
Computational Optimization and Applications, 29, 369–385, 2004

Computing the Initial Temperature of Simulated Annealing

WALID BEN-AMEUR walid.benameur@int-evry.fr
GET/INT—CNRS/SAMOVAR, Institut National des Télécommunications, 9, rue Charles Fourier, 91011 Evry, France

Received May 6, 2003; Revised December 30, 2003

Abstract. The classical version of simulated annealing is based on a cooling schedule. Generally, the initial temperature is set such that the acceptance ratio of bad moves is equal to a certain value χ_0. In this paper, we first propose a simple algorithm to compute a temperature which is compatible with a given acceptance ratio. Then, we study the properties of the acceptance probability. It is shown that this function is convex for low temperatures and concave for high temperatures. We also provide a lower bound for the number of plateaux of a simulated annealing based on a geometric cooling schedule. Finally, many numerical experiments are reported.

Keywords: simulated annealing, initial temperature, acceptance ratio

Introduction

Simulated annealing is a general probabilistic local search algorithm, proposed 20 years ago by Cerny [3] and Kirkpatrick et al. [10] to solve difficult optimization problems. Many large instances of practical difficult problems were successfully solved by simulated annealing (see, e.g., [2, 7–9]).

To use a simulated annealing algorithm, one has first to define a set of solutions, generally large, representing the solutions of an optimization problem. Then a neighborhood structure is defined. To find a good solution we move from a solution to one of its neighbors in accordance to a probabilistic criterion. If the cost decreases then the solution is changed and the move is accepted. Otherwise, the move is accepted only with a probability depending on the cost increase and a control parameter called temperature. Classically, the probability to accept bad moves, i.e. moves with increase in terms of cost, is high at the beginning to allow the algorithm to escape from local minimum. This probability decreases in a progressive way by reducing the temperature. The method used to decrease the temperature is generally called cooling schedule. The performance of the algorithm strongly depends on the choice of the cooling schedule and the neighborhood structure.

Many theoretical papers focused on an optimal cooling schedule (see, e.g., [1, 4, 6, 12, 13]). One of the most important results may be the proof of optimality of a logarithmic cooling schedule given in Hajek [6]. However, the number of iterations needed to guarantee to find of a global optimum is generally very large (see, e.g., [1]). The transition probability P_{ij} from state i to state j is defined as the product of a generation probability G_{ij} and an acceptance probability A_{ij}.

The acceptance probability considered in this paper is the one defined by Metropolis [11]:

$$A_{ij} = \exp\left(-\frac{E_j - E_i}{T}\right) \quad \text{if } E_j > E_i \quad \text{and} \quad A_{ij} = 1 \quad \text{otherwise} \tag{1}$$

where T is the current temperature and E_i (resp. E_j) is the energy of state i (resp. j).

A state is a solution of an optimization problem and energy is the cost function that has to be minimized. We indifferently use energy and cost to designate the same thing.

We also assume that the homogenous Markov chain representing the simulated annealing at a given temperature T is irreducible (i.e. all states can be reached from any other state with a positive probability) and aperiodic (see, e.g., [1]). These conditions are generally satisfied.

If we assume that the generation probabilities are symmetrical ($G_{ij} = G_{ji}$), the stationary distribution is nothing other than the Boltzmann distribution: $\pi_i = \frac{\exp(-\frac{E_i}{T})}{\sum_j \exp(-\frac{E_j}{T})}$.

Another generation strategy that is commonly used is given by

$$G_{ij} = \begin{cases} \dfrac{1}{|N(i)|} & \text{if } j \in N(i) \\ 0 & \text{else} \end{cases} \tag{2}$$

where $N(i)$ is the set of neighbors of i. The stationary distribution is then given by

$$\pi_i = \frac{|N(i)| \exp\left(-\frac{E_i}{T}\right)}{\sum_j |N(j)| \exp\left(-\frac{E_j}{T}\right)} \tag{3}$$

As previously said, one of the most important properties of simulated annealing is its hill climbing feature. This is achieved by accepting some increasing cost moves. Consequently, the average probability of accepting these moves is very important to evaluate the ability of simulated annealing to escape from local minimum.

This acceptation ratio strongly depends on the temperature. To allow the simulated annealing to find good solutions, one has to carefully compute the initial temperature. This parameter plays an important role in simulated annealing, but is of course only a piece of a large puzzle. This paper will focus on this initial temperature and some other properties of the acceptance ratio.

Many methods have been proposed in literature to compute the initial temperature T_0. It is suggested in Kirkpatrick et al. [10] to take $T_0 = \Delta E_{\max}$ where $\Delta E_{\max}$ is the maximal cost difference between any two neighboring solutions.

Another scheme based on a more precise estimation of the cost distribution is proposed with multiple variants (see, e.g., [1, 16]). It is recommended to choose $T_0 = K\sigma_\infty^2$ where K is a constant typically ranging from 5 to 10 and σ_∞^2 is the second moment of the energy distribution when the temperature is ∞. σ_∞ is estimated using a random generation of some solutions.

A more classical and intuitive method is described in Kirkpatrick et al. [10]. It consists in computing a temperature such that the acceptance ratio is approximately equal to a given value χ_0. First, we choose a large initial temperature. Then, we have to perform a number of transitions using this temperature. The ratio of accepted transitions is compared with χ_0. If it is less than χ_0, then the temperature is multiplied by 2. The procedure continues until the observed acceptance ratio exceeds χ_0. Other variants are proposed to obtain an acceptance ratio which is close to χ_0. It is, for example, possible to divide the temperature by 3 if the acceptance ratio is much higher than χ_0. Using this kind of rules, cycles are avoided and a good estimation of the temperature can be found.

Another procedure is proposed in Johnson et al. [7, 8]. Temperature is obtained using the formula $T_0 = -\frac{\overline{\Delta E}}{\ln(\chi_0)}$, where $\overline{\Delta E}$ is an estimation of the cost increase of strictly positive transitions. This estimation is again obtained by randomly generating some transitions. Notice that $-\frac{\delta_t}{\ln(\chi_0)}$, where δ_t is the cost increase induced by a transition t, is the temperature allowing this transition to be accepted with a probability χ_0. In other terms, $T_0 = -\frac{\overline{\Delta E}}{\ln(\chi_0)}$ is the average of these temperatures over a set of random transitions.

