On Logarithmic Simulated Annealing^{*}

Andreas Albrecht^{1,2} and Chak-Kuen Wong^{2**}

¹ University of Hertfordshire

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

² Dept. of Computer Science and Engineering,

The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Abstract. We perform a convergence analysis of simulated annealing for the special case of logarithmic cooling schedules. For this class of simulated annealing algorithms, B. HAJEK [7] proved that the convergence to optimum solutions requires the lower bound $\Gamma/\ln(k+2)$ on the cooling schedule, where k is the number of transitions and Γ denotes the maximum value of the escape depth from local minima. Let n be a uniform upper bound for the number of neighbours in the underlying configuration space. Under some natural assumptions, we prove the following convergence rate: After $k \geq n^{O(\Gamma)} + \log^{O(1)}(1/\varepsilon)$ transitions the probability to be in an optimum solution is larger than $(1-\varepsilon)$. The result can be applied, for example, to the average case analysis of stochastic algorithms when estimations of the corresponding values Γ are known.

1 Introduction

Simulated annealing was introduced independently by KIRKPATRICK et al. [9] and V. ČERNÝ [6] as a new class of algorithms computing approximate solutions of combinatorial optimisation problems. The general approach itself is derived from METROPOLIS' method [10] to calculate equilibrium states for substances consisting of interacting molecules.

Simulated annealing algorithms can be distinguished by the method that determines how the temperature is lowered at different times of the computation process (the so-called cooling schedule). When the temperature is kept constant for a (large) number of steps, one can associate a homogeneous Markov chain with this type of computation. Under some natural assumptions, the probabilities of configurations tend to the Boltzmann distribution in the case of homogeneous Markov chains (see [13,14,15]). This class of simulated annealing

---

^{*} Research partially supported by the Strategic Research Programme at The Chinese University of Hong Kong under Grant No. SRP 9505 and by a Hong Kong Government RGC Earmarked Grant, Ref. No. CUHK 4010/98E.

^{**} On leave from IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, N.Y., U.S.A.

algorithms has been studied intensely and numerous heuristics have been derived for a wide variety of combinatorial optimisation problems (see [2,5,11,12,16]).

When the temperature is lowered at any step, the probabilities of configurations are computed from an inhomogeneous Markov chain. The special case of logarithmic cooling schedules has been investigated in [1,4,5,7]. B. HAJEK [7] 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 depth from local minima of the underlying energy landscape. O. CATONI [4,5] obtained a convergence rate of $M \cdot k^{-\alpha}$, where M depends on the number of configurations with a certain value of the objective function and α is related to the parameter Γ . Since structural parameters are treated as constants, the estimation does not provide a time bound that depends on Γ and the confidence parameter ε only.

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 to find a lower bound for k such that $\sum_{f \in \mathcal{F}_{\min}} \mathbf{a}_f(k) > 1 - \varepsilon$ for $f \in \mathcal{F}_{\min} \subseteq \mathcal{F}$ minimising the objective function. By n we denote a uniform upper bound for the number of neighbours of configurations $f \in \mathcal{F}$.

We obtain a run-time of $n^{O(\Gamma)} + \log^{O(1)} 1/\varepsilon$ that is sufficient to approach with probability $1 - \varepsilon$ the minimum value of the objective function. The present approach is based on a very detailed analysis of transition probabilities and has been developed in the context of equilibrium computations for specific physical systems [3].

2 Preliminaries

Simulated annealing algorithms are acting within a configuration space in accordance with a certain neighbourhood relation, where the particular transitions between adjacent elements of the configuration space are governed by an objective function. In the following section we introduce these notations for our problem setting.

The configuration space is *finite* and denoted by $\mathcal{F}$. We assume an objective function $\mathcal{Z} : \mathcal{F} \rightarrow \mathbb{N}$ 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 $f \rightarrow f'$, $f' \in \mathcal{N}_f$, can be performed in the reverse direction, i.e., $f \in \mathcal{N}_{f'}$. We set $n := \max_{f \in \mathcal{F}} |\mathcal{N}_f|$.

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

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

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 $t_{\tilde{\sigma}} = x^{\tilde{\sigma}}$ and all subterms t that can be obtained by deleting a literal from $x^{\tilde{\sigma}}$ and that reject all negative examples. Each neighbourhood $\mathcal{N}_t$ 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 t and in this case the optimum is 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

$$(1) \quad G[f, f'] := \frac{1}{|\mathcal{N}_f|}.$$

The acceptance probabilities $A[f, f']$, $f' \in \subseteq \mathcal{F}$ are derived from the underlying analogy to thermodynamic systems [2,10]:

$$(2) \quad A[f, f'] := \begin{cases} 1, & \text{if } \mathcal{Z}(f') - \mathcal{Z}(f) \leq 0, \\ e^{-\frac{\mathcal{Z}(f') - \mathcal{Z}(f)}{c}}, & \text{otherwise,} \end{cases}$$

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

$$(3) \quad \mathbf{Pr}\{f \rightarrow f'\} = \begin{cases} G[f, f'] \cdot A[f, f'], & \text{if } f' \neq f, \\ 1 - \sum_{f' \neq f} G[f, f'] \cdot A[f, g], & \text{otherwise.} \end{cases}$$

By definition, the probability $\mathbf{Pr}\{f \rightarrow f'\}$ 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

$$(4) \quad \mathbf{a}_f(k) := \sum_h \mathbf{a}_h(k-1) \cdot \mathbf{Pr}\{h \rightarrow f\}.$$

