Simulated Annealing Beats Metropolis in Combinatorial Optimization

Ingo Wegener*

FB Informatik, LS2, Univ. Dortmund, Germany
ingo.wegener@uni-dortmund.de

Abstract. The Metropolis algorithm is simulated annealing with a fixed temperature. Surprisingly enough, many problems cannot be solved more efficiently by simulated annealing than by the Metropolis algorithm with the best temperature. The problem of finding a natural example (artificial examples are known) where simulated annealing outperforms the Metropolis algorithm for all temperatures has been discussed by Jerrum and Sinclair (1996) as "an outstanding open problem." This problem is solved here. The examples are instances of the well-known minimum spanning tree problem. Moreover, it is investigated which instances of the minimum spanning tree problem can be solved efficiently by simulated annealing. This is motivated by the aim to develop further methods to analyze the simulated annealing process.

1 Introduction

Simple randomized search heuristics like randomized local search (RLS), the Metropolis algorithm (MA), simulated annealing (SA), evolutionary algorithms (EA), and genetic algorithms (GA) find many applications. One cannot hope that they outperform sophisticated problem-specific algorithms for well-studied problems. They are easy to implement and good alternatives if one does not know efficient problem-specific algorithms and if one shies away from developing a clever algorithm. They are the tool of choice in black-box optimization where the problem instance is hidden from the algorithm. And they are useful as parts of hybrid algorithms combining general search principles with problem-specific modules.

Hence, it is interesting to understand the working principles behind these heuristics. The aim is to analyze the expected optimization time and the success probability within a given time bound of heuristics applied to specific problems. Up to now there are not many of such results. One reason is that the heuristics are not designed to support their analysis (in contrast to many problem-specific

* Supported in part by the Deutsche Forschungsgemeinschaft (DFG) as part of the Collaborative Research Center "Computational Intelligence" (SFB 531) and by the German-Israeli Foundation (GIF) in the project "Robustness Aspects of Algorithms".

L. Caires et al. (Eds.): ICALP 2005, LNCS 3580, pp. 589–601, 2005.

algorithms). To simplify the problem many authors have first investigated quite artificial problems hoping to develop methods which can be used in many other situations.

Here, we are interested in simulated annealing and the Metropolis algorithm (which can be defined as SA with a fixed temperature). Both algorithms are defined in Section 2. It is an obvious question how to use the freedom to choose a cooling schedule for SA and whether this option is essential. Little is known about this leading Jerrum and Sinclair (1996, page 516) to the following statement: "It remains an outstanding open problem to exhibit a natural example in which simulated annealing with any non-trivial cooling schedule provably outperforms the Metropolis algorithm at a carefully chosen fixed value of α." In their paper, α is the temperature. The notion of a "natural example" is vague, but the known examples are obviously artificial. Sorkin (1991) has proven the considered effect for a so-called fractal energy landscape. The chaotic behavior of this function asks for different temperatures in different phases of the search. The artificial example due to Droste, Jansen, and Wegener (2001) allows a simpler analysis.

Jerrum and Sorkin (1998) have analyzed the Metropolis algorithm for the graph bisection problem. They focus the interest on problems from combinatorial optimization: "Unfortunately no combinatorial optimization problem that has been subjected to rigorous theoretical analysis has been exhibited this phenomenon: those problems that can be solved efficiently by simulated annealing can be solved just as effectively by 'annealing' at a single carefully selected temperature. A rigorous demonstration that annealing is provably beneficial for some natural optimization problems would rate as a significant theoretical advance."

Our problem of choice is the minimum spanning tree problem (MSTP) which is contained in all textbooks on combinatorial optimization and should be accepted as "natural optimization problem." It should be obvious that SA cannot beat MA for each problem instance. E. g., for graphs where all edge weights equal 1 the frozen MA (at temperature 0) cannot be beaten by SA. In Section 3, we describe the notion of efficiency for randomized search heuristics and, in Section 4, we describe simple instances of the MSTP where SA outperforms MA. The underlying graphs will be so-called connected triangles (CT), see Figure 1.

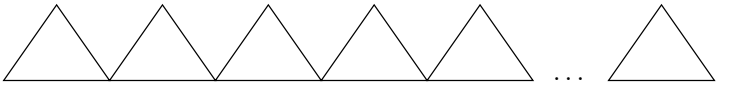

Fig. 1. Graphs called connected triangles.

