On the Complexity to Approach Optimum Solutions by Inhomogeneous Markov Chains

Andreas A. Albrecht

University of Hertfordshire
Dept. of Computer Science
Hatfield, Herts AL10 9AB, UK

Abstract. We analyse the probability $1-\delta$ to be in an optimum solution after k steps of an inhomogeneous Markov chain which is specified by a logarithmic cooling schedule $c(k) = \Gamma/\ln(k+2)$. We prove that after $k > (n/\delta)^{O(\Gamma)}$ steps the probability to be in an optimum solution is larger than $1-\delta$, where n is an upper bound for the size of local neighbourhoods and Γ is a parameter of the entire configuration space. By counting the occurrences of configurations, we demonstrate for an application with known optimum solutions that the lower bound indeed ensures the stated probability for a relatively small constant in $O(\Gamma)$.

Keywords: Markov Chains, Simulated Annealing, Cooling Schedules, Local Search, Convergence Analysis.

1 Introduction

Simulated annealing-based algorithms play an important role in the context of evolutionary algorithms, cf. [7]. Simulated annealing (cf. [9,5]) can be classified by the underlying cooling schedule, i.e., by the method that determines how the temperature is lowered at progressing steps of the computation. If the temperature is kept constant for a (large) number of steps, a homogeneous Markov chain can be associated with this type of computation. Under some natural assumptions, the probabilities of configurations tend to the Boltzmann distribution for homogeneous Markov chains (cf. [11,12,13]). This type of simulated annealing algorithms has been studied intensely and numerous heuristics have been devised for a wide variety of combinatorial optimisation problems [1,4,10].

If the temperature is lowered at any step of the computation, the transition probabilities represent an inhomogeneous Markov chain. The special case of logarithmic cooling schedules has been investigated in [2,3,4,6]. B. Hajek [6] proved that logarithmic simulated annealing tends to an optimum solution if and only if the cooling schedule is lower bounded by $\Gamma/\ln(k+2)$, where Γ is the maximum value of the escape height from local minima of the underlying energy landscape.

Given the configuration space $\mathcal{F}$, let $\mathbf{a}_f(k)$ denote the probability to be in configuration f after k steps of an inhomogeneous Markov chain. The problem is

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 642–653, 2004.

to find a lower bound for k such that $\sum_{f\in\mathcal{F}_{\min}} \mathbf{a}_f(k) > 1-\delta$ for $f \in \mathcal{F}_{\min} \subseteq \mathcal{F}$ minimising the objective function. Let n denote a uniform upper bound for the number of neighbours of configurations $f \in \mathcal{F}$. We obtain a run-time of $k \geq (n/\delta)^{O(\Gamma)}$ to ensure that with probability $1-\delta$ the minimum value of the objective function has been approached.

We briefly describe the derivation of the lower bound and then illustrate the approach by an example from machine learning: From a single positive example and a number of negative examples a conjunction of shortest length representing the examples has to be calculated. Counting of configurations demonstrates that the lower bound for k indeed ensures the stated probability for a relatively small constant in $O(\Gamma)$.

2 Preliminaries

The configuration space is *finite* and denoted by $\mathcal{F}$. We assume an objective function $\mathcal{Z} : \mathcal{F} \longrightarrow \mathbb{N}$ that for simplicity takes its values from the set of integers. By $\mathcal{N}_f$ we denote the set of neighbours of f, including f itself. We assume that $\mathcal{F}$ is *reversible*: Any transition $f \to f'$, $f' \in \mathcal{N}_f$, can be performed in the reverse direction, i.e., $f \in \mathcal{N}_{f'}$, and we set

$$n := \max_{f\in\mathcal{F}} |\mathcal{N}_f| \,. \tag{1}$$

The set of minimal elements (optimum solutions) is defined by

$$\mathcal{F}_{\min} := \big\{\, f \,:\, \forall f'\big(f' \in \mathcal{F} \to\ \mathcal{Z}(f) \,\leq\, \mathcal{Z}(f')\big)\big\}.$$

Example. We consider Boolean conjunctive terms defined on n variables. The conjunctions are of length $\lceil \log n \rceil$ and have to be guessed from negative examples and one positive example $\tilde{\sigma}$. Hence, $\mathcal{F}$ consists of $f_{\tilde{\sigma}} = x^{\tilde{\sigma}}$ and all sub-terms f that can be obtained by deleting a literal from $x^{\tilde{\sigma}}$ in such a way that all negative examples are rejected. Each neighbourhood $\mathcal{N}_f$ contains at most $n' \leq n+1$ elements, since deleted literals can be included again. The configuration space is therefore reversible. The objective function is given by the length of terms f and in this case the optimum in known and equal to $\lceil \log n \rceil$.