Finally, note that to accelerate the simulated annealing, a heuristic is sometimes used to find a good initial solution. Then, simulated annealing is applied with a low initial temperature (see, e.g., [5, 7, 15]). An algorithm is provided by Varanelli [15] to compute an initial temperature such that the expected cost of the best solution that can be found at this temperature is approximately equal to the cost of the solution given by the heuristic.

A new algorithm to compute the initial temperature is given in this paper. The algorithm is fast and accurate. It is presented in next section. The convergence is proved in Section 1. Some other properties of the acceptance probability are presented in Section 2. Many numerical experiments are reported and commented in Section 3. Finally, some concluding remarks are given in Section 4.

1. An efficient algorithm to compute the temperature

The initial temperature is often chosen such that the acceptance probability is approximately equal to a certain value, for example, 0.8 (see, e.g., [1]). Let t be a strictly positive transition and let $\max_t$ (resp. $\min_t$) be the state after (resp. before) the transition. As we assumed that the transition is strictly positive, then $E_{\max_t} > E_{\min_t}$. To simplify notation, we use δ_t to designate the cost difference $E_{\max_t} - E_{\min_t}$. Using the generation strategy (2), the acceptance probability is given by:

$$\chi(T) = \frac{\sum_{t \text{ positive}} \pi_{\min_t} \frac{1}{|N(\min_t)|} \exp\left(-\frac{\delta_t}{T}\right)}{\sum_{t \text{ positive}} \pi_{\min_t} \frac{1}{|N(\min_t)|}}. \tag{4}$$

Note that $\pi_{\min_t} \frac{1}{|N(\min_t)|}$ represents the probability to generate a transition t when the energy states are distributed in conformance with the stationary distribution (3). Moreover, $\exp(-\frac{\delta_t}{T})$ is the probability to accept a positive transition t. Thus, $\chi(T)$ is the conditional expectation of the acceptance of positive transitions.

We will use an estimation $\hat{\chi}(T)$ of this acceptance probability based on a random set S of positive transitions. $\hat{\chi}(T)$ is defined as follows:

$$\hat{\chi}(T) = \frac{\sum_{t \in S} \pi_{\min_t} \frac{1}{|N(\min_t)|} \exp\left(-\frac{\delta_t}{T}\right)}{\sum_{t \in S} \pi_{\min_t} \frac{1}{|N(\min_t)|}} = \frac{\sum_{t \in S} \exp\left(-\frac{E_{\max_t}}{T}\right)}{\sum_{t \in S} \exp\left(-\frac{E_{\min_t}}{T}\right)}. \quad (5)$$

Now, let us assume that we are looking for a temperature T_0 such that $\chi(T_0) = \chi_0$ where $\chi_0 \in]0, 1[$ is the wanted acceptance probability. We will propose a simple iterative method to compute such a temperature. In fact, we will consider $\hat{\chi}(T)$ instead of $\chi(T)$. First, we randomly generate a set of positive transitions S. This can be done, for example, by generating some states and a neighbor for each state. The energies $E_{\max_t}$ and $E_{\min_t}$ corresponding with the states of the subset S are stored. Then we choose a value T_1 for temperature. T_1 can be any positive number.

T_1 may be far from T_0. To find T_0 we use the recursive formula

$$T_{n+1} = T_n \left(\frac{\ln(\hat{\chi}(T_n))}{\ln(\chi_0)} \right)^{\frac{1}{p}}. \quad (6)$$

where p is a real number ≥ 1.

When $\hat{\chi}(T_n)$ becomes close to χ_0 we can stop: T_n is a good approximation of the wanted temperature T_0.

Please note that we use at each iteration the energy values previously stored. In other words, we do not have to generate new transitions.

Before proving the convergence of our procedure, let us give a summary of the whole process. ϵ denotes a small real number (e.g., 10^{-3}).

Computing the temperature of simulated annealing

Step 1.

(a) Estimate the number of samples $\|S\|$ needed to compute $\hat{\chi}(T)$.
(b) Generate and store $\|S\|$ random positive transitions.
(c) Set T_1 at any strictly positive number and set $n = 1$.

Step 2.

(a) Compute $\hat{\chi}(T_n) = \frac{\sum_{t \in S} \exp(-\frac{E_{\max_t}}{T_n})}{\sum_{t \in S} \exp(-\frac{E_{\min_t}}{T_n})}$.
(b) If $|\hat{\chi}(T_n) - \chi_0| \leq \epsilon$, return T_n.

Otherwise

– $T_{n+1} = T_n (\frac{\ln(\hat{\chi}(T_n))}{\ln(\chi_0)})^{\frac{1}{p}}$.
– $n = n + 1$.
– go to Step 2(a).

End.

Steps 1(a) and (b) will be discussed later.

As said before, the value of T_1 can be any strictly positive number. However, to slightly accelerate the whole process, we compute T_1 using the formula given in introduction Johnson [7, 8]:

$$T_1 = -\frac{\sum_{t \in S} \delta_t}{\|S\| \ln(\chi_0)}. \tag{7}$$

In the rest of this section, we first prove under some assumptions the convergence of the algorithm described above. Then we give some remarks about the sampling procedure needed by the algorithm.

1.1. Algorithm convergence

To show the convergence of the algorithm, we will prove that $T \to T(\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ is a non decreasing function and $T \to \hat{\chi}(T)$ is a strictly increasing function. This means that T_0 is a unique fixed point of function $T \to T(\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ and $\min(T_0, T_n) \leq T_{n+1} \leq \max(T_0, T_n)$. Notice that if $T \to T(\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ is a non decreasing function when $p = 1$, then it will have the same behavior for any $p \geq 1$. This can be seen by computing the derivative of the logarithm of this function: $\frac{1}{T} + \frac{1}{p}\frac{\hat{\chi}'(T)}{\hat{\chi}(T)\ln(\hat{\chi}(T))}$. If we assume that $\hat{\chi}'(T) \geq 0$, then $\frac{1}{T} + \frac{1}{p}\frac{\hat{\chi}'(T)}{\hat{\chi}(T)\ln(\hat{\chi}(T))}$ clearly increases when p increases. Therefore, we will focus on $p = 1$.