The idea is to produce examples as simple as possible. This allows proofs which can be taught in introductory courses on randomized search heuristics. Afterwards, we try to understand which instances of the MSTP can be solved efficiently by SA and MA, only by SA, or by none of them. Weights $w_1, \ldots, w_m$ are called $(1+\varepsilon)$-separated if $w_i > w_j$ implies $w_i \geq (1+\varepsilon) \cdot w_j$. For each $\varepsilon(m) = o(1)$ there are graphs with $(1+\varepsilon(m))$-separated weights such that SA

cannot attack them efficiently (Section 5). For each constant $\varepsilon > 0$, SA can attack all graphs with $(1 + \varepsilon)$-separated weights efficiently (Section 6). These results imply that SA outperforms MA on a much larger class of graphs than the connected triangles discussed in Section 4. We finish with some conclusions.

It should be obvious that we do not hope that SA or MA beats the well-known algorithms due to Kruskal and to Prim. Again we like to transfer a statement of Jerrum and Sorkin (1998) from minimum bisections to minimum spanning trees (MSTs): "Our main contribution is not, then, to provide a particularly effective algorithm for the minimum bisection problem ..., but to analyze the performance of a popular heuristic applied to a reasonably realistic problem in combinatorial optimization."

2 Metropolis Algorithm, Simulated Annealing, and Minimum Spanning Trees

An instance of the MSTP consists of an undirected graph $G = (V, E)$ with n vertices and m edges and a weight $w(e)$ for each edge e. Weights are positive integers. The problem is to find an edge set E' connecting all vertices with minimal total weight. The edges are numbered and edge sets are described as characteristic vectors, i. e., $x \in \{0,1\}^m$ describes the set of edges e_i where $x_i = 1$. This formalization is well-suited for MA and SA.

We describe the Metropolis algorithm with temperature T for minimization problems on $\{0,1\}^m$. The first search point x is chosen in some way discussed later. Each round of an infinite loop consists of local change and selection.

Local change: Let x be the current search point. Choose $i \in \{1, \ldots, m\}$ uniformly at random and flip x_i, i. e., let $x' = (x'_1, \ldots, x'_m)$ where $x'_j = x_j$, if $j \neq i$, and $x'_i = 1 - x_i$.

Selection of the new current search point with respect to a fitness function f:
if $f(x') \leq f(x)$: select x',
if $f(x') > f(x)$: select x' with probability $\exp\{-(f(x') - f(x))/T\}$, otherwise select x.

We have to discuss some details in order to ensure that our results are not based on too special choices. Randomized search heuristics do not produce a certificate that a search point is optimal. Therefore, the algorithm contains an infinite loop, but the run time is defined as the number of rounds until an optimal search point is produced. A round cannot be performed in time $O(1)$ but quite efficiently and people have agreed to count the number of rounds.

We choose 1^m as starting point. This is similar to the choice 0^m for the maximum matching problem (Sasaki and Hajek (1988)) and the maximum clique problem (Jerrum (1992)). The starting points are the worst legal solutions. This choice of the starting point implies that we cannot apply the lower bound technique for MA due to Sasaki (1991) which ensures only the existence of some bad starting point. It would be an alternative to start with a search point chosen uniformly at random. For many graphs, we then choose a non-legal solution (an

unconnected graph) and the fitness function has to contain hints directing the search to legal search points. It is not difficult to obtain similar results in this situation. However, most papers on MA and SA only work on legal search points.

We have chosen the fitness function f where $f(x) = \infty$ for search points x describing unconnected graphs and where $f(x)$ is the total weight of all chosen edges if x describes a connected graph. Unconnected graphs are never accepted as current search points. This again is in accordance with Sasaki and Hajek (1988) and Jerrum (1992). All search points are legal solutions in the graph bisection problem and therefore Jerrum and Sorkin (1993, 1998) start with randomly chosen search points.

We follow Sasaki and Hajek (1988) and Jerrum (1992) in allowing only 1-bit neighborhoods. Neumann and Wegener (2004) have analyzed RLS with 1-bit and 2-bit flips (RLS equals the frozen MA at temperature $T = 0$) and a simple EA for the MSTP. These algorithms do not select new search points which are worse than the old one. Hence, their search strategy is completely different from the strategy applied by MA and SA that have to accept sometimes worsenings to find the optimum. Flips of two bits allow to include an edge into a tree and to exclude simultaneously an edge of the newly created cycle. RLS and the simple EA find an MST in an expected number of $O(m^2(\log m + \log w_{\max}))$ steps, where $w_{\max}$ denotes the maximal weight. Note that we are not looking for a "best" algorithm for the MSTP. The main idea of an elitist EA is to reject worsenings and to escape from local optima by non-local steps. The main idea of MA and SA is to work with very local steps and to escape from local optima by accepting worsenings. The situation here is similar to the case of maximum matchings where also flips of 2 bits are helpful to shorten augmenting paths, compare Sasaki and Hajek (1988) who analyze SA with 1-bit flips only and Giel and Wegener (2003) who analyze RLS with 1-bit and 2-bit flips and a simple EA.