The recursive application of (4) defines a Markov chain of probabilities $\mathbf{a}_f(k)$, where $f \in \mathcal{F}$ and $k = 1, 2, \dots$. 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 are focusing on a special type of inhomogeneous Markov chains where the value $c(k)$ changes in accordance with

$$(5) \quad c(k) = \frac{\Gamma}{\ln(k+2)}, \quad k = 0, 1, \dots$$

The choice of $c(k)$ is motivated by HAJEK's Theorem [7] 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 $\exists f_0, f_1, \dots, f_r \in \mathcal{F}$ ($f_0 = f \wedge f_r = f'$) such that $G[f_u, f_{u+1}] > 0$, $u = 0, 1, \dots, (r-1)$ and $\mathcal{Z}(f_u) \leq h$, for all $u = 0, 1, \dots, r$.*

We use the notation $\text{height}(f \Rightarrow f') \leq h$ 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 $\text{depth}(g_{\min})$ denotes the smallest h such that there exists a $g' \in \mathcal{F}$, where $\mathcal{Z}(g') < \mathcal{Z}(g_{\min})$, which is reachable at height $\mathcal{Z}(g_{\min}) + h$.*

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

Theorem 1 [7] *Given a cooling schedule defined by*

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

the asymptotic convergence $\sum_{f \in \mathcal{F}} \mathbf{a}_f(k) \xrightarrow{k \rightarrow \infty} 1$ of the stochastic algorithm that is based on (2) and (3) is guaranteed if and only if

- (i) $\forall g, g' \in \mathcal{F} \exists g_0, g_1, \dots, g_r \in \mathcal{F}$ ($g_0 = g \wedge g_r = g'$):
 $G[g_u, g_{u+1}] > 0$, $u = 0, 1, \dots, (r-1)$;
- (ii) $\forall h : \text{height}(f \Rightarrow f') \leq h \iff \text{height}(f' \Rightarrow f) \leq h$;
- (iii) $\Gamma \geq \max_{g_{\min}} \text{depth}(g_{\min})$.

In the following, we assume that the conditions (i), ... , (iii) of HAJEK's Theorem are satisfied for our configurations space $\mathcal{F}$. Furthermore, we set the upper bound $|\mathcal{F}| \leq \exp(n^{O(1)})$ which is common in combinatorial optimisation. Additionally, we assume that the difference of the objective function is the same between neighbouring elements, like in the case of our Example.

3 Convergence Analysis

Our convergence result is derived from a careful analysis of the “exchange of probabilities” among configurations that belong to adjacent distance levels, where the distance is measured by the minimum number of steps to reach an element of $\mathcal{F}_{\min}$. Therefore, in addition to the value of the objective function, the elements of the configuration space are further distinguished by their distance to $\mathcal{F}_{\min}$.

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

$$L_0 := \mathcal{F}_{\min} \quad \text{and} \quad L_{h+1} := \{f : f \in \mathcal{F} \wedge \forall f' (f' \in \mathcal{F} \setminus \bigcup_{i=0}^h L_i \rightarrow \mathcal{Z}(f') \geq \mathcal{Z}(f))\}.$$

The highest level within $\mathcal{F}$ is denoted by L_{h_m} and we set

$$\begin{aligned} s(f) &:= |\{f' : f' \in \mathcal{N}_f \wedge \mathcal{Z}(f') > \mathcal{Z}(f)\}|, \\ q(f) &:= |\{f' : f' \in \mathcal{N}_f \wedge \mathcal{Z}(f') = \mathcal{Z}(f)\}|, \\ r(f) &:= |\{f' : f' \in \mathcal{N}_f \wedge (\mathcal{Z}(f') < \mathcal{Z}(f))\}|. \end{aligned}$$

As will be seen later, the $q(f)$ neighbors with the same value of the objective function can be treated in the same way as f itself (we have $q(f) = 0$ in our Example of Boolean conjunctions). Therefore, we will not distinguish between f and these neighbors, see Lemma 2 and the following equations. The introduced notations imply

$$(6) \quad s(f) + q(f) + r(f) = |\mathcal{N}_f| - 1.$$

We consider the probability $\mathbf{a}_f(k)$ to be in the configuration $f \in \mathcal{F}$ after k steps of an *inhomogeneous* Markov chain, i.e., the “temperature” is defined in accordance with (5). We observe that

$$(7) \quad \frac{1}{(k+2)^{(\mathcal{Z}(f) - \mathcal{Z}(f'))/\Gamma}} = e^{-\frac{\mathcal{Z}(f) - \mathcal{Z}(f')}{c(k)}}, \quad k \geq 0.$$

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

By using (1) till (3), one obtains from (4) by straightforward calculations

$$\begin{aligned} \mathbf{a}_f(k) &= \sum_{f' \in \mathcal{N}_f} \mathbf{a}_{f'}(k-1) \cdot \mathbf{Pr}_{c(k)}\{f' \rightarrow f\} \\ &= \mathbf{a}_f(k-1) \cdot \mathbf{Pr}_{c(k)}\{f \rightarrow f\} + \sum_{f' \neq f} \mathbf{a}_{f'}(k-1) \cdot \mathbf{Pr}_{c(k)}\{f' \rightarrow f\} \\ &= \mathbf{a}_f(k-1) \cdot \left(1 - \sum_{f' \neq f} \mathbf{Pr}_{c(k)}\{f \rightarrow f'\}\right) + \\ &\quad + \sum_{f' \neq f} \mathbf{a}_{f'}(k-1) \cdot \mathbf{Pr}_{c(k)}\{f' \rightarrow f\}. \end{aligned}$$

