1

Simulated Annealing

1.1 Introduction

The complex structures of the configuration space of a difficult optimization problem (as shown in the figure 0.2 of the foreword) inspired to draw analogies with physical phenomena, which led three researchers of IBM society — S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi — to propose in 1982, and to publish in 1983, a new iterative method: the simulated annealing technique [Kirkpatrick et al., 1983], which can avoid the local minima. A similar work, developed independently at the same time by V. Cerny [Cerny, 1985], was published in 1985.

Since its discovery, the simulated annealing method has proved its effectiveness in various fields such as the design of the electronic circuits, the image processing, the collection of the household garbage, or the organization of the data-processing network of French Loto... On the other hand it has been found too greedy or incapable of solving certain combinatorial optimization problems, which could be solved better by some specific heuristics.

This chapter starts with initially explaining the principle of the method, with the help of an example of the layout problem of an electronic circuit. This is followed by a simplified description of some theoretical approaches of simulated annealing, which underlines its strong points (conditional guaranteed convergence towards a global optimum) and its weak points (tuning of the parameters, which can be delicate in practice). Then various techniques of parallelization of the method are discussed. This is followed by the presentation of some applications. In conclusion, we recapitulate the advantages and the most significant drawbacks of simulated annealing. To conclude, we put forth some simple practical suggestions, intended for those users who plan to develop their first application based on simulated annealing. In appendix A of this book, we recapitulate the main results of the simulated annealing modeling based on Markov chains.

This chapter partly presents a summary of the synthesis book carried out on the simulated annealing technique [Siarry and Dreyfus, 1989], which we

published in the beginning of 1989; this presentation is properly augmented by mentioning the more recent developments [Siarry, 1995, Reeves, 1995]. The references mentioned in the text were selected either because they played a significant role, or because they illustrate a precise point of the discussion. A much more exhaustive bibliography — although old — can be found in the works [Siarry and Dreyfus, 1989, Van Laarhoven and Aarts, 1987] [Wong et al., 1988, Sechen, 1988] and in the article [Collins et al., 1988] published on the subject. Interested readers are also requested to go through the elaborate presentations of simulated annealing which appeared in the article [Pirlot, 1992] and in chapter 2 of the book [Reeves, 1995].

1.2 Presentation of the method

1.2.1 Analogy between an optimization problem and some physical phenomena

The idea of simulated annealing can be illustrated by a vision inspired by the layout problem and routing of the electronic circuits: let us assume that a relatively inexperienced electronics specialist randomly spread the components on a plane, and connections were established as indicated in figure 0.4a of the foreword.

It is clear that the solution obtained is an unacceptable one. The purpose of developing a layout-routing program is to transform this disordered situation to lead to an ordered electronic circuit diagram (figure 0.4b of the foreword), where all connections are rectilinear, components are aligned and placed so as to minimize the length of the connections. In other words, this program must carry out a disorder-order transformation which, on the basis of a "liquid of components", leads to an ordered "solid".

However such a transformation occurs spontaneously in nature if the temperature of a system is gradually lowered; there are computer based digital simulation techniques available, which exhibit the behavior of sets of particles in interaction according to the temperature. In order to apply these techniques to the optimization problems, an analogy can be established which is presented in table 1.1.

To lead a physical system to a low energy state, the physicists generally use the annealing technique: we will examine how this method of treatment

Table 1.1. Analogy between an optimization problem and a physical system.

Optimization problem	*physical system*
objective function	free energy
parameters of the problem	"coordinates" of the particles
find a "good" configuration (even optimal configuration)	find the low energy states

of materials (real annealing) is helpful to deal with an optimization problem (simulated annealing).

1.2.2 Real annealing and simulated annealing

To modify the state of a material, the physicists have an adjustable parameter: the temperature. To be specific, annealing is a strategy where an optimum state can be approached by controlling the temperature. To have a deeper understanding, let us consider the example of the growth of a monocrystal. The annealing technique consists in heating a material beforehand to impart high energy to it. Then the material is cooled slowly, by keeping at each stage a temperature of sufficient duration; if the decrease in temperature is too fast, it may cause defects which can be eliminated by local reheating. This strategy of a controlled decrease of the temperature leads to a crystallized solid state, which is a stable state, corresponding to an absolute minimum of energy. The opposite technique is that of the quenching, which consists in very quickly lowering the temperature of the material: this can lead us to an amorphous structure, a metastable state that corresponds to a local minimum of energy. In the annealing technique the cooling of a material caused a disorder-order transformation, while the quenching technique was responsible in solidifying a disordered state.

