pi \cdot \sqrt{\alpha \cdot k \cdot n}} \cdot \frac{1}{e^{\alpha}} \cdot \Big(1 + \frac{3}{8} \cdot (n-1) \cdot e^{-4 \cdot \alpha}\Big) \\ &=\; O\Big(\sqrt{\frac{n}{k}}\Big). \end{aligned} \tag{5.83}$$

2. $m = \frac{2^k}{k} \cdot n \cdot \beta(n)$ for $\beta(n) \to \infty$: we then have

$$\sum_{p=2}^{n-1} \frac{\alpha_p}{e^{\frac{a}{p-1}}} \;>\; O\Big(\sqrt{\frac{n}{k \cdot \beta(n)}} \cdot \frac{1}{e^{5 \cdot \beta(n)}}\Big). \tag{5.84}$$

From (5.21), (5.49), (5.51), (5.64), (5.66), (5.78), (5.83), and (5.84) we finally obtain

Lemma 5.9. *If $n \geq 24$, $k \cdot q \geq n$, $q_1 \leq q \leq q_1$ and $r_1(q) \leq r \leq r_2(q)$ for the values defined in (5.34) and (5.41), then*

$$Q(q,r) \;<\; \begin{cases} e^{-O\left(\sqrt{\frac{n}{k}}\right)}, & \text{if } m \leq \frac{2^k}{k} \cdot n \cdot \text{const}; \\ e^{-O\left(\sqrt{\frac{n}{k \cdot \beta(n)}} \cdot \frac{1}{e^{5 \cdot \beta(n)}}\right)}, & \text{if } m = \frac{2^k}{k} \cdot n \cdot \beta(n) \text{ for} \\ & \beta(n) \leq O\big(2^{-\frac{k}{2}} \cdot \big(\frac{n \cdot e}{k}\big)^{\frac{k-2}{2}}\big). \end{cases}$$

where $Q(q,r)$ is defined in (5.64).

We distinguish between the two cases in order to emphasise the behaviour of the upper bound for $m \leq O\Big(\frac{2^k}{k} \cdot n\Big)$. Furthermore, for slowly increasing functions $\beta(n)$, e.g. $\beta(n) = \ln \ln n$, the upper bound still decreases.

Taking into account the Vandermonde identity, Lemma 5.8, and Lemma 5.9 together with the lower bound (5.63), we finally arrive at

1. If $O\big(\sqrt{n \cdot \varphi(n)}\big) \leq m$ and $m/n \to 0$, i.e. $O(\ln n) \leq \varphi \leq O(\sqrt{n})$, then

$$M_{\tilde{\sigma}} \;<\; \binom{2^k \cdot \binom{n}{k}}{m} \cdot O\Big(\frac{1}{2^{\varphi(n)}}\Big). \tag{5.85}$$

Here, we employ that the upper bound on m implies $k \cdot q < n$.

2. If $m = \frac{2^k}{k} \cdot n \cdot \psi(n)$, where $\psi(n) \geq const > 0$, then

$$M_{\tilde{\sigma}} \quad < \quad \binom{2^k \cdot \binom{n}{k}}{m} \cdot \frac{1}{e^{O\left(\sqrt{\frac{n}{k \cdot \psi(n)}} \cdot \frac{1}{e^{5 \cdot \psi(n)}}\right)}}. \tag{5.86}$$

In Lemma 5.3 and (5.11) until (5.86) we exploit only information about $x_i^{\overline{\sigma_i}}$ vs. $x_i^{\sigma_i}$, i.e. information about the actual values of σ_i has no impact on $M_{\tilde{\sigma}}$ at all. Thus, $M_{\tilde{\sigma}}$ depends only on structural parameters (n, k, m):

Lemma 5.10. *If $\tilde{\sigma}, \tilde{\eta} \in \{0,1\}^n$, then $M_{\tilde{\sigma}} = M_{\tilde{\eta}}$ for a given class $\mathsf{F}_k(n,m)$.*