Finally, we introduce SA based on a cooling schedule $T(t)$. The initial temperature $T(1)$ may depend on m and the largest possible weight w_{max}. The temperature $T(t)$ applied by the selection operator in step t equals $\alpha^{t-1} \cdot T(1)$, where $\alpha < 1$ is a constant which may depend on m and an upper bound on w_{max}. This cooling schedule does not include any knowledge about the problem instance. We use a kind of "continuous cooling", other possibilities are longer phases with a constant temperature or dynamic cooling schedules that depend on the success rate (where a step is called successful if x' is selected) or the rate of f-improving steps.

3 Efficiency Measures

There are many well-known convergence results on MA and SA. We want to distinguish "efficient behavior" from non-efficient one. The first idea is to define efficiency as expected polynomial time. We think that this is not a good choice. There may be a small probability of missing a good event for temperatures in some interval $[T_1, T_2]$. For temperatures smaller than T_1 it may be very unlikely

that the good event happens. This may cause a superpolynomial or even exponential expected run time although the run time is polynomially bounded with overwhelming probability.

Definition 1. *Let A be a randomized search heuristic (RSH) running for a polynomial number of $p(m)$ rounds and let $s(m)$ be the success probability, i. e., the probability that A finds an optimal search point within this phase. A is called*

- *successful, if $s(m) \geq 1/q(m)$ for some polynomial $q(m)$,*
- *highly successful, if $s(m) \geq 1 - 1/q(m)$ for some polynomial $q(m)$, and*
- *successful with overwhelming probability, if $s(m) = 1 - e^{-\Omega(m^\varepsilon)}$ for some $\varepsilon > 0$.*

One can be satisfied with successful RSHs, since then multistart variants not depending on p and q are successful with overwhelming probability and have an expected polynomial run time. An RSH is called unsuccessful if, for each polynomial p, the success probability within $p(m)$ steps is $o(m^{-k})$ for each constant k. This implies a superpolynomial expected optimization time. Moreover, multistart variants do not help.

4 Simulated Annealing Beats Metropolis on Some Simple Graphs

Our plan is to present simple graphs where SA beats MA for each temperature. The graphs should allow proofs as simple as possible. The idea behind the chosen graphs is the following. The problem to compute an MST on graphs with many two-connected components is separable, i. e., an MST consists of MSTs on the two-connected components. We investigate graphs where each two-connected component can be handled easily by MA with a well-chosen temperature, but different components need different temperatures. To keep the analysis easy the components have constant size. This implies that, for high temperatures, each component can be optimized, but the solutions are not stable. They are destroyed from time to time and then reconstructed. Therefore, it is unlikely that all the components are optimized simultaneously. SA can handle these graphs efficiently.

As announced, we investigate connected triangles (CT), see Figure 1, with $m = 6n$ edges. The number of triangles equals $2n$ and the number of vertices equals $4n + 1$. The weight profile (w_1, w_2, w_3) of a triangle is simply the ordered vector of the three edge weights. We investigate CTs with n triangles with weight profile $(1, 1, m)$ and n triangles with weight profile (m^2, m^2, m^3). The unique MST consists of all edges of weight 1 or m^2.

Theorem 1. *The probability that the Metropolis algorithm applied to CTs with n triangles with weight profile $(1, 1, m)$ and n triangles with weight profile (m^2, m^2, m^3) computes the MST within e^{cm} steps (c a positive constant which is small enough) is bounded above by $e^{-\Omega(m)}$, i. e., MA is unsuccessful on these instances.*

Proof. We distinguish the cases of high temperature ($T \geq m$) and low temperature ($T < m$).

The low temperature case is easy. We do not care about the triangles with weight profile $(1, 1, m)$. For each other triangle, MA accepts the exclusion of the first flipping edge. By Chernoff bounds, with probability $1 - 2^{-\Omega(m)}$, we obtain $\Omega(m)$ triangles where the first spanning tree contains the heavy edge. In order to obtain the MST it is necessary to include the missing edge of weight m^2. If this edge is chosen to flip, the probability of selecting the new search point equals $e^{-m^2/T} \leq e^{-m}$. Hence, the success probability within $e^{m/2}$ steps is $e^{-\Omega(m)}$.