Before giving the proofs of the wanted results, we will present an hypothesis that will be used to simplify calculation.

Hypothesis 1.1. We assume that the energy levels $E_{\min_t}$ and the cost differences δ_t of the set of transitions S are independent.

More precisely, given a temperature T, we assume that the positive transitions are generated in conformance with the equilibrium distribution. As we focus here on S, we consider the conditional distribution where the probability to generate a transition t_0 is given by $\frac{\pi_{\min_{t_0}} \frac{1}{|N(\min_{t_0})|}}{\sum_{t \in S} \pi_{\min_t} \frac{1}{|N(\min_t)|}}$. It is natural to assume that there is no correlation between $\{\delta_i, E_{\min_i}\}$ and $\{\delta_j, E_{\min_j}\}$ where i and j are two transitions of S obtained by independent trials in conformance with the conditional equilibrium distribution. However, in Hypothesis 1.1 we also assume that $E_{\min_i}$ is independent with δ_i. This assumption is less easy to understand. In fact, it depends on the distribution which is related to temperature. Said another way, even if it is valid for some temperatures, it will be invalid for others. Note however that we do not need this assumption to be strictly satisfied. The convergence of the algorithm is obtained in almost all cases when $p = 1$. Moreover, it can be ensured by increasing the value of the parameter p. More details will be given in the end of this subsection.

Lemma 1.2. *Assuming hypothesis* 1.1 *is valid, then we have*

$$\frac{\sum_{i,j\in S, i<j}\exp\left(-\frac{E_{\min_i}+E_{\min_j}}{T}\right)\left(E_{\min_i}-E_{\min_j}\right)\left(\exp\left(-\frac{\delta_i}{T}\right)-\exp\left(-\frac{\delta_j}{T}\right)\right)}{\sum_{i,j\in S}\exp\left(-\frac{E_{\max_i}+E_{\min_j}}{T}\right)\delta_i}=0.$$

Proof: Let L (resp. R) be the numerator (resp. denominator) of the ratio given in the lemma. We want to show that $\frac{L}{R}=0$. In fact, L is nothing but $\frac{1}{2}\sum_{i,j\in S}\exp(-\frac{E_{\min_i}+E_{\min_j}}{T})(E_{\min_i}-E_{\min_j})(\exp(-\frac{\delta_i}{T})-\exp(-\frac{\delta_j}{T}))$.

Moreover, using formulas 3 and 2, the expectation of $(E_{\min_i}-E_{\min_j})(\exp(-\frac{\delta_i}{T})-\exp(-\frac{\delta_j}{T}))$ is given by $_{E_{\min_j}}E((E_{\min_i}-E_{\min_j})(\exp(-\frac{\delta_i}{T})-\exp(-\frac{\delta_j}{T}))\,|\,i,j\in S)=\sum_{i,j\in S}\frac{\exp(-\frac{E_{\min_i}}{T})}{T_{E_{\min_i}}}\times\frac{\exp(-\frac{E_{\min_j}}{T})}{\sum_{k\in S}\exp(-\frac{E_{\min_k}}{T})}\times((E_{\min_i}-E_{\min_j})(\exp(-\frac{\delta_i}{T})-\exp(-\frac{\delta_j}{T})))$. Note that we used here the fact that the transitions of S are independent. We obtain

$$L=\frac{1}{2}\left(\sum_{k\in S}\exp\left(-\frac{E_{\min_k}}{T}\right)\right)^2 E\left(\left(E_{\min_i}-E_{\min_j}\right)\left(\exp\left(-\frac{\delta_i}{T}\right)-\exp\left(-\frac{\delta_j}{T}\right)\right)\,\middle|\,i,j\in S\right).$$

On the other hand,

$$\begin{aligned}R&=\sum_{i,j\in S}\exp\left(-\frac{E_{\max_i}+E_{\min_j}}{T}\right)\delta_i\\&=\sum_{j\in S}\exp\left(-\frac{E_{\min_j}}{T}\right)\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)\delta_i\\&=\left(\sum_{j\in S}\exp\left(-\frac{E_{\min_j}}{T}\right)\right)^2\sum_{i\in S}\frac{\exp\left(-\frac{E_{\min_i}}{T}\right)}{\sum_{j\in S}\exp\left(-\frac{E_{\min_j}}{T}\right)}\exp\left(-\frac{\delta_i}{T}\right)\delta_i\\&=\left(\sum_{j\in S}\exp\left(-\frac{E_{\min_j}}{T}\right)\right)^2 E\left(\exp\left(-\frac{\delta_i}{T}\right)\delta_i\,\middle|\,i\in S\right)\end{aligned}$$

Combination of the previous expressions related to L and R leads to

$$\frac{L}{R}=\frac{1}{2}\frac{E\left(\left(E_{\min_i}-E_{\min_j}\right)\left(\exp\left(-\frac{\delta_i}{T}\right)-\exp\left(-\frac{\delta_j}{T}\right)\right)\,\middle|\,i,j\in S\right)}{E\left(\exp\left(-\frac{\delta_i}{T}\right)\delta_i\,\middle|\,i\in S\right)}$$

Now using Hypothesis 1.1, one can deduce that

$$E\left(\left(E_{\min_i} - E_{\min_j}\right)\left(\exp\left(-\frac{\delta_i}{T}\right) - \exp\left(-\frac{\delta_j}{T}\right)\right) \,\middle|\, i, j \in S\right)$$
$$= E\left(E_{\min_i} - E_{\min_j} \,\middle|\, i, j \in S\right) E\left(\exp\left(-\frac{\delta_i}{T}\right) - \exp\left(-\frac{\delta_j}{T}\right) \,\middle|\, i, j \in S\right)$$

Finally,

$$\frac{L}{R} = \frac{1}{2}\frac{E\left(E_{\min_i} - E_{\min_j} \,\middle|\, i, j \in S\right) E\left(\exp\left(-\frac{\delta_i}{T}\right) - \exp\left(-\frac{\delta_j}{T}\right) \,\middle|\, i, j \in S\right)}{E\left(\exp\left(-\frac{\delta_i}{T}\right)\delta_i \,\middle|\, i \in S\right)}$$
$$= 0$$

which means that $\frac{L}{R} = 0$. □

Note that it is possible to build a particular small example for which both Hypothesis 1.1 and Lemma 1.2 are not valid. However, our experimental results (Section 3) show that the algorithm works very well in practice, and the convergence is obtained in almost all cases. More details will be given in the end of this subsection.