The idea to use the annealing technique in order to deal with optimization problems gave rise to the simulated annealing technique. It consists in introducing a control parameter in optimization, which plays the role of the temperature. The "temperature" of the system to be optimized must have the same effect as the temperature of the physical system: it must condition the number of accessible states and lead towards the optimal state, if the temperature is lowered gradually in a slow and well controlled manner (as in the annealing technique) and towards a local minimum if the temperature is lowered abruptly (as in the quenching technique).

To conclude, we have to describe the algorithm in such a way that will enable us to implement the annealing in a computer.

1.2.3 Simulated annealing algorithm

The algorithm is based on two results of statistical physics.

On one hand, when thermodynamic balance is reached at a given temperature, the probability for a physical system to have a given energy E, is proportional to the Boltzmann factor: $e^{\frac{-E}{k_B T}}$, where k_B denotes the Boltzmann constant. Then, the distribution of the energy states is the Boltzmann distribution at the temperature considered.

In addition, to simulate the evolution of a physical system towards its thermodynamic balance at a given temperature, the Metropolis algorithm [Metropolis et al., 1953] can be utilized: on the basis of a given configuration

(in our case, an initial layout for all the components), the system is subjected to an elementary modification (for example, a component is relocated, or two components are exchanged); if this transformation causes a decrease in the objective function (or "energy") of the system, it is accepted; on the contrary, if it causes an increase ΔE of the objective function, it is also accepted, but with a probability $e^{\frac{-\Delta E}{T}}$ (in practice, this condition is realized in the following manner: a real number is drawn at random ranging between 0 and 1, and the configuration causing a ΔE degradation in the objective function is accepted, if the random number drawn is lower than or equal to $e^{\frac{-\Delta E}{T}}$). By repeatedly observing this Metropolis rule of acceptance, a sequence of configurations is generated, which constitutes a Markov chain (in a sense that each configuration depends on only that one which immediately precedes it). With this formalism in place, it is possible to show that, when the chain is of infinite length (in practical consideration, of "sufficient" length. . .), the system can reach (in practical consideration, can approach) thermodynamic balance at the temperature considered: in other words, this leads us to a Boltzmann distribution of the energy states at this temperature.

Hence the role entrusted to the temperature by the Metropolis rule is well understood . At high temperature, $e^{\frac{-\Delta E}{T}}$ is close to 1, therefore the majority of the moves are accepted and the algorithm becomes equivalent to a simple random walk in the configuration space . At low temperature, $e^{\frac{-\Delta E}{T}}$ is close to 0, therefore the majority of the moves increasing energy is refused. Hence the algorithm reminds us of a classical iterative improvement. At an intermediate temperature, the algorithm intermittently authorizes the transformations that degrade the objective function: hence it leaves a scope for the system to be pulled out of a local minimum.

Once the thermodynamic balance is reached at a given temperature, the temperature is lowered " slightly", and a new Markov chain is implemented at this new temperature stage (if the temperature is lowered too quickly, the evolution towards a new thermodynamic balance is slowed down: the theory establishes a narrow correlation between the rate of decrease of the temperature and the minimum duration of the temperature stage). By comparing the successive Boltzmann distributions obtained at the end of the various temperature stages, a gradual increase in the weight of the low energy configurations can be noted: when the temperature tends towards zero, the algorithm converges towards the absolute minimum of energy. In practice, the process is terminated when the system is "solidified" (it means that either the temperature reached the zero value or no more moves causing increase in energy were accepted during the stage). The flow chart of the simulated annealing algorithm has been presented in figure 0.4 of the foreword.