In the high temperature case, we do not care about the heavy triangles. For the light triangles, we distinguish between complete triangles (the search point chooses all three edges), optimal triangles (the two weight-1 edges are chosen), and bad triangles. The status of each triangle starts with "complete" and follows a Markov chain with the following transition probabilities:

	complete	optimal	bad
complete	$1 - 3/m$	$1/m$	$2/m$
optimal	$\frac{1}{m} \cdot e^{-m/T}$	$1 - \frac{1}{m} \cdot e^{-m/T}$	0
bad	$\frac{1}{m} \cdot e^{-1/T}$	0	$1 - \frac{1}{m} \cdot e^{-1/T}$

Let X_t be the number of optimal triangles after time step t, i. e., $X_0 = 0$. We are waiting for the first point of time t when $X_t = n$. Obviously, $|X_{t+1} - X_t| \leq 1$. Moreover,

$$\text{Prob}(X_{t+1} = a + 1 \mid X_t = a) \leq \frac{n - a}{m}$$

since it is necessary to flip the heaviest edge in one of the at most $n - a$ complete triangles, and

$$\text{Prob}(X_{t+1} = a - 1 \mid X_t = a) = \frac{a}{m} \cdot e^{-m/T} \geq \frac{a}{3m}$$

since $T \geq m$ and since it is necessary to flip the heaviest edge in one of the optimal triangles and to accept the new search point. Since we are interested in lower bounds, we use the upper bound for the probability of increasing a and the lower bound for the probability of decreasing a. Ignoring steps not changing a, we obtain the following transition probabilities for the new Markov chain Y_t:

$$\text{Prob}(Y_{t+1} = a - 1 | Y_t = a) = \frac{a/(3m)}{a/(3m) + (n - a)/m} = \frac{a}{3n - 2a}.$$

There has to be a phase where the Y-value increases from $(10/11)n$ to n without reaching $(9/11)n$. In such a phase the probability of decreasing steps is bounded below by $\frac{(9/11)n}{3n-(18/11)n} = \frac{3}{5}$. Applying results on the gambler's ruin problem, the

probability that one phase starting at $a = (10/11)n$ and finishing at $a = (9/11)n$ or $a = n$ stops at $a = n$ is bounded above by

$$((3/2)^{n/11} - 1)/((3/2)^{2n/11} - 1) = e^{-\Omega(m)}$$

since the probability of decreasing steps is at least by a factor of 3/2 larger than the probability of increasing steps. Hence, the probability of finding the MST within e^{cm} steps, $c > 0$ small enough, is bounded by $e^{-\Omega(m)}$. □

Theorem 2. *Let p be a polynomial and let the cooling schedule be described by $T(1) = m^3$ and $\alpha = 1 - 1/(cm)$ for some constant $c > 0$. If c is large enough, the probability that simulated annealing applied to CTs with n $(1, 1, m)$-triangles and n (m^2, m^2, m^3)-triangles computes the MST within $3cm \ln m$ steps is bounded below by $1 - 1/p(m)$.*

Proof. We only investigate the search until the temperature drops below 1. This phase has a length of at most $3cm \ln m$ steps and contains two subphases where the temperature is in the interval $[m^2, m^{5/2}]$ or in the interval $[1, m^{1/2}]$. The length of each subphase is at least $(c/4)m \ln m$.

If $T \leq m^{5/2}$, the probability of including an edge of weight m^3 is bounded above by $e^{-m^{1/2}}$. Each run where such an event happens is considered as unsuccessful. If $T \in [m^2, m^{5/2}]$ and an (m^2, m^2, m^3)-triangle is optimal, this triangle remains optimal unless the event considered above happens. Applying Chernoff bounds to each edge and choosing c large enough, the probability of not flipping edges of each triangle at least $c'' \log m$ times is bounded by m^{-k}, $c'' > 0$ and k arbitrary constants. This is a second source of bad behavior. Now, we investigate one (m^2, m^2, m^3)-triangle and the steps flipping one of its edges. For each complete or bad triangle, there is a chance that it turns into optimal within the next two steps concerning this triangle. This happens if the right two edges flip in the right order (probability 1/9) and the inclusion of the edge with weight m^2 is accepted (probability $e^{-m^2/T} \geq e^{-1}$). The probability of not having a good pair among the at least $(c''/2) \log m$ step pairs, can be made much smaller than m^{-k} by choosing c'' large enough. Altogether, the probability that the first subphase does not finish with MSTs on all (m^2, m^2, m^3)-triangles can be made smaller than $1/(3p(m))$.