In simulated annealing, the transitions between neighbouring elements are depending on the objective function $\mathcal{Z}$. Given a pair of configurations $[f, f']$, $f' \in \mathcal{N}_f$, we denote by $G[f,\, f']$ the probability of generating f' from f and by $A[f, f']$ the probability of accepting f' once it has been generated from f. Since we consider a single step of transitions, the value of $G[f,\, f']$ depends on the set $\mathcal{N}_f$. As in most applications of simulated annealing, we take a uniform probability which is given by

$$G[f,\, f'] := \frac{1}{|\mathcal{N}_f|}. \tag{2}$$

The acceptance probabilities $A[f,\, f']$, $f' \in \subseteq \mathcal{F}$ are derived from the underlying analogy to thermodynamic systems [1]:

$$A[f, f'] := \begin{cases} 1, & \text{if } \mathcal{Z}(f') - \mathcal{Z}(f) \leq 0, \\ e^{-\frac{\mathcal{Z}(f') - \mathcal{Z}(f)}{c}}, & \text{otherwise,} \end{cases} \tag{3}$$

where c is a control parameter having the interpretation of a *temperature* in annealing procedures. Finally, the probability of performing the transition between f and f' is defined by

$$\mathbf{Pr}\{f \to f'\} = \begin{cases} G[f,\, f'] \cdot A[f,\, f'], & \text{if } f' \neq f, \\ 1 - \sum\limits_{f' \neq f} G[f,\, f'] \cdot A[f,\, f'], & \text{otherwise.} \end{cases} \tag{4}$$

By definition, the probability $\mathbf{Pr}\{f \to f'\}$ depends on the control parameter c. Let $\mathbf{a}_f(k)$ denote the probability of being in the configuration f after k steps performed for the same value of c. The probability $\mathbf{a}_f(k)$ can be calculated in accordance with

$$\mathbf{a}_f(k) := \sum_h \mathbf{a}_h(k-1) \cdot \mathbf{Pr}\{h \to f\}. \tag{5}$$

The recursive application of (5) defines a Markov chain of probabilities $\mathbf{a}_f(k)$, where $f \in \mathcal{F}$ and $k = 1, 2, \ldots$. If the parameter $c = c(k)$ is a constant c, the chain is said to be a *homogeneous* Markov chain; otherwise, if $c(k)$ is lowered at any step, the sequence of probability vectors $\mathbf{a}(k)$ is an *inhomogeneous* Markov chain. In the present paper we focus on a special type of inhomogeneous Markov chains where the value $c(k)$ changes in accordance with

$$c(k) = \frac{\Gamma}{\ln(k+2)},\ k = 0, 1, \ldots. \tag{6}$$

The choice of $c(k)$ is motivated by Hajek's Theorem [6] on logarithmic cooling schedules for inhomogeneous Markov chains. To explain Hajek's result, we first need to introduce some parameters characterising local minima of the objective function:

Definition 1 *A configuration $f' \in \mathcal{F}$ is said to be reachable at height h from $f \in \mathcal{F}$, if $\exists f_0, f_1, \ldots, f_r \in \mathcal{F}\big(f_0 = f \wedge f_r = f'\big)$ such that $G[f_u, f_{u+1}] > 0$, $u = 0, 1, \ldots, (r-1)$ and $\mathcal{Z}(f_u) \leq h$, for all $u = 0, 1, \ldots, r$.*

We use the notation *height*$(f \Rightarrow f') \leq h$ for this property. The function f is a *local minimum*, if $f \in \mathcal{F} \setminus \mathcal{F}_{\min}$ and $\mathcal{Z}(f') > \mathcal{Z}(f)$ for all $f' \in \mathcal{N}_f \setminus f$.

Definition 2 *Let $g_{\min}$ denote a local minimum, then depth$(g_{\min})$ denotes the smallest h such that there exists a $g' \in \mathcal{F}$, where $\mathcal{Z}(g') < \mathcal{Z}(g_{\min})$, that is reachable at height $\mathcal{Z}(g_{\min}) + h$.*

The following convergence property has been proved by B. Hajek:

Theorem 1 [6] *Given a cooling schedule defined by*

$$c(k) = \frac{\Gamma}{\ln(k+2)},\ k = 0, 1, \ldots,$$

the asymptotic convergence $\sum_{f \in \mathcal{F}_{\min}} \mathbf{a}_f(k) \underset{k\to\infty}{\to} 1$ *of the stochastic algorithm that is based on (3) and (4) is guaranteed if and only if*

(i) $\forall f, f' \in \mathcal{F}\, \exists f_0, f_1, \ldots, f_r \in \mathcal{F}\big(f_0 = f \wedge f_r = f'\big)$*:* $G[f_u, f_{u+1}] > 0$, $u = 0, 1, \ldots, (r-1)$*;*
(ii) $\forall h$: *height*$(f \Rightarrow f') \leq h \iff$ *height*$(f' \Rightarrow f) \leq h$*;*
(iii) $\Gamma \geq \max_{g_{\min}}$ *depth*$(g_{\min})$.

In the following, we assume that the conditions (i), ... , (iii) of Hajek's Theorem are satisfied for our configurations space $\mathcal{F}$. Furthermore, for simplicity of notations we make the following

Basic Assumptions

The difference of the objective function is the same between neighbouring elements, like in the case of our Example;

For all neighbours $f' \in \mathcal{N}_f$ of f, $f' \neq f$, the value of the objective function is different from $\mathcal{Z}(f)$ (as in the Example).