We employ (7), and the resulting equation is given by

$$(8) \quad \mathbf{a}_f(k) = \left(\frac{s(f)+1}{|\mathcal{N}_f|} - \sum_{\substack{i=1 \\ \mathcal{Z}(f_i) \geq \mathcal{Z}(f)}}^{s(f)+q(f)} \frac{(k+1)^{-(\mathcal{Z}(f_i) - \mathcal{Z}(f))}}{|\mathcal{N}_f|} \right) \cdot \mathbf{a}_f(k-1) +$$

$$\begin{aligned}
 & + \sum_{\substack{i=1 \\ \mathcal{Z}(f_i) \geq \mathcal{Z}(f)}}^{s(f)+q(f)} \frac{\mathbf{a}_{f_i}(k-1)}{|\mathcal{N}_{f_i}|} + \\
 & + \sum_{\substack{j=1 \\ \mathcal{Z}(f'_j) < \mathcal{Z}(f)}}^{r(f)} \frac{\mathbf{a}_{f_j}(k-1)}{|\mathcal{N}_{f_j}|} \cdot \frac{1}{(k+1)^{\mathcal{Z}(f)-\mathcal{Z}(f_j)}}.
 \end{aligned}$$

This representation (expansion) will be used in the following as the main relation reducing $\mathbf{a}_f(k)$ to probabilities from previous steps.

Given $f \in \mathcal{F}$, we consider a shortest path of length $\text{dist}(f)$ from f to $\mathcal{F}_{\min}$, i.e., $f' \in \mathcal{F}_{\min}$. We introduce a partition of $\mathcal{F}$ with respect to $\text{dist}(f)$:

$$f \in M_i \iff \text{dist}(f) = i \geq 1, \quad \text{and} \quad \mathcal{M}_{\mathbf{s}} = \bigcup_{i=0}^{\mathbf{s}} M_i,$$

where $M_0 := L_0 = \mathcal{F}_{\min}$. Thus, we distinguish between distance levels M_i related to the minimal number of transitions required to reach an element of $\mathcal{F}_{\min}$ and the levels L_h that are defined by the objective function.

We suppose that $k \geq \mathbf{s}$, and we are going backwards from the k^{th} step. Our aim is to reduce the value

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

to the sum of probabilities from previous steps. First, we consider the expansion of only a particular probability $\mathbf{a}_f(k)$ as shown in (8). At the same step k , the neighbors 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 $\mathbf{a}_{f_j}(k-1)$. Since we consider the entire sum $\sum_f \mathbf{a}_f(k)$, the terms corresponding to a particular $\mathbf{a}_f(k-1)$ can be collected together in the expansion (8) to form a single term. Therefore, we obtain

$$\begin{aligned}
 (9) \quad & \mathbf{a}_f(k-1) \cdot \left\{ \left(\frac{|\mathcal{N}_f| - r(f)}{|\mathcal{N}_f|} - \sum_{\substack{i=1 \\ \mathcal{Z}(f_i) \geq \mathcal{Z}(f)}}^{s(f)+q(f)} \frac{1}{|\mathcal{N}_f|} \cdot \frac{1}{(k+1)^{\mathcal{Z}(f_i)-\mathcal{Z}(f)}} \right) + \right. \\
 & + \sum_{\substack{i=1 \\ \mathcal{Z}(f_i) > \mathcal{Z}(f)}}^{s(f)} \frac{1}{|\mathcal{N}_f|} \cdot \frac{1}{(k+1)^{\mathcal{Z}(f_i)-\mathcal{Z}(f)}} + \sum_{\substack{j=1 \\ \mathcal{Z}(f_j) < \mathcal{Z}(f)}}^{r(f)} \frac{1}{|\mathcal{N}_f|} \Big\} \\
 & = \mathbf{a}_f(k-1).
 \end{aligned}$$

However, it is important to distinguish between elements from M_1 and elements from M_i , $i \geq 2$: If $f \in M_1$, the neighbors from M_0 do not generate terms containing probabilities from higher levels because the $\mathbf{a}_{f'}(k-i)$ are *not* expanded during the particular steps of the recursive procedure since they are not present

in the sum $\sum_{f \notin M_0} \mathbf{a}_f(k)$. If $f \in M_1$, the terms related to $\mathcal{Z}(f_j) < \mathcal{Z}(f)$, where $f_j \in M_0$, are missing, and therefore the following arithmetic term is generated:

$$(10) \quad \left(1 - \frac{r'(f)}{|\mathcal{N}_f|}\right) \cdot \mathbf{a}_f(k-1),$$

where $r'(f) \leq r(f)$ is the number of neighbors from M_0 . We introduce the following abbreviations:

$$\varphi(f', f, v) := \frac{1}{(k+2-v)^{\mathcal{Z}(f')-\mathcal{Z}(f)}}, \quad D_f(k-v) := \frac{s(f)+1}{|\mathcal{N}_f|} - \sum_{i=1}^{s(f)} \frac{\varphi(f', f, v)}{|\mathcal{N}_f|}.$$

The relations expressed in (8) till (10) can be summarized to

Lemma 1 *A single step of the expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$ results in*

$$\begin{aligned} \sum_{f \notin M_0} \mathbf{a}_f(k) &= \sum_{f \notin M_0} \mathbf{a}_f(k-1) - \sum_{f \in M_1} \frac{r'(f)}{|\mathcal{N}_f|} \cdot \mathbf{a}_f(k-1) + \\ &\quad + \sum_{f \in M_1} \sum_{\substack{j=1 \\ f_j \in M_0 \cap \mathcal{N}_f}}^{r'(f)} \frac{\varphi(f, f_j, 1)}{|\mathcal{N}_{f_j}|} \cdot \mathbf{a}_{f_j}(k-1). \end{aligned}$$