1.3 Theoretical approaches

The simulated annealing algorithm gave rise to many theoretical works because of the two following reasons: on one hand, it was a new algorithm, for which it was necessary to establish the conditions of convergence; in addition, the method comprises of many parameters as well as variations, whose effect or influence on the mechanism should be properly understood, if one wishes to implement this method with maximum effect.

These approaches, specially those which appeared during the initial years of the formulation, are presented in detail in the book [Siarry and Dreyfus, 1989]. Here, we keep ourselves focused to emphasize on the principal aspects treated in the literature. The theoretical convergence of simulated annealing is analyzed first. Then those factors which are influential in the operation of the algorithm are analyzed in detail: the structure of the configuration space, the acceptance rules and the annealing program.

1.3.1 Theoretical convergence of simulated annealing

Many noted mathematicians have invested their research efforts in the convergence of the simulated annealing (see in particular [Aarts and Van Laarhoven, 1985, Hajek, 1988, Hajek and Sasaki, 1989]) or some of them even endeavored to develop a general model for the analysis of the stochastic methods for global optimization (notably [Rinnooy Kan and Timmer, 1987a, Rinnooy Kan and Timmer, 1987b]). The main outcome of these theoretical studies is the following: under certain conditions (discussed later), simulated annealing probably converges towards a global optimum, in a sense that it is made possible to obtain a solution arbitrarily close to this optimum, with a probability arbitrarily close to unity. This result is, in itself, significant because it distinguishes simulated annealing from other metaheuristic competitors, whose convergence is not guaranteed.

However, the establishment of the "conditions of convergence" is not unanimously accepted. Some of these, like those proposed by Aarts et al. [Aarts and Van Laarhoven, 1985], are based on the assumption of decreasing the temperature in stages. This property enables to represent the optimization process in the form of completely connected homogeneous Markov chains, whose asymptotic behavior can be simply described. It has also been shown that the convergence is guaranteed provided that on one hand the reversibility is respected (the opposite change of any change allowed must also be allowed) and on the other hand the connectivity (any state of the system can be reached starting from any other state with the help of a finite completed number of elementary changes) of the configuration space is also maintained. This formalization presents two advantages:

- it enables us to legitimize the lowering of the temperature in stages, which improves the convergence speed of the algorithm;

- it enables us to establish that a "good" quality solution (located significantly close to the global optimum) can be obtained by simulated annealing in a polynomial time, for certain NP-hard problems [Aarts and Van Laarhoven, 1985].

Some of the other authors, Hajek et al. [Hajek, 1988, Hajek and Sasaki, 1989] in particular, were interested in the convergence of the simulated annealing within the more general framework of the theory of the inhomogeneous Markov chains. In this case, the asymptotic behavior was the more sensitive aspect of study. The main result of this work is the following: the algorithm converges towards a global optimum with a probability of unity if, when time t tends towards infinity, the temperature $T(t)$ does not decrease more quickly than the expression $\frac{C}{\log(t)}$, where C is a constant, related to the depth of the "energy wells" of the problem. It should be stressed that the results of this theoretical work, till now, are not sufficiently generalized and unanimous to be used as a guide for the experimental approach, when one is confronted with a new problem. For example, the logarithmic law of decrease of the temperature, recommended by Hajek, is not used in practice for two major reasons: on one hand it is generally impossible to evaluate the depth of the energy wells of the problem, on the other hand this law introduces an unfavorable increase in computing time...

This analysis is now prolonged by careful, individual examination of the various components of the algorithm.

1.3.2 Configuration space

The configuration space plays a fundamental role in the effectiveness of the simulated annealing technique to solve a complex optimization problem. It is equipped with a "topology", originating from the concept of proximity between two configurations: the "distance" between two configurations represents the minimum number of elementary changes required to pass from one configuration to the other. Moreover, there is an energy associated with each configuration, so that the configuration space is characterized by an " energy landscape". All the difficulties of the optimization problem lie in the fact that the energy landscape comprises of a large number of valleys of varying depth, possibly relatively close to each other, which correspond to local minima of energy.