Proposition 1.3. *Assuming Hypothesis* 1.1 *is valid, then the derivative of* $\hat{\chi}(T)$ *is given by*:

$$\hat{\chi}'(T) = \frac{1}{T^2}\frac{\sum_{i \in S} \exp\left(-\frac{E_{\max_i}}{T}\right)\delta_i}{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)}.$$

Proof: Let us calculate $\hat{\chi}'(T)$.

$$\hat{\chi}'(T)$$
$$= \frac{\left(\sum_{i \in S} E_{\max_i} \exp\left(-\frac{E_{\max_i}}{T}\right)\right)\left(\sum_{j \in S} \exp\left(-\frac{E_{\min_j}}{T}\right)\right) - \left(\sum_{i \in S} E_{\min_i} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)\left(\sum_{j \in S} \exp\left(-\frac{E_{\max_j}}{T}\right)\right)}{T^2\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}$$
$$= \frac{\sum_{i,j} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right)\left(E_{\max_i} - E_{\min_j}\right)}{T^2\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}$$
$$= \frac{\sum_{i,j} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right)\left(E_{\min_i} - E_{\min_j} + \delta_i\right)}{T^2\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}$$
$$= \frac{\sum_{i,j} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right)\left(E_{\min_i} - E_{\min_j}\right) + \sum_{i,j} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right)\delta_i}{T^2\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}$$
$$= \frac{\sum_{i<j} \exp\left(-\frac{E_{\min_i} + E_{\min_j}}{T}\right)\left(E_{\min_i} - E_{\min_j}\right)\left(\exp\left(-\frac{\delta_i}{T}\right) - \exp\left(-\frac{\delta_j}{T}\right)\right) + \sum_{i,j} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right)\delta_i}{T^2\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}$$

Using Lemma 1.2, the previous expression becomes:

$$\begin{aligned}\hat{\chi}'(T) &= \frac{\sum_{i,j} \exp\left(-\frac{E_{\max_i}+E_{\min_j}}{T}\right)\delta_i}{T^2\left(\sum_{i\in S}\exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2} \\ &= \frac{1}{T^2}\frac{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)\delta_i}{\sum_{i\in S}\exp\left(-\frac{E_{\min_i}}{T}\right)}\end{aligned}$$

□

Please note that even if Hypothesis 1.1 is not valid, we can be satisfied with a small value of the ratio $\frac{L}{R}$ of Lemma 1.2 to obtain a good approximate value of $\hat{\chi}'(T)$.

Proposition 1.3 tells us that $\hat{\chi}'(T) > 0$. To finish our proof of convergence, we have to show that $T \to T\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)}$ is a non decreasing function.

Proposition 1.4. *Assuming Hypothesis* 1.1 *is valid, then* $(T\ln(\hat{\chi}(T)))' \leq 0$.

Proof: Derivative of $T\ln(\hat{\chi}(T))$ is given by $\ln(\hat{\chi}(T)) + T\frac{\hat{\chi}'(T)}{\hat{\chi}(T)}$.

Using expression (5), one can write:

$$\begin{aligned}\frac{1}{\hat{\chi}(T)} &= \frac{\sum_{i\in S}\exp\left(-\frac{E_{\min_i}}{T}\right)}{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)} \\ \frac{1}{\hat{\chi}(T)} &= \frac{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)\exp\left(\frac{\delta_i}{T}\right)}{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)} \\ &= \sum_{i\in S}\frac{\exp\left(-\frac{E_{\max_i}}{T}\right)}{\sum_{j\in S}\exp\left(-\frac{E_{\max_j}}{T}\right)}\exp\left(\frac{\delta_i}{T}\right)\end{aligned}$$

By concavity of logarithm, one can deduce that $\ln(\frac{1}{\hat{\chi}(T)}) \geq \sum_{i\in S}\frac{\exp(-\frac{E_{\max_i}}{T})}{\sum_{j\in S}\exp(-\frac{E_{\max_j}}{T})}\frac{\delta_i}{T}$. Said another way, we have

$$\ln(\hat{\chi}(T)) < \sum_{i\in S}\frac{\exp\left(-\frac{E_{\max_i}}{T}\right)}{\sum_{j\in S}\exp\left(-\frac{E_{\max_j}}{T}\right)}\frac{-\delta_i}{T}.$$

On the other hand, using Proposition 1.3, we obtain:

$$T\frac{\hat{\chi}'(T)}{\hat{\chi}(T)} = \frac{1}{T}\frac{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)\delta_i}{\sum_{i\in S}\exp\left(-\frac{E_{\max_i}}{T}\right)}.$$

Combination of the previous two results leads to $(T\ln(\hat{\chi}(T)))' \leq 0$. □

Propositions 1.3 and 1.4 clearly imply the convergence of the algorithm: $(T_n)_{n\in\mathbb{N}}$ is monotonous and bounded.

Note that even if the results of this subsection are based on Hypothesis 1.1, they are useful in a general context. Let us give an insight into this point. First, to show that $\hat{\chi}(T)$ is an increasing function, we only need to have the ratio of Lemma 1.2 close to 0. In other terms, we do not really require Hypothesis 1.1 to be strictly satisfied. Second, we already said in the beginning of this subsection that the derivative of the logarithm of the function $T \to T(\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ increases when p increases. Said another way, if we get some convergence problems when $p = 1$ due to the inaccuracy of Hypothesis 1.1, we can sufficiently increase p to allow $T \to T(\frac{\ln(\hat{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ to be an increasing function. Moreover, our experimental results (Section 3) show that in most of cases $p = 1$ is sufficient. We needed to take $p = 2$ in about 1 run per 1000 to guarantee the convergence. However, to strictly guarantee the convergence, we can slightly modify the algorithm of Section 1. If an oscillation is detected (i.e., $(T_{n+1} - T_n)(T_n - T_{n-1}) < 0$) then we multiply p by 2 and we continue the algorithm.