Sketch of Proof: We obtain from (9) and (10):

$$\begin{aligned} \sum_{f \notin M_0} \mathbf{a}_f(k) &= \sum_{f \notin M_0 \cup M_1} \mathbf{a}_f(k) + \sum_{f \in M_1} \mathbf{a}_f(k) \\ &= \sum_{f \notin M_0 \cup M_1} \mathbf{a}_f(k-1) + \sum_{f \in M_1} \left(1 - \frac{r'(f)}{|\mathcal{N}_f|}\right) \cdot \mathbf{a}_f(k-1) + \\ &\quad + \sum_{f \in M_1} \sum_{\substack{j=1 \\ f_j \in M_0 \cap \mathcal{N}_f}}^{r'(f)} \frac{\varphi(f, f_j, 1)}{|\mathcal{N}_{f_j}|} \cdot \mathbf{a}_{f_j}(k-1). \end{aligned}$$

The sum containing $\varphi(f, f_j, 1)/|\mathcal{N}_{f_j}|$ is generated from $\mathbf{a}_f(k)$, $f \in M_1$, in accordance with (8). The negative product $\mathbf{a}_f(k-1) \cdot r(f)/|\mathcal{N}_f|$ is taken separately, and one obtains the stated equation. **q.e.d.**

The diminishing factor $(1-r(f)/|\mathcal{N}_f|)$ 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 neighbor from M_{i-1} . The main task is now to analyze how this diminishing factor changes, if it is transmitted to the next higher distance level. We denote

$$(11) \quad \sum_{f \notin M_0} \mathbf{a}_f(k) = \sum_{f \notin M_0} \mu(f, v) \cdot \mathbf{a}_f(k-v) + \sum_{f' \in M_0} \mu(f', v) \cdot \mathbf{a}_{f'}(k-v),$$

i.e., the coefficients $\mu(\tilde{f}, v)$ are the factors at probabilities after v steps of an expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$. For $f \in M_1$ we have $\mu(f, 1) = 1 - r'(f)/|\mathcal{N}_f|$,

and $\mu(f, 1) = 1$ for the remaining $f \in \mathcal{M}_s \setminus (M_0 \cup M_1)$. For $f' \in M_0$, Lemma 1 implies:

$$(12) \quad \mu(f', 1) = \sum_{\substack{i=1 \\ f_i \in M_1 \cap \mathcal{N}_{f'}}}^{s(f')} \frac{\varphi(f_i, f', 1)}{|\mathcal{N}_{f'}|}$$

Starting from step $(k-1)$, the generated probabilities $\mathbf{a}_{f'}(k-u)$ are expanded in the same way as all other probabilities.

Except for the initial values, the coefficients $\mu(f, v)$, $f \notin M_0$, and $\mu(f', v)$, $f' \in M_0$, can be treated in the same way because the expansion (8) is valid for both $\mathbf{a}_f(k)$ and $\mathbf{a}_{f'}(k)$. Therefore, we can apply the same considerations which led to equation (9) to $\mu(f, v) \cdot \mathbf{a}_f(k-v)$ and $\mu(f', v) \cdot \mathbf{a}_{f'}(k-v)$ from (11). We perform an inductive step from $(k-v+1)$ to $(k-v)$ and we suppose that all terms within $\{\dots\}$ of (9) are multiplied by the corresponding $\mu(f, v-1)$:

$$\begin{aligned} & \mathbf{a}_f(k-v) \cdot \left\{ \mu(f, v-1) \cdot \left(\frac{|\mathcal{N}_f| - r(f)}{|\mathcal{N}_f|} - \sum_{\substack{i=1 \\ f_i > f}}^{s(f)} \frac{(k+1)^{-(\mathcal{Z}(f_i) - \mathcal{Z}(f))}}{|\mathcal{N}_f|} \right) + \right. \\ & + \sum_{\substack{i=1 \\ f_i \geq f}}^{s(f)+q(f)} \frac{\mu(f_i, v-1)}{|\mathcal{N}_f|} \cdot \frac{1}{(k+1)^{\mathcal{Z}(f_i) - \mathcal{Z}(f)}} + \sum_{\substack{j=1 \\ f_j < f}}^{r(f)} \frac{\mu(f_j, v-1)}{|\mathcal{N}_f|} \left. \right\} \\ & = \mathbf{a}_f(k-v) \cdot \mu(f, v). \end{aligned}$$

We obtain for $v \geq 2$:

$$\mu(f, v) = \mu(f, v-1) \cdot D_f(k-v) + \sum_{f' \leq f} \frac{\mu(f', v-1)}{|\mathcal{N}_f|} + \sum_{f'' > f} \frac{\mu(f'', v-1)}{|\mathcal{N}_f|} \cdot \varphi(f'', f, v).$$

It is important to note that the summands are divided by the same value $|\mathcal{N}_f|$, see (9). We set $\mu(f, j) := 1 - \nu(f, j)$ because we are interested in the convergence of both values $\mu(f, j)$ and $\nu(f, j)$ to zero. Substituting $\mu(\tilde{f}, j)$ by $1 - \nu(\tilde{f}, j)$ and rearranging both sides leads to the same structure of the equation as for $\mu(\tilde{f})$:

Lemma 2 *The following recurrent relation is valid for the coefficients $\nu(f, v)$:*

$$\nu(f, v) = \nu(f, v-1) \cdot D_f(k-v) + \sum_{f' \leq f} \frac{\nu(f', v-1)}{|\mathcal{N}_f|} + \sum_{f'' > f} \frac{\nu(f'', v-1)}{|\mathcal{N}_f|} \cdot \varphi(f'', f, v).$$

Furthermore,

$$\nu(f, i) = \begin{cases} 0, & \text{if } f \in M_j, j > i; \\ \frac{r(f)}{|\mathcal{N}_f|}, & \text{if } f \in M_1, i = 1; \\ 1 - \sum_{\substack{l=1 \\ f_l \in M_1 \cap \mathcal{N}_f}}^{s(f)} \frac{\varphi(f_l, f, 1)}{|\mathcal{N}_f|}, & \text{if } f \in M_0, i = 1. \end{cases}$$

We note that the representation from Lemma 2 is a linear function with respect to the values $\nu(f, v)$, i.e., once an arithmetic term f has been added (or generated) by the previous value of $\nu(f, v-1)$ or “neighbouring” values $\nu(f', v-1)$, this term is multiplied by factors that only depend either on structural properties of the neighborhood relation (that is the factor $(s(f) + 1)/|\mathcal{N}_f|$ from D_f) or on differences of the objective function which are expressed by $\varphi(f', f, v)$. The differences $\mathcal{Z}(f') - \mathcal{Z}(f)$ do not change during the recursive calculation of values $\nu(f, v)$; the only parameter that changes in $\varphi(f', f, v)$ is v . Therefore, any $\nu(f, v)$ and $\mu(f', v)$ can be expressed by a sum $\sum_i T_i$ of arithmetic terms.

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 $\nu(f, v-1) = \sum_u T_u$. Since there are no $f'' < f'$ for $f' \in M_0$, we obtain:

$$\begin{aligned} \mu(f', v) &= D_{f'}(k - v) \cdot \sum_{u'} T'_{u'}(f') + \sum_{f > f'} \frac{1 - \sum_u T_u(f)}{|\mathcal{N}_{f'}|} \cdot \varphi(f, f', v) \\ &= \sum_{f > f'} \frac{\varphi(f, f', v)}{|\mathcal{N}_{f'}|} + D_{f'}(k - v) \cdot \sum_{u'} T'_{u'}(f') - \sum_{f > f'} \frac{\sum_u T_u(f)}{|\mathcal{N}_{f'}|} \cdot \varphi(f, f', v). \end{aligned}$$

As can be seen, the values $\sum_{f > f'} \varphi(f, f', v)/|\mathcal{N}_{f'}|$ are generated at each time step with the corresponding $\varphi(f, f', v)$. The remaining summands of $\mu(f', v)$ are individual arithmetic terms f_u and $f'_{u'}$ from the previous step, multiplied by a positive or a negative factor.

In the same way we obtain for elements of M_1 :

$$\begin{aligned} \nu(f, v) &= \frac{r(f)}{|\mathcal{N}_f|} + \nu(f, v-1) \cdot D_f(k - v) + \sum_{f'' > f} \frac{\nu(f'', v-1)}{|\mathcal{N}_f|} \cdot \varphi(f'', f, v) - \\ &\quad - \sum_{\substack{f' < f \\ f' \in M_0}} \frac{\sum_{u'} T'_{u'}(f')}{|\mathcal{N}_f|}, \end{aligned}$$

where $r(f)/|\mathcal{N}_f|$ is from $(1 - (s(f) + 1)/|\mathcal{N}_f|)$. Again, the sign changes only for terms $f'_{u'}(f')$ because $D_f(k - v) > 0$. 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 $\sum_{f > f'} \varphi(f, f', v)/|\mathcal{N}_{f'}|$ in all $\mu(f', v)$, $f' \in M_0$. Therefore, all arithmetic terms f are derived from expressions of the type $r(f)/|\mathcal{N}_f|$ and $\sum_{f > f'} \varphi(f, f', v)/|\mathcal{N}_{f'}|$. This justifies the following

Definition 3 *The expressions $r(f)/|\mathcal{N}_f|$, $f \in M_1$, and $\sum_{f > f'} \varphi(f, f', v)/|\mathcal{N}_{f'}|$, $f' \in M_0$, are called source terms of $\nu(f, v)$ and $\mu(f', v)$, respectively.*

During an expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$ backwards according to (8), the source terms are distributed permanently to higher distance levels M_j . That means, in the same way as for M_1 , the calculation of $\nu(f, v)$ is repeated almost identically at any step, only the “history” of generations becomes longer. Therefore, at

higher distance levels the notion of a source term can be defined by an inductive step:

Definition 4 *If $\tilde{f} \in M_i$, $i > 1$, any term that is generated according to the equation of Lemma 2 from a source term of $\nu(f, v-1)$, where $f \in M_{i-1} \cap \mathcal{N}_{\tilde{f}}$, is said to be a source term of $\nu(\tilde{f}, v)$.*

We introduce a counter $\mathbf{r}(T)$ to terms f that indicates the step at which the term has been generated from source terms. The value $\mathbf{r}(T)$ is called the rate of a term and we set $\mathbf{r}(T) = 1$ for source terms f .

The value $\mathbf{r}(T) > 1$ is assigned to terms related to M_0 and M_1 in a slightly different way compared to higher distance levels because at the first step the $f' \in M_0$ do not participate in the expansion of $\sum_{f \notin M_0} \mathbf{a}_f(k)$. Furthermore, in the case of M_0 and M_1 we have to take into account the changing signs of terms that result from the simultaneous consideration of $\nu(f, v)$ (for M_1) and $\mu(f', v)$ (for M_0). Basically, a rate $\mathbf{r}(T) > 1$ is assigned to a term that was at the preceding step an $(\mathbf{r}(T) - 1)$ -term or $(\mathbf{r}(T) - 2)$ -term at the same or a higher distance level, respectively, or an $\mathbf{r}(T)$ -term at a lower distance level.