The same calculations for $T \in [1, m^{1/2}]$ and the $(1, 1, m)$-triangles show that the probability that the second subphase does not finish with MSTs on all $(1, 1, m)$-triangles can be made smaller than $1/(3p(m))$. Finally, the probability that an (m^2, m^2, m^3)-triangle has turned from optimal into non-optimal after the first subphase is smaller than $1/(3p(m))$. This proves the theorem. □

We have proved that SA is highly successful for the considered graph instances. It is easy to choose a cooling schedule such that SA is even successful with overwhelming probability, e. g., $T(1) = m^3$ and $\alpha = 1 - 1/m^2$. See Neumann and Wegener (2004) to compare SA with simple evolutionary algorithms.

This section contains the result announced in the title of the paper. In the remaining sections, we investigate which graphs can be handled efficiently by MA and SA, only by SA, or by none of them.

5 Connected Triangles with the Same Weight Profile

It is interesting to understand how much different weights have to differ such that MA or SA are able to construct efficiently an MST. For this reason, we investigate graphs consisting of connected triangles in more detail. In this section, we consider the case of n CTs with the same weight profile $(w, w, (1+\varepsilon(m)) \cdot w)$ where $\varepsilon(m) > 0$. We distinguish the cases where $\varepsilon(m)$ is bounded below by a positive constant ε and the case where $\varepsilon(m) = o(1)$.

Theorem 3. *If $\varepsilon(m) \geq \varepsilon > 0$, MA with an appropriate temperature finds the MST on CTs with n $(w, w, (1+\varepsilon(m)) \cdot w)$-triangles in expected polynomial time and is successful with overwhelming probability.*

Proof. A good temperature has to fulfil two properties:

- It has to be low enough to distinguish w-edges effectively from $(1+\varepsilon) \cdot w$-edges.
- It has to be high enough to allow the inclusion of a w-edge in expected polynomial time.

We choose $\gamma := 3/\varepsilon$ and $T := w/(\gamma \cdot \ln m)$. The probability to accept the inclusion of a w-edge equals $e^{-w/T} = m^{-\gamma}$ while the corresponding probability for a $((1 + \varepsilon(m)) \cdot w)$-edge equals $m^{-\gamma \cdot (1+\varepsilon(m))} \leq m^{-\gamma-3}$. We analyze the success probability of a phase of length $m^{\gamma+2}$ starting with an arbitrary connected graph. The event to accept the inclusion of a heavy edge is considered as an unsuccessful phase. The probability of this event is bounded above by $1/m$. Following the lines of the proof of Theorem 2 we have for each triangle with overwhelming probability $\Omega(m^{\gamma+1})$ steps flipping an edge of this triangle which we partition into $\Omega(m^{\gamma+1})$ pairs of consecutive steps. The probability that a complete or bad triangle is turned within such two steps into an optimal one is $\Omega(m^{-\gamma})$. Hence, with overwhelming probability, all triangles turn into optimal during this phase and with probability at least $1 - 1/m$ none of them is turned into non-optimal. Hence, the expected number of phases is $O(1)$ and the probability that a sequence of m phases is unsuccessful is exponentially small. □

It is obvious how to tune the parameters in order to get improved run times. We omit such calculations which do not need new ideas. SA finds the MST in polynomial time with a probability exponentially close to 1 if it starts with $T(1) := w/(\gamma \cdot \ln m)$ and has a cooling schedule that cools down the temperature sufficiently slow. This follows in the same way as Theorem 3.

Theorem 4. *If $\varepsilon(m) = o(1)$, MA and SA are unsuccessful on CTs with n $(w, w, (1+\varepsilon(m)) \cdot w)$-triangles.*

Proof. First, we investigate MA. The search starts with n complete triangles and each one has a probability of 2/3 to be turned into a bad one before it is turned into an optimal one. With overwhelming probability, at least $n/2$ bad triangles are created where the missing w-edge has to be included in order to be able to turn it into an optimal triangle. The probability of including a w-edge within a polynomial number of $p(m)$ steps is bounded above by $p(m) \cdot e^{-w/T}$. This is bounded below by $\Omega(m^{-k})$ only if $e^{-w/T} = \Omega(m^{-\gamma})$ for some constant $\gamma > 0$. Hence, we can assume that $T \geq w/(\gamma \cdot \ln m)$ for some constant $\gamma > 0$.