1.2. On the sampling procedure

The first steps of the algorithm (1(a) and (b)) can be called the sampling procedure.

Even if the convergence of the algorithm is shown for a set S of random transitions satisfying Hypothesis 1.1 (and experimentally in Section 3), the set S must be representative to allow the algorithm to give a temperature which is close to the wanted temperature. Obviously, the exact temperature is given when S contains all positive transitions. However, it is generally not possible to consider all transitions.

We will not give a definitive description of the sampling procedure: we think that it depends on the nature and the size of the problem that we are solving.

One can, for example, begin with a small value of $\|S\|$, compute the temperature, and increase the number of transitions until the temperature becomes stable.

It is also possible to use the temperature T_1 of Eq. (7) to perform a first simulated annealing plateau. All positive transitions considered during this plateau can be stored and then used to compute a more accurate temperature using our algorithm.

Numerical experiments that will be presented in Section 3, are based, for each value of $\|S\|$, on a random generation of independent transitions. Notice that when we use the transitions encountered during a plateau, transitions may not be independent.

2. Other properties

More properties of the acceptance probability are given in this section.

Proposition 2.1. *Assuming Hypothesis* 1.1 *is valid, then* $\hat{\chi}'(T) \leq \frac{1}{eT}$.

Proof: It was shown in Proposition 1.3 that $\hat{\chi}'(T) = \frac{1}{T^2} \frac{\sum_{i \in S} \exp(-\frac{E_{\max_i}}{T}) \delta_i}{\sum_{i \in S} \exp(-\frac{E_{\min_i}}{T})}$.

It implies that $\hat{\chi}'(T) = \frac{1}{T} \frac{\sum_{i \in S} \exp(-\frac{E_{\min_i}}{T}) \exp(-\frac{\delta_i}{T}) \frac{\delta_i}{T}}{\sum_{i \in S} \exp(-\frac{E_{\min_i}}{T})}$

Moreover, the function $x \to x \exp(-x)$ is bounded by $1/e$. Using this upper bound in the previous approximation leads to the wanted result. □

An important straightforward corollary dealing with the evolution of the acceptance probability is given below.

Corollary 2.2. *Assuming Hypothesis* 1.1 *is valid, then* $\hat{\chi}(T+\Delta T)-\hat{\chi}(T) \leq \frac{1}{e}\ln(1+\frac{\Delta T}{T})$.

Proof: A simple integration of the inequality $\hat{\chi}'(T) \leq \frac{1}{eT}$ gives the wanted result. □

Using the fact that $\ln(1+x) \leq x$, one can deduce that $\hat{\chi}(T+\Delta T) - \hat{\chi}(T) \leq \frac{1}{e}\frac{\Delta T}{T}$.

Corollary 2.2 implies that even if you divide the temperature by 2, you can not expect to reduce the acceptance probability by more than $\frac{\ln(2)}{e} \approx 0,255$.

It is also possible to use the previous corollary to have an indication about the number of iterations of a classical simulated annealing with a geometric cooling schedule. Assume that the temperature is multiplied by $\alpha < 1$ at the end of each plateau. In most of cases, the initial temperature is chosen such that the acceptance probability of positive moves is equal to χ_0. The stopping criterion can also be a low acceptance probability χ_f. Using Corollary 2.2, one can easily show that the number of plateaux N is higher than $\frac{e(\chi_0-\chi_f)}{\ln(1/\alpha)}$.

Proposition 2.3. *Assuming Hypothesis* 1.1 *is valid, then the number of plateaux is higher than* $\frac{e(\chi_0-\chi_f)}{\ln(1/\alpha)}$.

Assume, for example, that $\chi_0 = 0.9$, $\chi_f = 0.05$ and $\alpha = 0.95$. The number of plateaux is then higher than 46. If $\alpha = 0.99$, we need more than 230 plateaux. More precisely, if $\alpha = 1-\epsilon$ where $\epsilon \ll 1$, then the number of plateaux is approximately higher than $\frac{e(\chi_0-\chi_f)}{\epsilon}$.

Note that one of the advantages of the upper bound $\frac{1}{eT}$ given in Proposition 2.1 is its independence with energy. However, this upper bound is bad for low temperatures. In fact, one can easily see that $\hat{\chi}'(T) \approx \frac{C}{T^2}\exp(-\frac{\Delta}{T})$ where C is a constant depending on the energies and the transitions and Δ is the difference between the smallest $E_{\max_i}$ and the smallest $E_{\min_i}$. This clearly implies that $\hat{\chi}'(T)$ is approximately equal to 0 when T is close to 0.

To finish our study of the acceptance probability, let us consider the second derivative $\hat{\chi}''(T)$.

First, another simple lemma will be stated.

Lemma 2.4. *Assuming Hypothesis* 1.1 *is valid, then*

$$\frac{\sum_{i,j\in S, i<j} \exp\left(-\frac{E_{\min_i}+E_{\min_j}}{T}\right)\left(E_{\min_i} - E_{\min_j}\right)\left(\delta_i \exp\left(-\frac{\delta_i}{T}\right) - \delta_j \exp\left(-\frac{\delta_j}{T}\right)\right)}{\sum_{i,j\in S} \exp\left(-\frac{E_{\max_i}+E_{\min_j}}{T}\right)\delta_i{}^2} = 0.$$

This lemma can be easily proved using the same kind of arguments as those given to prove the validity of Lemma 1.2.