Definition 5 *A term f' is called a j^{th} rate term of $\mu(f', v)$, $f' \in M_0$ and $j \geq 2$, if either $f' = -T$ and $\mathbf{r}(T) = j - 1$ for some $\nu(f, v - 1)$, $f \in M_1 \cap \mathcal{N}_{f'}$, or $\mathbf{r}(T') = j - 1$ for some $\mu(f'', v - 1)$, $f'' \in M_0 \cap \mathcal{N}_{f'}$.*

A term f is called a j^{th} rate term of $\nu(f, v)$, $f \in M_1$ and $j \geq 2$, if $\mathbf{r}(T) = j - 2$ for some $\nu(\tilde{f}, v - 1)$, $\tilde{f} \in M_2 \cap \mathcal{N}_f$, $\mathbf{r}(T) = j - 1$ for some $\nu(\tilde{f}, v - 1)$, $\tilde{f} \in M_1 \cap \mathcal{N}_f$, or $f = -T'$ and $\mathbf{r}(T') = j - 1$ for some $f' \in M_0 \cap \mathcal{N}_f$ with respect to $\mu(f', v - 1)$.

A term f is called a j^{th} rate term of $\nu(\tilde{f}, v)$, $\tilde{f} \in M_i$ and $i, j \geq 2$, if $\mathbf{r}(T) = j - 1$ for some $\nu(\tilde{f}', v - 1)$, $\tilde{f}' \in M_{i+1} \cap \mathcal{N}_{\tilde{f}}$, $\mathbf{r}(T) = j - 1$ for some $\nu(\tilde{f}', v - 1)$, $\tilde{f}' \in M_i \cap \mathcal{N}_{\tilde{f}}$, or f is a j^{th} rate term of $\nu(f, v - 1)$ for some $f \in M_{i-1}$.

The classification of terms will be used for a partition of the summation over all terms which constitute particular values $\nu(f, v)$ and $\mu(f', v)$. Let $\mathcal{T}_j(\tilde{f}, v)$ be the set of j^{th} rate arithmetic terms of $\nu(f, v)$ ($\mu(f', v)$) related to $\tilde{f} \in \mathcal{M}_{\mathbf{s}}$. We set

$$(13) \quad \mathbf{S}_j(\tilde{f}, v) := \sum_{f \in \mathcal{T}_j(\tilde{f}, v)} T.$$

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

$$(14) \quad \nu(\tilde{f}, v) = \sum_{j=1}^{v-i+1} \mathbf{S}_j(\tilde{f}, v) \quad \text{and} \quad \mu(f', v) = \sum_{j=1}^v \mathbf{S}_j(f', v).$$

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 , respectively. Since the source terms depend on v only,

see Definition 3, we obtain from the simple relation $k_2 - (k_2 - k_1 + v) = k_1 - v$ the following equality:

Lemma 3 *Given $k_2 \geq k_1$ and $1 \leq j \leq k_1$, then*

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

We recall that our main goal is to upper bound the sum $\sum_{f \notin M_0} \mathbf{a}_f(k)$. When $\mathbf{a}(0)$ denotes the initial probability distribution, we can derive

$$(15) \quad \left| \sum_{f \notin M_0} \mathbf{a}_f(k_2) - \sum_{f \notin M_0} \mathbf{a}_f(k_1) \right| \leq \sum_{f \notin M_0} |(\nu(f, k_2) - \nu(f, k_1)) \cdot \mathbf{a}_f(0)| + \\ + \sum_{f' \in M_0} |(\mu(f', k_2) - \mu(f', k_1)) \cdot \mathbf{a}_{f'}(0)|.$$

Any element of the initial distribution $\mathbf{a}_f(0)$ is simply replaced by 1 and we obtain for the first part of the sum in accordance with Lemma 3:

$$(16) \quad \sum_{f \notin M_0} |(\nu(f, k_2) - \nu(f, k_1)) \cdot \mathbf{a}_f(0)| \leq \sum_{f \notin M_0} \sum_{j=k_1+1}^{k_2} |\mathbf{S}_j^2(f, k_2)|.$$

For $\tilde{f} \in M_i \neq M_1, M_0$ and $j \geq 2$ we obtain immediately from Lemma 2 and Definition 5:

$$(17) \quad \mathbf{S}_j(\tilde{f}, v) = \mathbf{S}_{j-1}(\tilde{f}, v-1) \cdot D_{\tilde{f}}(k-v) + \sum_{\substack{\tilde{f}' < \tilde{f} \\ \tilde{f}' \in M_{i+1}}} \frac{\mathbf{S}_{j-2}(\tilde{f}', v-1)}{|\mathcal{N}_{\tilde{f}}|} + \\ + \sum_{\substack{\tilde{f}'' > \tilde{f} \\ \tilde{f}'' \in M_{i+1}}} \frac{\mathbf{S}_{j-2}(\tilde{f}'', v-1)}{|\mathcal{N}_{\tilde{f}}|} \cdot \varphi(\tilde{f}'', \tilde{f}, v) + \sum_{\substack{\tilde{f}' < \tilde{f} \\ \tilde{f}' \in M_{i-1}}} \frac{\mathbf{S}_j(\tilde{f}', v-1)}{|\mathcal{N}_{\tilde{f}}|} + \\ + \sum_{\substack{\tilde{f}' < \tilde{f} \\ \tilde{f}' \in M_{i-1}}} \frac{\mathbf{S}_j(\tilde{f}', v-1)}{|\mathcal{N}_{\tilde{f}}|} + \sum_{\substack{\tilde{f}'' > \tilde{f} \\ \tilde{f}'' \in M_{i-1}}} \frac{\mathbf{S}_j(\tilde{f}'', v-1)}{|\mathcal{N}_{\tilde{f}}|} \cdot \varphi(\tilde{f}'', \tilde{f}, v).$$