It is clear that the shape of this landscape is not specific to the problem under study, but to a large extent depends on the choice of the cost function and the choice of the elementary changes. On the other hand, the required final solution i.e. the global minimum (or one of the global minima of comparable energy) must depend primarily on the nature of the problem considered, and not (or very little) on the preceding choices. We showed, with the help of an example problem of placement of building blocks, considered specifically for this purpose, that an apparently difficult problem can be largely simplified,

either by widening the allowable configuration space, or by choosing a better adapted topology [Siarry and Dreyfus, 1989].

Several authors endeavored to establish general analytical relations between certain properties of the configuration space and the convergence of simulated annealing. In particular, some of these works were directed towards the analysis of the energy landscapes, and they searched to develop any link between the "ultrametricity" and simulated annealing [Kirkpatrick and Toulouse, 1985, Rammal et al., 1986, Solla et al., 1986]: the simulated annealing method would be more effective for those optimization problems whose low local minima (i.e. the required solutions) form an ultrametric set. Thereafter, G.B. Sorkin [Sorkin, 1991] showed that certain fractal properties of the energy landscape induce a polynomial convergence of simulated annealing; such an explanation was provided by the author on the basis of the effectiveness of the method in the field of the electronic circuit layouts. In addition, Azencott et al. [Azencott, 1992] utilized the "theory of the cycles" (originally developed in the context of the dynamic systems) to establish general explicit relations between the geometry of the energy landscape and the expected performances of simulated annealing. This work led them to propose the "method of the distortions" for the objective function, which significantly improved the quality of the solutions for certain difficult problems [Delamarre and Virot, 1998]. However, all these approaches of simulated annealing are still in a nascent stage, and their results are not yet generalized.

Lastly, another aspect of immediate practical interest relates to the adaptation of the simulated annealing for the solution of the continuous optimization problems [Siarry, 1994, Courat et al., 1994]. This subject is examined more in detail in chapter 6 of this book. Here, we put stress only on the transformations necessary to graduate from the "combinatorial simulated annealing" to the " continuous simulated annealing". Indeed, the method was originally developed for application in the domain of the combinatorial optimization problems, where the free parameters can take discrete values only. In the majority of these types of problems encountered in practice, topology is considered almost always as data for the problem: for example, in the traveling salesman problem, the permutation of two cities is a natural way to generate the rounds close to a given round. It is the same in the layout of the components for the exchange of two blocks. On the other hand, when the objective is to optimize a function of continuous variables, the topology has to be updated. This gives rise to the concept of "adaptive topology": here the length of the elementary steps is not imposed any more by the problem. This choice must be dictated by a compromise between two extreme situations: if the step is too small, the program explores only a limited region of the configuration space; the cost function is then improved very often, but in negligible amount. On the contrary, if the step is too large, the tests are accepted only seldom, and they are almost independent of each other. We will examine in the chapter 6 some of the published algorithms , which are generally developed utilizing an empirical step. From the point of mathematical interest, it is necessary to

underline the work of L Miclo [Miclo, 1991], which was directed towards the convergence of the simulated annealing in the continuous case.

1.3.3 Rules of acceptance

The principle of simulated annealing requires that one accepts, occasionally and under the control of the "temperature", an increase in the energy of the current state, which enables it to be pulled out of a local minimum. The rule of acceptance generally used is the Metropolis rule described in section 1.2.3. It possesses an advantage that it originates directly from statistical physics. There are however several variations of this rule [Siarry and Dreyfus, 1989], which can be more effective from the point of view of the computing time.

Another aspect can be the examination of the following problem: at low temperature the rate of acceptance of the algorithm becomes very low, hence the method is ineffective. It is a well-known problem encountered in simulated annealing, which can be solved by substituting the traditional Metropolis rule with an accelerated alternative, called "thermostat" [Siarry and Dreyfus, 1989], as soon as the rate of acceptance falls too low. In practice, this methodology is rarely employed.

1.3.4 Annealing scheme

The convergence speed of the simulated annealing methodology depends primarily on two factors: the configuration space and the program of annealing. With regard to the configuration space, the effects on convergence of topology and the shape of the energy landscape were described above. Let us discuss the influence of the "program of annealing": it addresses the problem of controlling the "temperature" as well as the possibility of a system to reach, as quickly as possible, a solution. The program of annealing must specify the following values of the control parameters of the temperature:

- the initial temperature;
- the length of the homogeneous Markov chains, i.e. the criterion for change of temperature stage;
- the law of decrease of the temperature;
- the criterion for program termination.