Proposition 2.5. *Assuming Hypothesis* 1.1 *is valid, then the second derivative is given by:*

$$\hat{\chi}''(T) = \frac{1}{T^4} \frac{\sum_{i \in S} \exp\left(-\frac{E_{\max_i}}{T}\right) \delta_i (\delta_i - 2T)}{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)}.$$

Proof: A simple derivation of $T^2 \hat{\chi}'(T)$ using Proposition 1.3 gives the following:

$$\begin{aligned}
&(T^2 \hat{\chi}'(T))' \\
&= \frac{1}{T^2} \frac{\sum_{i \in S} E_{\max_i} \exp\left(-\frac{E_{\max_i}}{T}\right) \delta_i \sum_{j \in S} \exp\left(-\frac{E_{\min_j}}{T}\right) - \sum_{i \in S} E_{\min_i} \exp\left(-\frac{E_{\min_i}}{T}\right) \delta_i \sum_{j \in S} \exp\left(-\frac{E_{\max_j}}{T}\right)}{\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2} \\
&= \frac{1}{T^2} \frac{\sum_{i,j \in S, i<j} \exp\left(-\frac{E_{\min_i} + E_{\min_j}}{T}\right) \left(E_{\min_i} - E_{\min_j}\right) (\delta_i \exp\left(-\frac{\delta_i}{T}\right) - \delta_j \exp\left(-\frac{\delta_j}{T}\right))}{\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2} \\
&+ \frac{1}{T^2} \frac{\sum_{i,j \in S} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right) {\delta_i}^2}{\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2}
\end{aligned}$$

Using Lemma 2.4, we obtain:

$$\begin{aligned}
(T^2 \hat{\chi}'(T))' &= \frac{1}{T^2} \frac{\sum_{i,j \in S} \exp\left(-\frac{E_{\max_i} + E_{\min_j}}{T}\right) {\delta_i}^2}{\left(\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)\right)^2} \\
&= \frac{1}{T^2} \frac{\sum_{i \in S} \exp\left(-\frac{E_{\max_i}}{T}\right) {\delta_i}^2}{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)}
\end{aligned}$$

Using again Proposition 1.3 gives the wanted result. □

One can easily see that the expression given above is positive when T is close to 0 and negative when T is sufficiently high.

Corollary 2.6. *Assuming Hypothesis* 1.1 *is valid, the probability to accept positive transitions is convex for low temperatures and concave for high temperatures.*

Finally, we give here simple bounds for the second derivative $\hat{\chi}''(T)$.

Corollary 2.7. *Assuming Hypothesis* 1.1 *is valid, then*

$$\frac{1}{T^2}(2 - 2\sqrt{2}) \exp(\sqrt{2} - 2) \leq \hat{\chi}''(T) \leq \frac{1}{T^2}(2 + 2\sqrt{2}) \exp(-2 - \sqrt{2}).$$

Proof: Proposition 2.5 tells us that $\hat{\chi}''(T) = \frac{1}{T^4} \frac{\sum_{i \in S} \exp(-\frac{E_{\max_i}}{T}) \delta_i (\delta_i - 2T)}{\sum_{i \in S} \exp(-\frac{E_{\min_i}}{T})}$. Simple calculation leads to:

$$\begin{aligned}
\hat{\chi}''(T) &= \frac{1}{T^2} \frac{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right) \exp\left(-\frac{\delta_i}{T}\right) \frac{\delta_i}{T} \left(\frac{\delta_i}{T} - 2\right)}{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)} \\
&= \frac{1}{T^2} \frac{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right) f\left(\frac{\delta_i}{T}\right)}{\sum_{i \in S} \exp\left(-\frac{E_{\min_i}}{T}\right)} \\
&= \frac{1}{T^2} E\left(f\left(\frac{\delta_i}{T}\right) \middle| i \in S \right)
\end{aligned}$$

where f denotes the function $x \in \mathbb{R}_+ \to x(x-2)\exp(-x)$. One can easily see that the minimum of f is obtained for $x = 2 - \sqrt{2}$ and the maximum is reached for $x = 2 + \sqrt{2}$. Thus, for any $x \geq 0$ we have $(2 - 2\sqrt{2})\exp(\sqrt{2} - 2) \leq f(x) \leq (2 + 2\sqrt{2})\exp(-2 - \sqrt{2})$.

Combinations of these inequalities and the expression of $\hat{\chi}''(T)$ leads to the wanted result. □

Notice that $(2 - 2\sqrt{2})\exp(\sqrt{2} - 2) \approx -0.462$ and $(2 + 2\sqrt{2})\exp(-2 - \sqrt{2}) \approx 0.159$.

3. Numerical experiments

To illustrate the results given in the previous sections, extensive numerical experiments are carried out. Two kind of problems are considered: random problems and traveling salesman problems (TSP).

3.1. Random problems

A random problem is represented by a symmetric graph $G = (V, E)$ where V is the set of vertices corresponding with the solutions of the problem, and E is the set of edges representing the neighborhood relationship. Each solution (vertex) has a random cost. Simulated annealing is applied to find a minimum cost solution.

The graphs used to represent random problems are described using three parameters: the number of vertices $\|V\|$, the graph density $d = \frac{\|E\|}{\frac{\|V\|\|V-1\|}{2}}$ and an upper bound U for the maximum degree. Two sets of problems are considered in this section where $(\|V\|, d, U) = (5 \times 10^4, 10^{-4}, 30)$ in the first case and $(2 \times 10^6, 10^{-6}, 5)$ in the second one. Notice that these problems are small and can be solved by enumeration. However, the aim of this section is only to study the procedure proposed in this paper to compute the temperature of a simulated annealing.

Due to space limitation, we only give a summary of the results: more details will be provided on Kluwer's web site.

We consider 4 different values of the number of samples $\|S\|$: 20, 100, 500 and 2500. We also try 8 values of the acceptance probability χ_0: 0.99, 0.9, 0.7, 0.5, 0.3, 0.1, 0.05 and 0.01. Positive transitions are randomly and independently generated. For each value of χ_0 and $\|S\|$ the algorithm of Section 1 is used to give a temperature. Convergence was always obtained with $p = 1$ (Formula 6).

Simulated annealing is then applied using the given temperature, without any decrease, to provide the experimental acceptance probability $\bar{\chi}$. We also apply simulated annealing using the temperature T_1 defined by Eq. (7) to obtain $\overline{\chi(T_1)}$ defined as the experimental acceptance probability corresponding with T_1. Recall that this temperature is commonly used by simulated annealing practitioners.

All experiments are repeated 200 times (200 runs for each value of χ_0 and $\|S\|$). Results are expressed in terms of average and standard deviation.