In case of $f \in M_1$ and $j \geq 2$, we have in accordance with Definition 5:

$$(18) \quad \mathbf{S}_j(f, v) = \mathbf{S}_{j-1}(f, v-1) \cdot D_f(k-v) + \\ + \sum_{\substack{\tilde{f}'' > f \\ \tilde{f}'' \in M_2}} \frac{\mathbf{S}_{j-2}(\tilde{f}'', v-1)}{|\mathcal{N}_f|} \cdot \varphi(\tilde{f}'', t, v) - \sum_{f' \in M_0 \cap \mathcal{N}_f} \frac{\mathbf{S}_{j-1}(f', v-1)}{|\mathcal{N}_f|}.$$

Finally, the corresponding relation for $f' \in M_0$ is given by

$$(19) \quad \mathbf{S}_j(f', v) = \mathbf{S}_{j-1}(f', v-1) \cdot D_{f'}(k-v) - \sum_{f > f'} \frac{\mathbf{S}_{j-1}(f, v-1)}{|\mathcal{N}_{f'}|} \cdot \varphi(f, f', v).$$

The recursive application of (17) till (19) generates a tree-like structure $\mathbf{T}(f, j, v)$, where a particular $\mathbf{S}_j(f, v)$ represents the root of the tree. The leaves are source terms from M_1 and M_0 after a full recursion of v steps.

Positive and negative terms are considered separately, and a uniform upper bound of $|\mathbf{S}_j(f, k_2)|$ is calculated from the leaves to the root. Here, we distinguish between M_1 and M_0 . The terms from M_0 are multiplied by an extra factor $\varphi(f_1, f, 1)/|\mathcal{N}_f|$, and therefore it is sufficient to consider the terms generated by elements of M_1 .

Since any element of M_1 has a neighbour from M_0 and the terms arising from the lowest level are treated separately, the terms from M_0 (i.e., generated by f itself and neighbours $\mathcal{N}_f \cap M_0$) are multiplied by a factor smaller than 1. At each step towards the root, we take the values from elements of M_1 as a new summand to the terms generated at higher levels M_i , $i > 1$. Therefore, the already existing summands are also multiplied by a factor smaller than 1, when summands from neighbouring elements with the same generation step are taken together. The factor is determined by $D' := D_f - (s_{i-1} + r_{i-1})/n$, when n denotes for short an upper bound for $|\mathcal{N}_f|$.

Hence, the analysis of terms can be reduced to the terms (new summands) generated at level M_0 . The particular products (summands) can be classified by the number a of self-transitions (i.e., factors $D_f(k - v)$), the number b of transitions to higher levels (factors $\varphi(f', f, v)$), and c steps to the same level or lower levels of the objective function. Thus, we have $a + b + c = v$ for the three types of transitions. For fixed a and b , the products can be upper bounded by

$$(20) \quad \binom{v}{a} \cdot \binom{v-a}{2 \cdot b} \cdot \binom{2 \cdot b}{b} \cdot \max_{i,j} \prod_{i=1}^b \left(\frac{t_i}{(n)^2} \cdot \frac{1}{(k+2-u_i)^\omega} \right) \cdot \prod_{j=1}^a D'_{f_j}(k-u_j),$$

where $t_i \leq (n-1)^2$ and ω denotes the uniform difference between neighbouring elements (see the remark after Theorem 1). We employ the STIRLING formula for the following estimation:

$$(21) \quad \prod_{u=1}^b \varphi(f_u, f'_u, v_u) \leq \prod_{u=1}^b \frac{1}{(k+2-v_u)^\omega} \leq \frac{1}{(b!)^\omega} \leq \left(\frac{e}{b}\right)^{b \cdot \omega}.$$

The product (20) is analysed for both cases $a \leq v/2$ and $a > v/2$ separately. In the first case, $v \geq 2^{O(\Gamma)}$ is sufficient to obtain an exponentially small bound for (20). In the second case, we need $k \geq n^{O(\Gamma)}$ for the factors $\varphi(f_u, f'_u, v_u)$ to be sufficiently small in order to compensate the binomial coefficients.

The analysis of (20) leads to a uniform upper bound of $|\mathbf{S}_j(f, k_2)|$ and we obtain

Lemma 4 *Given the maximum escape depth Γ , then $n^{O(n)} \geq k_2 \geq k_1 > n^{O(\Gamma)}$ implies*

$$\left| \sum_{f \notin M_0} (\nu(f, k_2) - \nu(f, k_1)) \cdot \mathbf{a}_f(0) \right| < k_2^{O(1)} \cdot 2^{-k_1^{O(1-1/\Gamma)}}.$$

In the same way, the corresponding upper bound for $(\mu(f', k_1) - \mu(f', k_2))$ is derived.