For two neighbouring elements, the one with the smaller value of the objective function has more neighbours with a higher value of the objective function and less neighbours with a smaller value of the objective function (as in the Example).

Let K_0 denote the maximum of the minimum number of transitions to reach an optimum solution starting from an arbitrary $f \in \mathcal{F}$. In Section 3, we will prove the following convergence result:

Theorem 2 *Given the configuration space* $\mathcal{F}$ *and* Γ *as defined in Theorem 1, then* $k \geq K_0$ *and* $k > (n/\delta)^{O(\Gamma)}$ *imply for arbitrary initial probability distributions* $\mathbf{a}(0)$ *the relation*

$$\sum_{\tilde{f} \notin \mathcal{F}_{\min}} \mathbf{a}_{\tilde{f}}(k) \leq \delta, \quad \textit{and therefore} \quad \sum_{f' \in \mathcal{F}_{\min}} \mathbf{a}_{f'}(k) \geq 1 - \delta.$$

3 Convergence Analysis

We introduce the following partition of the set of configurations with respect to the value of the objective function:

$L_0 := \mathcal{F}_{\min}$ and

$$L_{h+1} := \{f : f \in \mathcal{F} \wedge \forall f'\big(f' \in \mathcal{F} \setminus \bigcup_{i=0}^{h} L_i \rightarrow \mathcal{Z}(f') \geq \mathcal{Z}(f)\}.$$

For any particular element $f \in \mathcal{F}$, we introduce notations for the number of neighbours with a certain length. We recall that the definition of the neighbourhood relation implies that $\mathcal{N}_f$ contains only one element with $\mathcal{Z}(f)$ - the element f itself. We denote

$$s(f) := |\{\, f' \,:\, f' \in \mathcal{N}_f \wedge \mathcal{Z}(f') \,>\, \mathcal{Z}(f)\}|, \tag{7}$$

$$r(f) := |\{\, f' \,:\, f' \in \mathcal{N}_f \wedge \mathcal{Z}(f') \,<\, \mathcal{Z}(f)\}|\,. \tag{8}$$

Thus, from the definition of $\mathcal{N}_f$ we have

$$s(f) \,+\, r(f) = |\mathcal{N}_f| - 1. \tag{9}$$

We consider the probability $\mathbf{a}_f(k)$ to be in the configuration $f \in \mathcal{F}$ after k transitions of an *inhomogeneous* Markov chain that is defined in accordance with (6). We observe that for $\mathcal{Z}(f') > \mathcal{Z}(f)$ the acceptance probability (3) can be rewritten as

$$e^{-(\mathcal{Z}(f')-\mathcal{Z}(f))/c(k)} = \frac{1}{\big(k+2\big)^{(\mathcal{Z}(f')-\mathcal{Z}(f))/\Gamma}},\ k \,\geq\, 0. \tag{10}$$

To simplify notations, we define a new objective function where we maintain the same notation: $\mathcal{Z}(f) := \mathcal{Z}(f)/\Gamma$.

In (5), we separate the probabilities according to whether or not f' equals f, and we obtain:

Lemma 1 *The value of* $\mathbf{a}_f(k)$ *can be calculated from probabilities of the previous step by*

$$\mathbf{a}_f(k) = \left(\frac{s(f)+1}{|\mathcal{N}_f|} - \sum_{i=1}^{s(f)} \frac{\big(k+1\big)^{-\,(\mathcal{Z}(f_i)-\mathcal{Z}(f))}}{|\mathcal{N}_f|}\right) \cdot \mathbf{a}_f(k-1) + \sum_{i=1}^{s(f)} \frac{\mathbf{a}_{f_i}(k-1)}{|\mathcal{N}_{f_i}|} + \\ + \sum_{j=1}^{r(f)} \frac{\mathbf{a}_{f_j}(k-1)}{|\mathcal{N}_{f_j}|} \cdot \frac{1}{\big(k+1\big)^{\mathcal{Z}(f)-\mathcal{Z}(f_j)}}.$$

The representation (expansion) from Lemma 1 will be used in the following as the main relation reducing $\mathbf{a}_f(k)$ to probabilities from previous steps. Besides taking into account the value of the objective function in classes L_h, the elements of the configuration space are distinguished additionally by their minimum distance to $\mathcal{F}_{\min}$: Given $f \in \mathcal{F}$, we consider a shortest path of length $\mathrm{dist}(f)$ with respect to neighbourhood transitions from f to $\mathcal{F}_{\min}$. We introduce a partition of $\mathcal{F}$ in accordance with $\mathrm{dist}(f)$:

$$f \in M_i \Longleftrightarrow \mathrm{dist}(f) \,=\, i \,\geq\, 0, \quad \text{and} \quad \mathcal{M}_{\mathrm{d\_m}} \quad = \quad \bigcup_{i=0}^{\mathrm{d\_m}} M_i, \tag{11}$$

where $M_0 := L_0 = \mathcal{F}_{\min}$ and d_m is the maximum distance. Thus, we distinguish between distance levels M_i related to the minimum number of transitions

required to reach an element of $\mathcal{F}_{\min}$ and the levels L_h that are defined by the objective function.