Let $p^*(T)$ be the probability of accepting the inclusion of a w-edge and $p^{**}(T)$ the corresponding probability for a $((1+\varepsilon(m)) \cdot w)$-edge. Since $T \geq w/(\gamma \cdot \ln m)$ and $\varepsilon(m) = o(1)$,

$$\begin{aligned} p^*(T)/p^{**}(T) &= e^{-w/T} \cdot e^{(1+\varepsilon(m)) \cdot w/T} \\ &= e^{\varepsilon(m) \cdot w/T} \\ &\leq e^{\varepsilon(m) \cdot \gamma \cdot \ln m} \\ &= m^{\varepsilon(m) \cdot \gamma}. \end{aligned}$$

Choosing m large enough, this gets smaller than any $m^\delta, \delta > 0$. It will turn out that this advantage of w-edges against $((1 + \varepsilon(m)) \cdot w)$-edges is too small. The stochastic process behind MA can be described by the parameters b (number of bad triangles) and c (number of complete triangles). We use the potential function $2b + c$ which starts with the value n and has the value 0 for the MST. The value of the potential function changes in the following way:

- It increases by 1 if a complete triangle turns into a bad one or an optimal one turns into a complete one. The probability of the first event equals $2c/m$, since we have to flip one of the two light edges of one of the complete triangles. The probability of the second event equals $p^{**}(T) \cdot (n - b - c)/m$ since we have to flip the heavy edge in one of the $n - b - c$ optimal triangles and to accept this flip.
- It decreases by 1 if a complete triangle turns into an optimal one (probability c/m) or a bad triangle turns into a complete one (probability $p^*(T) \cdot b/m$).
- It remains unchanged, otherwise.

Since we are interested in lower bounds on the optimization time, we can ignore all non-accepted steps, i. e., all steps not changing the potential. If $b \leq n^{1/2}$ and m is large enough, the probability that an accepted step increases the potential is at least 3/5. This claim is equivalent to

$$\frac{2c/m + p^{**}(T) \cdot (n - b - c)/m}{2c/m + p^{**}(T) \cdot (n - b - c)/m + c/m + p^*(T) \cdot b/m} \geq \frac{3}{5}$$

which is equivalent to

$$2c + p^{**}(T) \cdot (n - b - c) \geq \frac{9}{5}c + \frac{3}{5} \cdot p^{**}(T) \cdot (n - b - c) + \frac{3}{5}p^*(T) \cdot b$$

and

$$\frac{1}{5}c + \frac{2}{5}p^{**}(T) \cdot (n - b - c) \geq \frac{3}{5} \cdot p^*(T) \cdot b.$$

This is obviously true if $c \geq 3 \cdot b$. Otherwise, $n - b - c \geq n - 4b \geq n - 4n^{1/2}$ and it is sufficient to show that

$$2 \cdot p^{**}(T) \cdot (n - 4n^{1/2}) \geq 3 \cdot p^{*}(T) \cdot n^{1/2}$$

or

$$p^{*}(T)/p^{**}(T) \leq \frac{2}{3}(n^{1/2} - 4).$$

We have shown that this holds for large enough m, since $n = \Omega(m)$. The claim for MA follows now from results on the gambler's ruin problem. The probability to start with a potential of $n^{1/2}/2$ and to reach the value 0 before the value $n^{1/2}$ is exponentially small. Finally, we investigate a polynomial number of $p(m)$ steps of SA. Let d be chosen such that $p(m) \leq m^d$. We claim that it is unlikely that the potential drops below $n^{1/2}/4$ within m^d steps. With overwhelming probability, we produce a bad triangle. Therefore, it is necessary to accept the inclusion of a w-edge. Hence, as seen above, only steps where the temperature is at least $w/(\gamma \cdot \ln m)$ for some appropriate constant $\gamma > 0$ have to be considered. However, the analysis of MA treats all these temperatures in the same way. The probability to start with a potential of $n^{1/2}/2$ and to reach the value $n^{1/2}/4$ before $(3/4)n^{1/2}$ is still exponentially small. □

The proof also shows that SA with an arbitrary cooling schedule is unsuccessful in the considered situation.