The ratios corresponding with Lemmas 1.2 and 2.4 are also considered here in order to check the validity of our hypothesis. We also focus on the number of iterations of the algorithm needed to compute the temperature. This number is null if the temperature T_1 given by Eq. (7) obtained in Step 1(c) is the final result of the algorithm. The precision term ϵ used in the algorithm is here equal to 10^{-3}.

To summarize, we focus on the average and the standard deviation of the following quantities: the experimental acceptance probability $\bar{\chi}$, the experimental acceptance probability obtained with T_1 of Eq. (7), the ratios corresponding with Lemmas 1.2 and 2.4 and the number of iterations of the algorithm of Section 1.

First, the algorithm used to compute the temperature converges at each run. The average and the standard deviation values corresponding with Lemmas 1.2 and 2.4 are generally low. They are not null because Hypothesis 1.1 is not always valid. When both χ_0 and $\|S\|$ are very low, Lemma 1.2 does not seem to be satisfied. In fact, the denominator of the fraction defined in Lemma 1.2 is close to 0 when T is very low. Moreover, Hypothesis 1.1 is unlikely to be satisfied when $|S|$ is very small. Nevertheless, as previously said, convergence is always obtained, even if Hypothesis 1.1 is not satisfied.

The number of iterations needed by the algorithm to achieve convergence is small for high values of χ_0. In fact, when χ_0 is high, temperature T_1 seems to be a good one. This is shown by $\overline{\chi(T_1)}$ which is very close to χ_0 when χ_0 is high.

$\overline{\chi(T_1)}$ becomes far from χ_0 when χ_0 is low. We also observed that the standard deviation of $\overline{\chi(T_1)}$ generally decreases when $\|S\|$ increases. However, the average value of $\overline{\chi(T_1)}$ can be considered as stable. This is due to the fact that T_1 is based on the average of the cost variations. Although, $\bar{\chi}$ is generally close to χ_0 more than $\overline{\chi(T_1)}$.

Moreover, our numerical experiments show that for high values of χ_0, a small value of $\|S\|$ can be sufficient to obtain a temperature achieving the goal. However, if χ_0 is low, we need higher values of $\|S\|$.

The difference between χ_0 and $\bar{\chi}$ decreases when the problem size decreases. The problem size considered in the first case is smaller than the second problem size and the results are slightly better: the standard deviation of $\bar{\chi}$ is lower in the first case than in the second one. Said another way, when the problem size is larger, we may need more transitions to compute a temperature.

Table 1. $T_1 = 500$ and $p = 1$.

χ	0.9352	0.5114	0.5007	0.5000
T	500.0	48.3411	46.7738	46.6768

Table 2. $T_1 = 5$ and $p = 1$.

χ	0.0461	0.2661	0.4687	0.4978	0.4999	0.5000
T	5.0	22.1953	42.3929	46.3526	46.6492	46.6687

Table 3. $T_1 = 500$ and $p = 2$.

χ	0.9352	0.8067	0.6818	0.5993	0.5536	0.5286	0.5152	0.5081
	0.5043	0.5023	0.5012	0.5006	0.5003	0.5002	0.5000	
T	500.0	155.4690	86.5412	64.3252	55.2274	51.0096	48.9224	47.8533
	47.2957	47.0020	46.8464	46.7639	46.72	46.6966	46.6842	

Table 4. $T_1 = 5$ and $p = 2$.

χ	0.0461	0.1141	0.2208	0.3291	0.4043	0.4482	0.4722	0.4852
	0.4921	0.4958	0.4976	0.4988	0.4994	0.4997	0.4998	0.4999
T	5.0	10.5345	18.6429	27.5214	34.8496	39.8314	42.8594	44.5904
	45.5479	46.0682	46.3483	46.4984	46.5785	46.6213	46.6441	46.6562

Finally, some sequences of temperatures and acceptance ratios obtained by the algorithm are given in Tables 1–4. We intend to compute a temperature corresponding with an acceptance ratio $\chi_0 = 0.5$. Instead of using T_1, defined in Eq. (7), we take $T_1 = 500$ to perform the experiments of Tables 1 and 3, and $T_1 = 5$ for the experiments of Tables 2 and 4. The precision ϵ is here 10^{-4}. Parameter p used in the recursive formula 6 is equal to 1 in Tables 1 and 2. We take $p = 2$ in Tables 3 and 4.

We can see that the convergence of the algorithm is slower for $p = 2$ than for $p = 1$. This can be easily understood from formula 6. However, recall that when p increases then the derivative of the function $T \to T(\frac{\ln(\tilde{\chi}(T))}{\ln(\chi_0)})^{\frac{1}{p}}$ is more likely to be positive (Section 1.1). Said another way, as we know that Hypothesis 1.1 is not always valid, it may be more advisable to take $p > 1$. Although we did not need to take $p > 1$ to compute the temperature in the case of these random problems, we will see in the next section that this may be necessary in very few of cases.

3.2. Traveling salesman problems

The algorithm of Section 1 is applied here for the traveling salesman problem. We consider transitions based on the very classical 2-OPT moves (see, for example [9]). Two sets of

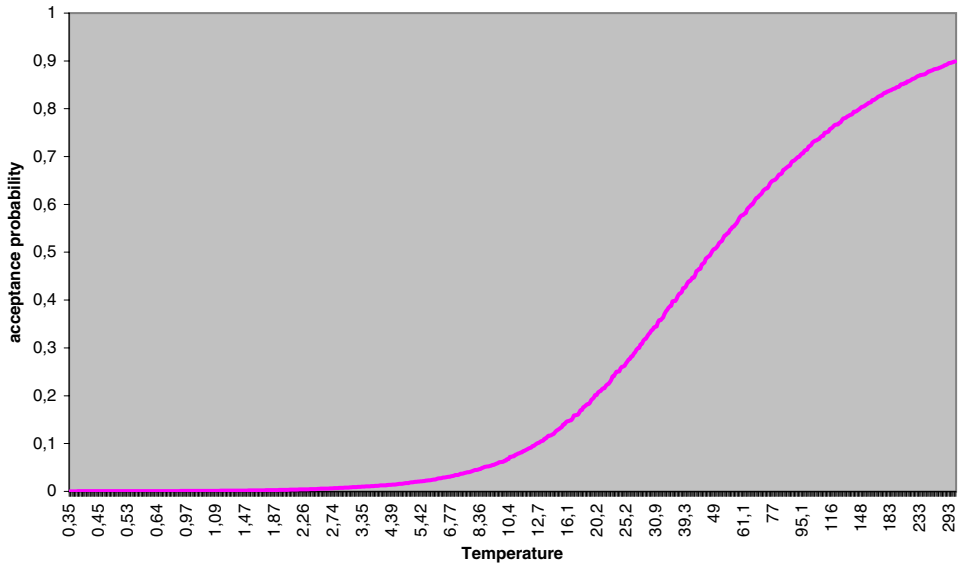

Figure 1. Evolution of $\overline{\chi(T)}$ for TSP(100).