Since we want to analyze the convergence to elements from $M_0 = L_0 = \mathcal{F}_{\min}$, we have to show that the value

$$\sum_{f \notin M_0} \mathbf{a}_f(k) \tag{12}$$

becomes small for large k. We suppose that $k \geq d\_m$, and we are going backwards from the k^{th} step. We consider the expansion of a particular probability $\mathbf{a}_f(k)$ as shown in Lemma 1. At the same step k, the neighbours of f are generating terms containing $\mathbf{a}_f(k-1)$ as a factor, in the same way, as $\mathbf{a}_f(k)$ generates terms with factors $\mathbf{a}_{f_i}(k-1)$ and $\mathbf{a}_{f_j}(k-1)$ in Lemma 1. If we consider the entire sum $\sum_{f \notin M_0} \mathbf{a}_f(k)$, the terms corresponding to a particular $\mathbf{a}_f(k-1)$ can be collected together to form a single term.

Firstly, we consider $f \in M_i$, $i \geq 2$. In this case, f does not have neighbours from M_0, i.e., the expansion from Lemma 1 appears for all neighbours of f in the reduction of $\sum_{f \notin M_0} \mathbf{a}_f(k)$ to step $(k-1)$. Therefore, taking all terms together that contain $\mathbf{a}_f(k-1)$, we obtain

$$\begin{aligned} &\mathbf{a}_f(k-1) \cdot \left\{ \left(\frac{|\mathcal{N}_f| - r(f)}{|\mathcal{N}_f|} - \sum_{i=1}^{s(f)} \frac{1}{|\mathcal{N}_f|} \cdot \frac{1}{\big(k+1\big)^{\mathcal{Z}(f_i)-\mathcal{Z}(f)}} \right) + \right. \\ &\left. + \sum_{i=1}^{s(f)} \frac{1}{|\mathcal{N}_f|} \cdot \frac{1}{\big(k+1\big)^{\mathcal{Z}(f_i)-\mathcal{Z}(f)}} + \sum_{j=1}^{r(f)} \frac{1}{|\mathcal{N}_f|} \right\} \\ &= \mathbf{a}_f(k-1). \end{aligned} \tag{13}$$

Secondly, if $f \in M_1$, the neighbours from M_0 are missing in $\sum_{f \notin M_0} \mathbf{a}_f(k)$ at the step to $(k-1)$, i.e., they do not generate terms containing probabilities from higher levels. For $f' \in M_0$, the expansion from Lemma 1 contains the terms $\mathbf{a}_{f_i}(k-1)/\,|\,\mathcal{N}_{f_i}\,|$ for $f_i \in M_1$ (and there are no terms for f_j with a smaller value of the objective function since $f' \in M_0$). Thus, the terms $\mathbf{a}_{f_i}(k-1)/\,|\,\mathcal{N}_{f_i}\,|$ are not available for $f = f_i \in M_1$ in the reduction of $\sum_{f \notin M_0} \mathbf{a}_f(k)$ to step $(k-1)$, when one tries to establish a relation like (13) for elements of M_1. For each $f \in M_1$, there are $r(f)$ such terms related to neighbours from M_0, see (8). Therefore, in the expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$, the following arithmetic term is generated when the particular f is from M_1:

$$\left(1 - \frac{r(f)}{|\,\mathcal{N}_f\,|}\right) \cdot \mathbf{a}_f(k-1). \tag{14}$$

We introduce the following abbreviations:

$$\varphi(f', f, v) := \frac{\big(k+2-v\big)^{-(\mathcal{Z}(f')-\mathcal{Z}(f))}}{|\mathcal{N}_f|}, \tag{15}$$

$$D_f(k-v) := \frac{s(f)+1}{|\mathcal{N}_f|} - \sum_{i=1}^{s(f)} \varphi(f', f, v). \tag{16}$$

Now, the relations expressed in (13) and (14) can be summarized to

Lemma 2 *A single step of the expansion of* $\sum_{f \notin M_0} \mathbf{a}_f(k)$ *results in*

$$\sum_{f \notin M_0} \mathbf{a}_f(k) = \sum_{f \notin M_0} \mathbf{a}_f(k-1) \; - \sum_{f \in M_1} \frac{r(f)}{|\mathcal{N}_f|} \cdot \mathbf{a}_f(k-1) \; + \\ + \sum_{f' \in M_1} \sum_{i=1}^{s(f')} \varphi(f_i, f', 1) \cdot \mathbf{a}_{f_j}(k-1).$$