6 Simulated Annealing Is Successful for $(1+\varepsilon)$-Separated Weights

We have seen in Theorem 4 that MA and even SA are unsuccessful on certain graphs if we allow that different weights may differ by a factor of $1+\varepsilon(m)$ where $\varepsilon(m)$ is an arbitrary function such that $\varepsilon(m) = o(1)$. Here, we prove that SA is highly successful on all graphs if the different weights differ at least by a factor of $1 + \varepsilon$ for some constant $\varepsilon > 0$.

Before proving this result, we repeat some well-known facts about MSTs. Let $E_1, \ldots, E_r$ be the partition of the edge set E such that all edges in E_i have the same weight W_i and $W_1 > \cdots > W_r$. Let $c_i, 1 \leq i \leq r+1$, be the number of connected components of $G_i := (V, E_i \cup \cdots \cup E_r)$. Each MST contains exactly $a_i := c_{i+1} - c_i$ E_i-edges such that the chosen edges from $E_i \cup \cdots \cup E_r$ span the connected components of G_i. A set E_i^* of a_i E_i-edges is called optimal if $G_i^* := (V, E_i^* \cup E_{i+1} \cup \cdots \cup E_r)$ has the same connected components as G_i. An MST contains exactly the edges of optimal sets $E_1^*, \ldots, E_r^*$. The set E_i^* is not necessarily uniquely defined. The idea of the proof is the following. There is some point of time $t_i, 1 \leq i \leq r+1$, such that, with large probability, the following holds. After step t_i, no inclusion of an edge from $E_1 \cup \cdots \cup E_i$ is accepted and at step t_i the current search point has chosen among all E_j-edges, $1 \leq j \leq i-1$, an optimal subset E_j^*. This implies that after step t_i no edges from $E_1 \cup \cdots \cup E_{i-1}$ are included (the first property) or excluded (this would destroy the connectedness

of the graph described by the search point). Moreover, no edges from E_i are included and we hope to exclude enough E_i-edges until step t_{i+1} such that then the search point chooses an optimal set E_i^* of E_i-edges. Note that after time step t_i the set of chosen E_i-edges is always a superset of an optimal set E_i^* since, otherwise, the considered graph would be unconnected. Finally, the properties imply that at step t_{r+1} the search point describes an MST.

Theorem 5. *Let the weights of the edges be bounded by 2^m and $(1+\varepsilon)$-separated for some constant $\varepsilon > 0$, i. e., $w_i > w_j$ implies $w_i \geq (1+\varepsilon) \cdot w_j$. SA with an appropriate cooling schedule is highly successful when searching for an MST on such graphs.*

Proof. Let $T(1) := 2^m$, $\gamma := 8/\varepsilon$, α be the cooling factor such that it takes $m^{\gamma+7}$ steps to decrease the temperature from T to $T/(1+\varepsilon/2)$, and β be defined by $(1+\varepsilon/2)^\beta = 2$. Then we set $t_{r+1} := 2\beta m^{\gamma+8}$. Until step t_{r+1}, the temperature has dropped (far) below $1/m$. Our claim is that, with a probability of $1-O(1/m)$, the search point at step t_{r+1} describes an MST.

To follow the proof strategy discussed above let $t_i, 1 \leq i \leq r$, be the earliest point of time when $T(t_i) \leq W_i/((1+\varepsilon) \cdot \gamma \cdot \ln m)$. The probability of accepting the inclusion of an edge of weight W_i after step t_i is bounded above by $m^{-\gamma-8}$. During the next $m^{\gamma+7}$ steps, with overwhelming probability, there are $O(m^{\gamma+6})$ steps flipping a specified edge and the probability to accept this edge at least once is $O(1/m^2)$. Afterwards, the temperature has dropped by a factor of $1/(1+\varepsilon/2)$. The probability to accept this edge is then bounded by $m^{-\gamma-12}$ and the probability to accept the edge during the next $m^{\gamma+7}$ steps is $O(1/m^5)$. This argumentation can be continued implying that the probability to accept the inclusion of the considered edge after step t_i is $O(1/m^2)$. Hence, with probability $1 - O(1/m)$, it holds that, for all i, edges of weight W_i are not included after step t_i. In the following, we assume that this event holds.