randomly generated Euclidean instances are used: 20-city and 100-city problems. Notice that these problems can now be solved by efficient cutting plane algorithms.

Three values of $\|S\|$ are considered: 20, 2500 and 62500. Eight values of χ_0 are used: 0.99, 0.9, 0.7, 0.5, 0.3, 0.1, 0.05 and 0.01. 200 experiments are performed for each value of χ_0 and $\|S\|$.

Comments given in the previous subsection about random problems are still valid here. However, we noticed that $\|S\| = 2500$ was sufficient to give a good approximation of the temperature in the previous case, but does not seem to be sufficient here for some values of χ_0. In other words, when the size of problems increases, we need larger size samples to obtain a good approximation of the temperature.

The procedure used to compute the temperature does not converge in about 1 run per 1000 when $p = 1$. If $p = 2$, the algorithm always converges. In fact, when the algorithm is applied, it is easy to check whether there is an oscillation in terms of temperature. In this case, we multiply p by 2 and we continue the algorithm.

The experimental acceptance ratio is plotted as a function of temperature. The graph of figure 1 corresponds with the 100-city problem. This ratio is, as claimed in Corollary 2.6, convex for low temperatures and concave for high temperatures.

Finally, we studied the experimental number of plateaux of simulated annealing when a geometric cooling schedule ($\alpha = 0.95$) is used. Different values of the initial acceptance ratio χ_0 and the final acceptance ratio χ_f are considered. The number of Plateaux is compared with the lower bound of Proposition 2.3. This lower bound seems to be good for intermediate acceptance ratios and bad for extremal acceptance ratios (either very low or very high ratios).

4. Conclusion

A simple algorithm is proposed to compute a temperature such that the acceptance ratio of increasing cost moves is equal to a given value χ_0. We also presented some properties of the acceptance ratio.

We think that this algorithm can be used as a component of either classical or modern simulated annealing schemes for which the cooling schedule is not necessarily monotonous.

The procedure proposed in this paper can be modified in different ways. First, the formula linking T_n and T_{n+1} can be changed. Said another way, even if the algorithm is very fast, one can find another formula allowing a faster convergence. Second, We assumed in this paper that transitions are accepted in accordance with the Metropolis criterion. A further research direction may consist in introducing some modifications and studying the convergence of the algorithm when other acceptance probabilities are considered.

Finally, we considered the acceptance ratio of positive transitions. Although, we may want to focus on the acceptance of all transitions. A similar algorithm allowing the computation of a temperature that is compatible with a given acceptation probability of all transitions is now under study.

Acknowledgment

I would like to thank an anonymous referee for his valuable comments.

References

1. E. Aarts, J. Korst, and P. van Laarhoven, "Simulated annealing," in Local Search in Combinatorial Optimization, E.H.L. Aarts and J.K. Lenstra (Eds.), John Wiley and Sons, Ltd., 1997, pp. 91–120.
2. E. Bonomi and J.-L. Lutton, "The N-city traveling salesman problem: Statistical mechanisms and the metropolis algorithm," SIAM Review, vol. 26, pp. 551–568, 1984.
3. V. Cerny, "A thermodynamical approach to the traveling salesman problem: An efficient simulated algorithm," Journal of Optimization Theory and Applications, vol. 45, pp. 41–51, 1985.
4. H. Cohn and M. Fielding, "Simulated annealing: Searching for an optimal temperature schedule," SIAM J. Optim, vol. 3, pp. 779–802, 1999.
5. L. Grover, "A new simulated annealing algorithm for standard cell placement," in Proc. IEEE ICCAD-86, Santa Clara, CA, 1986.
6. B. Hajek, "Cooling schedules for optimal annealing," Math. Oper. Res., vol. 13, pp. 311–329, 1988.
7. D.S. Johnson, C.R. Aragon, L.A. McGeoch, and C. Schevon, "Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning," Operations Research, vol. 37, pp, 865–892, 1989.
8. D.S. Johnson, C.R. Aragon, L.A. McGeoch, and C. Schevon, "Optimization by simulated annealing: An experimental evaluation; part II, graph coloring and number partitioning," Operations Research, vol. 39, pp. 378–406, 1991.
9. D.S. Johnson and L.A. McGeoch, "The traveling salesman problem: A case study," in Local Search in Combinatorial Optimization, E.H.L. Aarts and J.K. Lenstra (Eds.), John Wiley and Sons, Ltd., 1997, pp. 215–310.
10. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by simulated annealing," Science, vol. 220, pp. 671–680, 1983.
11. N.A. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys, vol. 21, pp. 1087–1092, 1953.

12. D. Mitra, F. Romeo, and A.L. Sangiovanni-vincentelli, "Convergence and finite-time bahavior of simulated annealing," Advances in Applied Probability, vol. 18, pp. 747–771, 1986.
13. F. Romeo and A.L. Sangiovanni-Vincentelli, "A theoretical framework for simulated annealing," Algorithmica, vol. 6, pp. 302–345, 1991.
14. P. Van Laarhoven and E. Aarts, Simulated Annealing: Theory and Applications, D. Reidel Publishing Company, 1988.
15. J. Varanelli, "On the acceleration of simulated annealing," PhD Dissertation, University of Virginia, 1996.
16. S. White, "Concepts of scale in simulated annealing," in Proc. IEEE Int. Conference on Computer Design, Port Chester, 1984.