The diminishing factor $\big(1 - r(f)/|\mathcal{N}_f|\big)$ appears by definition for all elements of M_1. At subsequent reduction steps, the factor is "transmitted" successively to all probabilities from higher distance levels M_i because any element of M_i has at least one neighbour from M_{i-1}. The main task is now to analyze how this diminishing factor changes when it is transmitted to higher distance levels. We denote

$$(17) \quad \sum_{f \notin M_0} \mathbf{a}_f(k) = \sum_{f \notin M_0} \mu(f, v) \cdot \mathbf{a}_f(k-v) \; + \sum_{f' \in M_0} \mu(f', v) \cdot \mathbf{a}_{f'}(k-v),$$

i.e., the coefficients $\mu(\tilde{f}, v)$ are the factors at probabilities after v steps of an expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$. Starting from step $(k-1)$, the probabilities $\mathbf{a}_{f'}(k-v)$, $f' \in M_0$, from (17) are expanded in the same way as the probabilities for all other $f \notin M_0$. We establish a recursive relation for the coefficients $\mu(\tilde{f}, v)$ defined in (17), where we apply the same expansion that resulted in equation (13) to the products $\mu(\tilde{f}, v) \cdot \mathbf{a}_{\tilde{f}}(k-v)$. For neighbouring elements we use $f' < \tilde{f}$ if $\mathcal{Z}(f') < \mathcal{Z}(\tilde{f})$, and $f' > \tilde{f}$ for the reverse relation of the objective function. Thus, taking into account (15) and (16), we obtain the following parameterized representation:

Lemma 3 *The following recurrent relation is valid for the coefficients* $\mu(\tilde{f}, v)$*:*

$$(18) \qquad \mu(\tilde{f}, v) = \mu(\tilde{f}, v-1) \cdot D_{\tilde{f}}(k-v) + \sum_{f' < \tilde{f}} \frac{\mu(f', v-1)}{|\mathcal{N}_{\tilde{f}}|} + \\ + \sum_{f'' > \tilde{f}} \mu(f'', v-1) \cdot \varphi(f'', \tilde{f}, v).$$

For $f \notin M_0$, we consider $\nu(f, v) = 1 - \mu(f, v)$ instead of $\mu(f, v)$ itself; for elements from M_0 we take the original value. When $\mu(f, v)$ is substituted in (18) by $1 - \nu(f, v)$, we obtain the same relation for $\nu(f, v)$ because the sum of transition probabilities equals 1 within the neighbourhood $\mathcal{N}_f$. We consider in more details the terms associated with elements of M_0 and M_1. We assume a representation $\mu(f', v-1) = \sum_{u'} T'_{u'}$ and $\nu(f, v-1) = \sum_u T_u$, where $T'_{u'}$ and T_u are arithmetic terms that have been generated at previous steps from the

elementary terms listed in Lemma 3 for $v = 1$. Since there are no $f'' < f'$ for $f' \in M_0$, we obtain:

$$\mu(f', v) = D_{f'}(k - v) \cdot \sum_{u'} T'_{u'}(f') + \sum_{f > f'} \Big(1 - \sum_{u} T_u(f)\Big) \cdot \varphi(f, f', v) \tag{19}$$

$$= \sum_{f > f'} \varphi(f, f', v) + D_{f'}(k - v) \cdot \sum_{u'} T'_{u'}(f') - \tag{20}$$

$$- \sum_{f > f'} \sum_{u} T_u(f) \cdot \varphi(f, f', v).$$

As can be seen, the term $\sum_{f > f'} \varphi(f, f', v)$ is generated at each time step v with the corresponding $\varphi(f, f', v)$. When (18) is written for $\nu(f, v)$, we obtain in the same way for elements of M_1:

$$\nu(f, v) = \frac{r(f)}{|\mathcal{N}_f|} + D_f(k - v) \cdot \nu(f, v - 1) + \tag{21}$$

$$+ \sum_{f'' > f} \nu(f'', v - 1) \cdot \varphi(f'', f, v) - \sum_{f' < f} \frac{\sum_{u'} T'_{u'}(f')}{|\mathcal{N}_f|},$$

where $r(f)/|\mathcal{N}_f|$ is from $\sum_{f' < \tilde{f}} 1/|\mathcal{N}_{\tilde{f}}|$. The term $r(f)/|\mathcal{N}_f|$ appears in all recursive equations of $\nu(f, v)$, $f \in M_1$ and $v \geq 1$, and the same is valid for the value $\sum_{f > f'} \varphi(f, f', v)$ in all $\mu(f', v)$, $f' \in M_0$. Therefore, all arithmetic terms T are derived from terms of the type $r(f)/|\mathcal{N}_f|$ and $\sum_{f > f'} \varphi(f, f', v)$. We try to keep track for each individual term that is generated by a recursive step as given in (18). For this purpose, the coefficients $\nu(f, v)$ are represented by a sum $\sum_i T_i$ of arithmetic terms (as in the derivation of (19), ..., (21)), and we are now going to define the terms T_i in more details by an inductive procedure.

Definition 3 *The terms $r(f)/|\mathcal{N}_f|$ (the first in (21)), $f \in M_1$, and $\sum_{f > f'} \varphi(f, f', v)$ (the first sum in (20)), $f' \in M_0$, are called* source terms *of $\nu(f, v)$ and $\mu(f', v)$, respectively, where $v \geq 1$.*

During an expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$ backwards according to (17), the source terms are distributed permanently to higher distance levels M_j as well as to elements from M_0. That means, in the same way as for M_1, the calculation of $\nu(f, v)$ ($\mu(f', v)$ for M_0) is repeated almost identically at any step, only the "history" of generations becomes longer. We introduce a counter $\mathbf{r}(f)$ to terms T that indicates the step at which the term has been generated from source terms. The value $\mathbf{r}(f)$ is called the rank of a term and we set $\mathbf{r}(f) = 1$ for source terms T from Definition 3. Basically, the rank $\mathbf{r}(f) \geq 1$ indicates the number of factors when T is represented by the subsequent multiplications according to the recurrent generation rules (20) and (21).