In the absence of general theoretical results, which can be readily exploited, the user has to take resort to an empirical adjustment of these parameters. For certain problems, the task is even complicated by the great sensitivity of the result (and the computing time) to this adjustment. This aspect — that unites simulated annealing with other metaheuristics — is an indisputable disadvantage of this method.

To elaborate the subject a little more, we deliberate on the characteristic of the program of annealing which drew most attention: the law of decrease

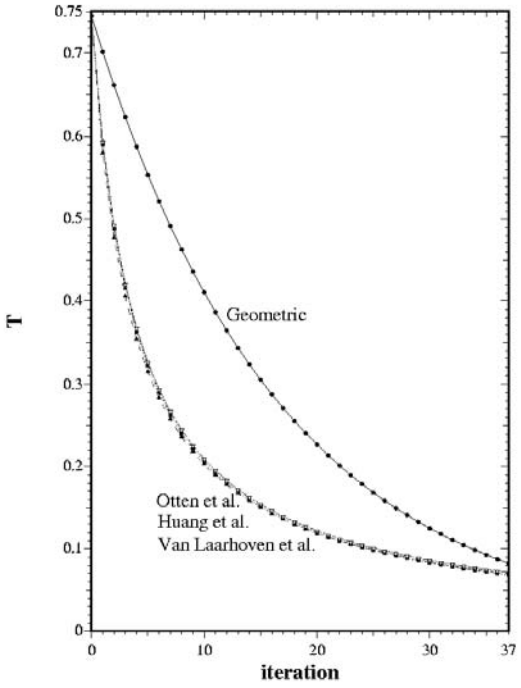

Fig. 1.1. Lowering of the temperature according to the number of stages for the geometrical law and several traditional laws.

of the temperature. The geometrical law of decrease: $T_{k+1} = \alpha \cdot T_k$, $\alpha = constant$, is a widely accepted one, because of its simplicity. An alternative solution, potentially more effective, consists in resorting to an adaptive law of the form: $T_{k+1} = \alpha(T_k) \cdot T_k$, but it is then necessary to exercise a choice among several laws suggested in the literature. One can show, however, that several traditional adaptive laws, having quite different origins and mathematical expressions are, in practice, equivalent (see figure 1.1), and can be expressed in the following generic form:

$$T_{k+1} = \left(1 - T_k \cdot \frac{\Delta\left(T_k\right)}{\sigma^2\left(T_k\right)}\right) \cdot T_k$$

where:
$\sigma^2\left(T_k\right) = \left\langle f_{T_k}^2 \right\rangle - \left\langle f_{T_k} \right\rangle^2$,

f denotes the objective function,
$\Delta(T_k)$ depends on the adaptive law selected.

The simplest adjustment, $\Delta(T_k) = constant$, can then be made effective, although it does not correspond to any of the traditional laws.

Due to the inability in synthesizing the results (theoretical and experimental) showing some disparities presented in the literature, the reader is redirected to the paragraph 1.7, where we propose a suitable tuning algorithm for the four parameters of the program of annealing, which can often be useful, at least to start with.

Those readers who are interested in the mathematical modeling of simulated annealing are advised to refer to the appendix A at the end of this book: the principal results produced by the Markov formalism are described there.

1.4 Parallelization of the simulated annealing algorithm

Often, the computing time becomes a critical factor in the economic evaluation of the utility of a simulated annealing technique, for applications in real industrial problems. To reduce this time, a promising research direction is the parallelization of the algorithm, which consists in simultaneously carrying out several calculations necessary for its realization. This step can be considered in the context of the significant activities which have developed around the algorithms and architectures of parallel computation for quite some time now. However, this may appear paradoxical, because of the sequential structure of the algorithm. Nevertheless, several types of parallelization have been considered to date. A book [Azencott, 1992] has been published which was completely devoted to this direction; it has described at once the rigorous mathematical results available and the simulation results, executed on parallel or sequential computers. To have a concrete idea, we shall describe the idea behind two principal modes of parallelization, which are independent of the problem dealt with and were suggested very soon after the invention of simulated annealing. The distinction of these two modes remains relevant to date, as has been shown in the recent status of the state of the art described by Delamarre and Virot in [Delamarre and Virot, 1998].