We assume that at step t_i the search point chooses optimal sets $E_1^*, \ldots, E_{i-1}^*$ and a superset E_i' of an optimal set E_i^*. This is obviously true for $i = 1$. We analyze the steps $t_i, \ldots, t_i + m^{\gamma+7} - 1$. The probability to accept an edge with weight $w \leq W_{i+1}$ in one step is bounded below by $m^{-\gamma-4}$ during this phase. By our assumption, we do not include edges of weight $w \geq W_i$. Let $b_i := |E_i'| - |E_i^*|$ at step t_i. As long as $|E_i'| > |E_i^*|$ there are at least $|E_i'| - |E_i^*|$ candidate E_i-edges whose exclusion is possible. The exclusion of such an edge is only accepted if this edge lies on a cycle. Either the edge lies on a cycle or there is a missing edge of weight $w \leq W_{i+1}$ whose inclusion creates a cycle containing the considered E_i-edge. If no cycle with an E_i-edge exists, the probability of creating such a cycle in the next step is at least $m^{-\gamma-5}$. If a cycle with an E_i-edge exists, the probability to destroy the cycle by excluding an E_i-edge is at least $1/m$ (there may be more than one E_i-edge on the cycle). Let us assume that we do not exclude b_i E_i-edges within the considered $m^{\gamma+7}$ steps. Let s be the number of steps in this phase where a cycle with an E_i-edge exists.

If $s \geq m^{3/2}$, then the probability of less than $b_i \leq m$ steps excluding an E_i-edge on the cycle is exponentially small. If $s < m^{3/2}$, then the probability that among the at least $m^{\gamma+7} - m^{3/2}$ steps without a cycle with E_i-edges there are less than $m^{3/2}$ steps creating such a cycle is exponentially small. Hence, with overwhelming probability, enough E_i-edges are excluded and the claim holds for step t_{i+1}.

Altogether, with a probability of $1 - O(1/m)$, SA has found an MST after $O(m^{\gamma+8})$ steps. □

It is easy to see that we can generalize the result to weights up to $2^{p(m)}$ for a polynomial p. The run time increases by a factor of $O(p(m)/m)$. It is possible to tune the parameters to obtain better run times. However, the purpose of Theorem 4 and Theorem 5 was to identify the border (with respect to quotients of different weights) between cases where SA is highly successful and cases where SA can be unsuccessful. With respect to these aims we have obtained optimal results. It is easy to generalize our results to prove that SA is always highly successful if one is interested in $(1 + \varepsilon)$-optimal spanning trees. It remains an open problem to find other sufficient conditions implying that MA or SA is successful or unsuccessful on the MSTP.

7 Conclusions

The paper contributes to the theory of randomized search heuristics, in particular, the Metropolis algorithm and simulated annealing. The problem to present a natural example from combinatorial optimization where simulated annealing beats the Metropolis algorithm is solved by investigating the problem of computing minimum spanning trees. Moreover, the minimal factor between different weights to guarantee that simulated annealing finds minimum spanning trees efficiently is determined.

References

1. Droste, S., Jansen, T., and Wegener, I. (2001). Dynamic parameter control in simple evolutionary algorithms. FOGA'2000. Foundations of Genetic Algorithms 6 (Eds. Martin, W. N. and Spears, W. M.), 275–294. Morgan Kaufmann.
2. Giel, O. and Wegener, I. (2003). Evolutionary algorithms and the maximum matching problem. Proc. of 20th Symp. on Theoretical Aspects of Computer Science (STACS), LNCS 2607, 415–426.
3. Jerrum, M. (1992). Large cliques elude the Metropolis process. Random Structures and Algorithms 3, 347–359.
4. Jerrum, M. and Sinclair, A. (1996). The Markov chain Monte Carlo method. An approach to approximate counting and integration. Ch. 12 of Hochbaum, D. (Ed.). *Approximation Algorithms for NP-hard Problems*, 482–522. PWS Publishing Company.
5. Jerrum, M. and Sorkin, G. B. (1993). Simulated annealing for graph bisection. Proc. of 37th Symp. Foundations of Computer Science (FOCS), 94–103.

6. Jerrum, M. and Sorkin, G. B. (1998). The Metropolis algorithm for graph bisection. Discrete Applied Mathematics 82, 155–175.
7. Neumann, F. and Wegener, I. (2004). Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. Proc. of Genetic and Evolutionary Computation. GECCO 2004. LNCS 3102, 713–724.
8. Sasaki, G. (1991). The effect of the density of states on the Metropolis algorithm. Information Processing Letters 37, 159–163.
9. Sasaki, G. and Hajek, B. (1988). The time complexity of maximum matching by simulated annealing. Journal of the ACM 35, 387–403.
10. Sorkin, G. B. (1991). Efficient simulated annealing on fractal energy landscapes. Algorithmica 6, 367–418.