Let $\mathcal{T}_j(\tilde{f}, v)$ be the set of j^{th} rate arithmetic terms from $\nu(\tilde{f}, v)$ with the same rank $\mathbf{r}(f)$, where $\tilde{f} \in \mathcal{M}_{d\_m} \backslash M_0$. We set

$$\mathbf{S}_j(\tilde{f}, v) := \sum_{T \in \mathcal{T}_j(\tilde{f}, v)} T. \tag{22}$$

The same notation is used in case of $f' = \tilde{f} \in M_0$ with respect to $\mu(f', v)$. Now, the coefficients $\nu(\tilde{f}, v)$, $\mu(f', v)$, can be represented by

$$\nu(\tilde{f}, v) = \sum_{j=1}^{v} \mathbf{S}_j(\tilde{f}, v) \quad \text{and} \quad \mu(f', v) \;=\; \sum_{j=1}^{v} \mathbf{S}_j(f', v). \tag{23}$$

We compare the computation of $\nu(f, v)$ and $\mu(f', v)$ for two different values $v = k_1$ and $v = k_2$, i.e., $\nu(f, v)$ is calculated backwards from k_1 and k_2, respectively. Let $\mathbf{S}_j^1$ and $\mathbf{S}_j^2$ denote the corresponding sums of terms related to two different starting steps k_1 and k_2. From Definition 3 we see that the source term $r(f)/|\mathcal{N}_f|$ does not depend on k. For the second type of source terms, we employ the simple equation $k_2 - (k_2 - k_1 + v) = k_1 - v$, which leads to

Lemma 4 *Given $k_2 \geq k_1 \geq K_0$ and $1 \leq j \leq k_1$, then for each $f \in \mathcal{M}$:*

$$\mathbf{S}_j^1(f, v) = \mathbf{S}_j^2(f, k_2 - k_1 + v).$$

We use (17) and obtain:

$$\sum_{f \notin M_0} \mathbf{a}_f(k_1) = \sum_{f \notin M_0} \big(\mathbf{a}_f(k_1) - \mathbf{a}_f(k_2) \big) + \sum_{f \notin M_0} \mathbf{a}_f(k_2) \tag{24}$$

$$\begin{aligned} &= \sum_{f \notin M_0} \big(\nu(f, k_2 - k_1) - \nu(f, 0) \big) \cdot \mathbf{a}_f(k_1) + \\ &+ \sum_{f' \in M_0} \big(\mu(f', 0) - \mu(f', k_2 - k_1) \big) \cdot \mathbf{a}_{f'}(k_1) + \sum_{f \notin M_0} \mathbf{a}_f(k_2). \end{aligned} \tag{25}$$

For the first part of the sum we obtain:

$$\begin{aligned} &\sum_{f \notin M_0} \big(\nu(f, k_2 - 1) - \nu(f, 0) \big) \cdot \mathbf{a}_f(k_1) \\ &= \sum_{f \notin M_0} \Big(\sum_{j=1}^{k_2 - k_1} \mathbf{S}_j^2(f, k_2 - k_1) - \mathbf{S}_0^1(f, 0) \Big) \cdot \mathbf{a}_f(k_1), \end{aligned} \tag{26}$$

and Lemma 4 leads to:

$$\sum_{f \notin M_0} \big(\nu(f, k_2 - k_1) - \nu(f, 0) \big) \cdot \mathbf{a}_f(k_1) = \sum_{f \notin M_0} \sum_{j=1}^{k_2 - k_1} \mathbf{S}_j^2(f, k_2 - k_1) \cdot \mathbf{a}_f(k_1). \tag{27}$$

The same applies to configurations $f' \in M_0$.

To find upper bounds for (27), we estimate $\mathbf{a}_f(k_1)$ for configurations different from global and local minima, and the $\mathbf{S}_j^2(f, k_2 - k_1)$ are then estimated for global and local minima separately. To distinguish between the two cases is necessary since for small j and f different from global and local minima, the values $\mathbf{S}_j^2(f, k_2 - k_1)$ are relatively large (cf. Definition 3). We note that the recursive application of (18) generates negative summands in the representation

of values $\mathbf{S}_j(f,v)$, as can be seen from Definition 3 (cf. also (20) and (21)). We set $\mathbf{S}_j(f,v) = \mathbf{S}_j^+(f,v) - \mathbf{S}_j^-(f,v)$ and $\mathbf{S}_j(f',v) = \mathbf{S}_j^+(f',v) - \mathbf{S}_j^-(f',v)$ for $f \in M_1$ and $f' \in M_0$, where the partial sums consist of positive products only.