In accordance with Definition 2, we finally obtain

Theorem *The average number $\widehat{\mathsf{N}^1_{\mathrm{lm}}}$ of one-step local maxima of $\mathsf{F}_k(n,m)$ is upper bounded by*

$$\widehat{\mathsf{N}^1_{\mathrm{lm}}} \quad < \quad \begin{cases} O\left(\frac{2^n}{2^{\varphi(n)}}\right), & \text{if } O\left(\sqrt{n \cdot \varphi(n)}\right) \leq m << n \text{ and} \\ & O(\ln n) \leq \varphi \leq O(\sqrt{n}); \\ \frac{2^n}{e^{O\left(\sqrt{\frac{n}{k \cdot \psi(n)}} \cdot \frac{1}{e^{5 \cdot \psi(n)}}\right)}}, & \text{if } m = \frac{2^k}{k} \cdot n \cdot \psi(n) \text{ and} \\ & 0 < \text{const} \leq \psi(n) \leq O\left(2^{-\frac{k}{2}} \cdot \left(\frac{ne}{k}\right)^{\frac{k-2}{2}}\right). \end{cases}$$

The upper bounds imply that for m is in the region of $2^k \cdot n/k$ the average number of local maxima is bounded by $2^{n-O(\sqrt{n/k})}$.

6. Concluding Remarks

The Garnier/Kallel-approach requires a partition of the search space into attraction basins, i.e. within each neighbourhood a single element with the maximum value of the objective function is assumed. This assumption does not apply to the neighbourhood in our study. Nevertheless, our computational experiments provide evidence that the sampling-based method for the approximation of the number of local maxima seems to work in the context of k-SAT instances. The quality of approximations is steady for an increasing size of sampling information and the maximum deviation from the true values is below 15%, with a typical value in the region of 10%. The theoretical analysis confirms a decreasing number of local maxima in the region of the phase transion for increasing values m of the number of clauses. We intend to analyse a variety of neighbourhood relations proposed in the literature [14, 28, 29], where it would be interesting to find out if the average number of local maxima can be related to the quality of the associated local search procedures. We intend to apply the Garnier/Kallel-method in a completely different context, namely to structure prediction problems in Computational Biology, such as RNA secondary structure prediction and protein folding simulation in various lattice models and for different types of the objective function. RNA secondary structure prediction with pseudo-knots as well as protein folding simulation in various lattice models are known to be

NP-complete and population-based heuristics are an obvious choice to tackle these problems [12]. The standard method for identifying local minima in folding landscapes are barrier trees [33]. As pointed out in [10], "... from a practical point of view, the tree describing the repartition of local optima is unknown and too expensive in terms of computational cost to determine for a given landscape." Thus, approximations as described in the present paper might be helpful for the analysis of energy landscapes induced by structure prediction problems.

References

[1] M. Alava, J. Ardelius, E. Aurell, P. Kaski, S. Krishnamurthy, P. Orponen, S. Seitz, *Circumspect descent prevails in solving random constraint satisfaction problems.* PNAS USA **105** (2008), 15253–15257.

[2] D. Achlioptas, Y. Peres, *The threshold for random k-SAT is* $2^k \cdot \log 2 - O(k)$. J. American Mathematical Society **17** (2004), 947–973.

[3] A.A. Albrecht, *A stopping criterion for logarithmic simulated annealing.* Computing **78** (2006), 55–79.

[4] A.A. Albrecht, P.C.R. Lane, K. Steinhöfel, *Combinatorial landscape analysis for k-SAT instances.* Proc. IEEE Congress on Evolutionary Computation (2008), 2498–2504.

[5] A.A. Albrecht, A. Skaliotis, K. Steinhöfel, *Stochastic protein folding simulation in the d-dimensional HP-model.* Computational Biology and Chemistry **32** (2008) 248–255.

[6] A. Barvinok, *On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries.* (2008), arXiv:0806.1480v2.

[7] T. Brueggemann, W. Kern, *An improved local search algorithm for 3-SAT.* Theoretical Computer Science **329** (2004), 303–313.

[8] E.R. Canfield, C. Greenhill, B.D. McKay, *Asymptotic enumeration of dense 0–1 matrices with specified line sums.* J. Combinatorial Theory (Series A) **115** (2008), 32–66.

[9] E. Dantsin, A. Wolpert, *An improved upper bound for SAT.* Proc. SAT 2005, LNCS 3569 (2005), 400–407.

[10] J. Garnier, L. Kallel, *Efficiency of local search with multiple local optima.* SIAM J. Discrete Mathematics **15** (2002), 122–141.

[11] A. Gerevini, I. Serina, *Planning as propositional CSP: From WalkSAT to local search techniques for action graphs.* Constraints **8** (2003), 389–413.