By our assumption about the size of $\mathcal{F}$ (see Section 2), a value of $k_2 < n^{O(n)}$ allows a complete search of $\mathcal{F}$. We assume that such an upper bound is sufficient for a convergence according to Theorem 1. Now, we can immediately prove

Theorem 2 *When the convergence is guaranteed for $k_2 \leq n^{O(n)}$, then $k > n^{O(\Gamma)} + \log^{O(1)} 3/\varepsilon$ implies for arbitrary initial probability distributions $\mathbf{a}(0)$ and $\varepsilon > 0$ the relation*

$$\sum_{\tilde{f} \notin M_0} \mathbf{a}_{\tilde{f}}(k) < \varepsilon \quad \text{and therefore,} \quad \sum_{f' \in M_0} \mathbf{a}_{f'}(k) \geq 1 - \varepsilon.$$

Sketch of Proof: We use the following representation:

$$\begin{aligned} \sum_{\tilde{f} \notin M_0} \mathbf{a}_{\tilde{f}}(k) &= \sum_{\tilde{f} \notin M_0} (\mathbf{a}_{\tilde{f}}(k) - \mathbf{a}_{\tilde{f}}(k_2)) + \sum_{\tilde{f} \notin M_0} \mathbf{a}_{\tilde{f}}(k_2) \\ &= \sum_{\tilde{f} \notin M_0} (\nu(\tilde{f}, k_2) - \nu(\tilde{f}, k)) \cdot \mathbf{a}_{\tilde{f}}(0) + \\ &\quad + \sum_{f' \in M_0} (\mu(f', k) - \mu(f', k_2)) \cdot \mathbf{a}_{f'}(0) + \sum_{\tilde{f} \notin M_0} \mathbf{a}_{\tilde{f}}(k_2). \end{aligned}$$

The value k_2 from Lemma 4 is larger but independent of $k_1 = k$, i.e., we can take a $k_2 > k$ such that

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

We employ Theorem 1, i.e., if the constant Γ from (5) is sufficiently large, the inhomogeneous simulated annealing procedure defined by (1), (2), and (3) tends to the global minimum of $\mathcal{Z}$ on $\mathcal{F}$. We obtain the stated inequality, if additionally both differences $\sum_{\tilde{f} \notin M_0} (\nu(\tilde{f}, k_2) - \nu(\tilde{f}, k))$ and $\sum_{f' \in M_0} (\mu(f', k) - \mu(f', k_2))$ are smaller than $\varepsilon/3$. Lemma 4 implies that the condition on the differences is satisfied in case of

$$2^{-k^{1/\gamma}} < \frac{\varepsilon}{3},$$

which is valid, if $\log^\gamma(3/\varepsilon) < k$.

q.e.d.

4 Concluding Remarks

We analysed the convergence of logarithmic cooling schedules in terms of the underlying energy landscape. Under natural assumption about the configuration space we succeeded to prove a lower bound on the number of cooling steps that are sufficient to approach with probability $1-\varepsilon$ the set of optimum solutions. The lower bound is given by $n^{O(\Gamma)} + \log^{O(1)}(1/\varepsilon)$, where Γ is the maximum escape depth from local minima and n describes the size of the neighbourhood of configurations. When Γ can be calculated or estimated, the result can provide good or improved upper bounds for the time complexity of stochastic algorithms.

References

1. S. Azencott (editor). *Simulated Annealing: Parallelization Techniques*. Wiley & Sons, New York, 1992.
2. E.H.L. Aarts and J.H.M. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach*, Wiley & Sons, New York, 1989.
3. A. Albrecht, S.K. Cheung, K.S. Leung, and C.K. Wong. Stochastic Simulations of Two-Dimensional Composite Packings. *J. of Computational Physics*, 136(2):559 – 579, 1997.
4. O. Catoni. Rough Large Deviation Estimates for Simulated Annealing: Applications to Exponential Schedules. *Annals of Probability*, 20(3):1109 – 1146, 1992.
5. O. Catoni. Metropolis, Simulated Annealing, and Iterated Energy Transformation Algorithms: Theory and Experiments. *J. of Complexity*, 12(4):595 – 623, 1996.
6. V. Černý. 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).
7. B. Hajek. Cooling Schedules for Optimal Annealing. *Mathem. Oper. Res.*, 13:311 – 329, 1988.
8. M. Kearns, M. Li, L. Pitt, and L.G. Valiant. Recent Results on Boolean Concept Learning. In *Proc. 4th 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. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. of Chemical Physics*, 21(6):1087 – 1092, 1953.
11. S. Rajasekaran and J.H. Reif. Nested Annealing: A Provable Improvement to Simulated Annealing. *J. of Theoretical Computer Science*, 99(1):157-176, 1992.
12. F. Romeo and A. Sangiovanni-Vincentelli. A Theoretical Framework for Simulated Annealing. *Algorithmica*, vol. 6, no. 3, pp. 302 – 345, 1991.
13. E. Seneta. *Non-negative Matrices and Markov Chains*. Springer-Verlag, New York, 1981.
14. A. Sinclair and M. Jerrum. Approximate Counting, Uniform Generation, and Rapidly Mixing Markov Chains. *Information and Computation*, 82:93 – 133, 1989.
15. A. Sinclair and M. Jerrum. Polynomial-Time Approximation Algorithms for the Ising Model. *SIAM J. Comput.*, 22(5):1087 – 1116, 1993.
16. G. Sorkin. Efficient Simulated Annealing on Fractal Energy Landscapes. *Algorithmica*, 6:367-418, 1991.
17. K. Verbeurgt. Learning DNF under the Uniform Distribution in Quasi-Polynomial Time. In *Proc. of the 3rd Annual Workshop on Computational Learning Theory*, pp. 314 – 326, 1990.