When $\mathbf{S}_j(f,v)$, $f \in M_1$, and $\mathbf{S}_j(f',v)$, $f' \in M_0$, are calculated, the negative products of $\mathbf{S}_{j-1}(f,v-1)$ become positive for $\mathbf{S}_j(f',v)$, and the negative products of $\mathbf{S}_{j-1}(f',v-1)$ become positive for $\mathbf{S}_j(f,v)$; see (20) and (21). The negative products of $\mathbf{S}_{j-1}(f,v-1)$ remain negative in the calculation of $\mathbf{S}_j(\tilde{f},v)$, $\tilde{f} \in M_2$, and the same applies to higher distance levels. Hence, the negative and positive products can be considered separately at all distance levels. Thus, we concentrate on upper bounds of $\mathbf{S}_j^+(f,v)$ only. To simplify notations, we use $\mathbf{S}_j(f,v)$ instead of $\mathbf{S}_j^+(f,v)$. Furthermore, we use instead of $n+1$ from $N_{\tilde{f}} \leq n+1$ (see (1)) the value $n' = n+1$, and for convenience n again for n'. We set $\widehat{\mathcal{M}} := \{f : r(f) \geq 1\}$, and for a constant $a > 0$ we can prove

$$\sum_{f \in L_h \cap \widehat{\mathcal{M}}} \mathbf{a}_f(k) < \frac{2 \cdot (n+1-h) \cdot n^a}{(k+2-n^a)^\gamma}. \tag{28}$$

Now, we estimate $\mathbf{S}_j(t,v)$ specifically for local and global minima. Here, we use the property that backwards expansions "entering" a local or global minimum are multiplied by $1/(k+2-v)^\gamma$, i.e., the upper bound is of the the type $\Pi/(k+2-v)^\gamma$, where Π represents the sum of products leading from M_1 (or M_0) to the local or global minimum. From Lemma 3 we conclude

$$\mathbf{S}_j(f,v) = \sum \Phi_1 \cdot \Phi_2 \cdots \Phi_j \;\leq \sum_{[d,g,h_1,h_2]} \sum_{\substack{\text{Possible Positions of} \\ D, G, H_1, H_2}} D^d \cdot G^g \cdot H_1^{h_1} \cdot H_2^{h_2}, \tag{29}$$

$d + g + h_1 + h_2 = j$, where D is the probability to stay in a local minimum, G corresponds to steps decreasing the objective function (we recall, that we are going backwards in the expansion of $\mathbf{a}_f(k-v)$), H_1 is associated with steps increasing the objective function, and H_2 is from the probability to stay in the same configuration which is not a local minimum.

For $f \in L_h$ we set $h(f) = h$ and we consider $f \in \mathcal{M} \backslash \widehat{\mathcal{M}}$. We set $k_1 := k + 2 - v + j$ and $k_2 := k + 2 - v$, and by induction on j we show

Lemma 5 *For $f \in \mathcal{M} \backslash \widehat{\mathcal{M}}$, $\Gamma > 3$, and $k \geq n^{2 \cdot \Gamma}$, the following inequality holds:*

$$\mathbf{S}_j(f,v) < e^{-\frac{j}{k_1^{3 \cdot \gamma}}} \cdot \left(1 + \frac{1}{k_2^\gamma}\right)^{h(f)}. \tag{30}$$

Based on (27) and Lemma 5, we derive an upper bound for $\sum_{f \in \mathcal{M} \backslash \widehat{\mathcal{M}}} \mathbf{a}_f(k)$, which leads to

$$\sum_{f \in \mathcal{M} \backslash \widehat{\mathcal{M}}} \mathbf{a}_f(k) < \frac{n^b}{(k+2-n^b)^\gamma},\ b = \text{const.} > 0. \tag{31}$$

Now, (28) and (31) are used to prove for $c \geq \max\{a+1, b\}$:

$$(32) \qquad \Big| \sum_{f \notin M_0} \big(\nu(f, k_2 - k_1) \; - \; \nu(f, 0)\big) \cdot \mathbf{a}_f(k_1) \Big| < O\Big(\frac{n^c}{(k + 2 - n^c)^\gamma}\Big).$$

Here, we consider $\mathbf{S}_j^+$, and (27) has been applied to these values only. But the same holds for $\mathbf{S}_j^-$, with even a smaller first factor of the expansion, see Lemma 3. Thus, in the same way we obtain the corresponding upper bound for $\big(\mu(f', 0) - \mu(f', k_2 - k_1)\big)$, and we finally complete the

Proof of Theorem 2: We utilise (24) until (26) and employ Theorem 1, i.e., if the constant Γ from (6) is sufficiently large, the inhomogeneous simulated annealing procedure defined by (2), (3), and (4) tends to the global minimum of $\mathcal{Z}$ on $\mathcal{F}$. The value k_2 from (32) is larger but independent of $k_1 = k$, i.e., we can take a $k_2 > k$ such that

$$\sum_{\tilde{f} \notin M_0} \mathbf{a}_{\tilde{f}}(k_2) < \frac{\delta}{3}.$$

Additionally, we require that both differences $\sum_{\tilde{f} \notin M_0} \big(\nu(\tilde{f}, k_2 - k) - \nu(\tilde{f}, 0)\big)$ and $\sum_{f' \in M_0} \big(\mu(f', 0) - \mu(f', k_2 - k)\big)$ are smaller than $\delta/3$. From (32) we obtain the condition

$$O\Big(\frac{n^c}{(k + 2 - n^c)^\gamma}\Big) < \frac{\delta}{3}.$$

We finally arrive at

$$k > \Big(\frac{n}{\delta}\Big)^{O(\Gamma)} \; \geq \; n^c - 2 + O\Big(\frac{3 \cdot n^c}{\delta}\Big)^\Gamma.$$

q.e.d.