[12] H.J. Greenberg, W.E. Hart, G. Lancia, *Opportunities for combinatorial optimization in computational biology.* INFORMS J. Computing **16** (2004), 211–231.

[13] T. Hagerup, C. Rüb, *A guided tour of Chernoff bounds.* Information Processing Letter **33** (1990), 305–308.

[14] H.H. Hoos, T. Stützle, *Towards a characterisation of the behaviour of stochastic local search algorithms for SAT.* Artificial Intelligence **112** (1999), 213–232.

[15] P. Kaski, *Barriers and local minima in energy landscapes of stochastic local search.* (2006) arXiv:cs/0611103v1.

[16] O.C. Martin, R. Monasson, R. Zecchina, *Statistical mechanics methods and phase transitions in optimization problems.* Theoretical Computer Science **265** (2001), 3–67.

[17] M. Mézard, R. Zecchina, *Random K-satisfiability problem: from an analytic solution to an efficient algorithm.* Physical Review E **66** (2002), 056126-1–27.

[18] D. Mitchell, B. Selman, H. Levesque, *Hard and easy distributions of SAT problems.* Proc. 10^{th} National Conference on Artificial Intelligence (1992), 459–465.

[19] R. Monasson, R. Zecchina, *Statistical mechanics of the random K-SAT problem.* Physical Review E **56** (1997), 1357–1361.

[20] A. Montanari, G. Parisi, F. Ricci-Tersenghi, *Instability of one-step replica-symmetry-broken phase in satisfiability problems.* J. Physics A: Mathematics and General **37** (2004), 2073–2091.

[21] R. Paturi, P. Pudlák, M.E. Saks, F. Zane, *An improved exponential-time algorithm for k-SAT.* J. ACM **52** (2005), 337–364.

[22] C.R. Reeves, A.V. Eremeev, *Statistical analysis of local search landscapes.* J. Operational Research Society **55** (2004), 687–693.

[23] C.M. Reidys, P.F. Stadler, *Combinatorial landscapes.* SIAM Review **44** (2002), 3–54.

[24] H. Robbins, *A remark on Stirling's formula.* American Mathematical Monthly **62** (1955), 26–29.

[25] U. Schöning, *A probabilistic algorithm for k-SAT based on limited local search and restart.* Algorithmica **32** (2002), 615–623.

[26] P. Schuler, *An algorithm for the satisfiability problem of formulas in conjunctive normal form.* J. Algorithms **54** (2005), 40–44.

[27] D. Schuurmans, F. Southey, *Local search characteristics of incomplete SAT procedures.* Artificial Intelligence **132** (2001), 121–150.

[28] S. Seitz, M. Alava, P. Orponen, *Threshold behaviour of WalkSAT and focused Metropolis search on random 3-satisfiability.* Proc. SAT 2005, LNCS 3569 (2005), 475–481.

[29] S. Seitz, P. Orponen, *An efficient local search method for random 3-satisfiability.* Electronic Notes in Discrete Mathematics **16** (2003), 71–79.

[30] B. Selman, H.A. Kautz, B. Cohen, *Noise strategies for improving local search.* Proceedings AAAI (1994) 337–343, MIT Press, Cambridge, MA.

[31] B. Selman, H. Levesque, D. Mitchell, A new method for solving hard satisfiability problems, Proc. AAAI (1992), 440–446.

[32] K. Steinhöfel, A. Skaliotis, A.A. Albrecht, *Relating time complexity of protein folding simulation to approximations of folding time.* Computer Physics Communications **176** (2007), 165–170.

[33] D. Wales, *Energy landscapes.* Cambridge University Press, (2003).

[34] W. Zhang, *Configuration landscape analysis and backbone guided local search. Part I: Satisfiability and maximum satisfiability.* Artificial Intelligence **158** (2004), 1–26.

A.A. Albrecht
CCRCB
Queen's University Belfast
97 Lisburn Road, Belfast BT9 7BL, UK
e-mail: A.Albrecht@qub.ac.uk

P.C.R. Lane
School of Computer Science
University of Hertfordshire
College Lane, Hatfield AL10 9AB, UK
e-mail: P.C.Lane@herts.ac.uk

K. Steinhöfel
Department of Computer Science
King's College London
Strand, London WC2R 2LS, UK
e-mail: Kathleen.Steinhofel@kcl.ac.uk