4 Computational Experiments

We are given a set $\mathcal{S} \subseteq \{0,1\}^n$ of uniformly distributed binary n-tuples $\tilde{\eta} = \eta_1 \cdots \eta_n$ that represent negative examples for an unknown target conjunction $C_\ell = x_{i_1}^{\sigma_{i_1}} \& x_{i_2}^{\sigma_{i_2}} \& \cdots \& x_{i_\ell}^{\sigma_{i_\ell}}$ (here, we use $x^1 \equiv x$ and $x^0 \equiv \overline{x}$, i.e., $x^0 = 1$ for $x = 0$, and $x^0 = 0$ for $x = 1$), and a single positive example $\tilde{\sigma} = \sigma_1 \cdots \sigma_n$: $C_\ell(\tilde{\sigma}) = 1$ and $\forall \tilde{\eta} (\tilde{\eta} \in \mathcal{S} \rightarrow C_\ell(\tilde{\eta}) = 0)$. The task is to find a conjunction C_l of length $l \leq \ell$ that matches all of the samples, i.e., from C_ℓ generating the samples we do know only the length ℓ; cf. [8] and the Example in Section 2.

As explained in Section 2, we have $\Gamma \leq \lceil \log n \rceil$ for the problem to find a conjunction of length $\ell = \lceil \log n \rceil$. We implemented the search procedure for $m = 32$ negative examples, and for each element of $\mathcal{F}$ we counted the number of occurrences during the search procedure, in particular, for $\mathcal{F}_{\min}$. The calculations were repeated three times, and we present the average values (we obsereved only small deviations). The constant c in $O(\Gamma) = c \cdot \Gamma$ was set to $c = 1$.

		$n = 8$ and $\Gamma = 3$		$n = 16$ and $\Gamma = 4$	
δ	$1 - \delta$	k according to Theorem 2 ($c = 1$)	Frequency of $f \in \mathcal{F}_{\min}$	k according to Theorem 2 ($c = 1$)	Frequency of $f \in \mathcal{F}_{\min}$
0.50	0.50	4096	0.739	1048576	0.786
0.25	0.75	32768	0.812	16777216	0.895
0.10	0.90	512000	0.945	655360000	0.953
0.01	0.99	512000000	0.996	———	——

Frequencies of $f \in \mathcal{F}_{\min}$.

As we can see, the experimental results are in compliance with Theorem 2 for the small constant $c = 1$.

References

1. E.H.L. Aarts and J.H.M. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach*, Wiley & Sons, New York, 1989.
2. S. Azencott (editor). *Simulated Annealing: Parallelization Techniques.* Wiley & Sons, New York, 1992.
3. O. Catoni. Rough Large Deviation Estimates for Simulated Annealing: Applications to Exponential Schedules. *Annals of Probability*, 20(3):1109 – 1146, 1992.
4. O. Catoni. Metropolis, Simulated Annealing, and Iterated Energy Transformation Algorithms: Theory and Experiments. *J. of Complexity*, 12(4):595 – 623, 1996.
5. V. Černy. A Thermodynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm. Preprint, Inst. of Physics and Biophysics, Comenius Univ., Bratislava, 1982 (see also: *J. Optim. Theory Appl.*, 45:41 – 51, 1985).
6. B. Hajek. Cooling Schedules for Optimal Annealing. *Mathem. Oper. Res.*, 13:311 – 329, 1988.
7. W.E. Hart. A Theoretical Comparison of Evolutionary Algorithms and Simulated Annealing. In *Proc. of the 5^{th} Annual Conf. on Evolutionary Programming*, pp. 147-154, 1996.
8. M. Kearns, M. Li, L. Pitt, and L.G. Valiant. Recent Results on Boolean Concept Learning. In *Proc. 4^{th} Int. Workshop on Machine Learning*, pp. 337 – 352, 1987.
9. S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671 – 680, 1983.
10. F. Romeo and A. Sangiovanni-Vincentelli. A Theoretical Framework for Simulated Annealing. *Algorithmica*, vol. 6, no. 3, pp. 302 – 345, 1991.
11. E. Seneta. *Non-negative Matrices and Markov Chains.* Springer-Verlag, New York, 1981.
12. A. Sinclair and M. Jerrum. Approximate Counting, Uniform Generation, and Rapidly Mixing Markov Chains. *Information and Computation*, 82:93 – 133, 1989.
13. A. Sinclair and M. Jerrum. Polynomial-Time Approximation Algorithms for the Ising Model. *SIAM J. Comput.*, 22(5):1087 – 1116, 1993.