The first type of parallelization [Aarts et al., 1986] consists in implementing several Markov chain computations in parallel, by using K elementary processors. To implement this, the algorithm is decomposed into K elementary processes, constituting K Markov chains. If L be the length of these Markov chains, assumed constant, each chain is divided into K sub-chains of length $\frac{L}{K}$. The first processor executes the first chain at the initial temperature, and implements the first $\frac{L}{K}$ elements of this chain (i.e. the first sub-chain); then it calculates the temperature of the following Markov chain, starting from the states already obtained. The second elementary processor then begins executing the second Markov chain at this temperature, starting

from the final configuration of the first sub-chain of the first chain. During this time, the first processor begins the second sub-chain of the first chain. This process continues for the K elementary processors. It has been shown that this mode of parallelization — described in more detail in the reference [Siarry and Dreyfus, 1989] — allows to divide the total computing time by a factor K, if K is small compared to the total number of Markov chains carried out. However, the procedure presents a major disadvantage: its convergence towards an optimum is not guaranteed. Indeed, the formalism of the Markov chains enables to establish that the convergence of simulated annealing is assured provided that the distribution of the states, at the end of each Markov chain, is close to the stationary distribution. In the case of the algorithm described, this proximity is not established at the end of each sub-chain, and larger the number K of the processors in parallel, larger is the deviation from the proximity.

The second type of parallelization [Kravitz and Rutenbar, 1987] [Roussel-Ragot et al., 1990] consists in carrying out the computation in parallel for several states of the same Markov chain, while the following condition must always be kept in mind: at a low temperature, the number of elementary transformations rejected becomes very important; it is thus possible to consider that these moves are produced by independent elementary processes, which may likely be implemented in parallel. Then the computing time can be divided approximately by the number of processes. A strategy consists in subdividing the algorithm into K elementary processes, each of which is responsible to calculate the energy variations corresponding to one or more elementary moves, and to carry out the corresponding Metropolis tests. Two operating modes are considered:

- at "high temperature", a process corresponds to only one elementary move. Each time K elementary processes were implemented in parallel, one can randomly choose a transition among those which were accepted, and the memory, containing the best solution known, is updated with the new configuration;
- at "low temperature", the accepted moves become very rare: less than one transition is accepted for K moves carried out. Each process then consists in calculating the energy variations corresponding to a succession of disturbances, until one of them is accepted. As soon as any of the processes succeeds, the memory is updated.

These two operating modes can ensure a behavior, and in particular a convergence, which is strictly identical to those of the sequential algorithms. This type of parallelization was tested by experimenting on the optimization problem of the placement of connected blocks [Roussel-Ragot et al., 1990]. We estimated the amount of computing time saved in two cases: the placement of presumed point blocks in predetermined sites and the placement of real blocks on a plane. With 5 elementary processes in parallel, the saving in computing time lies between 60 % and 80 %, according to the program of

annealing used. This work was then continued, within the scope of the thesis work of P. Roussel-Ragot [Roussel-Ragot, 1990], by considering a theoretical model, which was validated by programming the simulated annealing using a network of "Transputers".

In addition to these two principal types of parallelization of simulated annealing, which should be applicable for any optimization problem, other methods were proposed to deal with specific problems. Some of these problems are problems of placement of electronic components, problems of image processing and problems of meshing (for the finite element method). In each of these three cases, information is distributed in a plane or in space, and each processor can be entrusted with the task to optimize the data pertaining to a geographical area by simulated annealing; here information is exchanged periodically between the neighboring processors.

Another step was planned to reduce the cost of synchronizations between the processors: the algorithms known as "asynchronous" agree to calculate the energy variations starting from partially out-of-date data. However it seems very complex and sensitive to control the admissible error, except for certain particular problems [Durand and White, 1991].

As an example, let us describe the asynchronous parallelization technique, suggested by Casotto et al. [Casotto et al., 1987] to deal with the problem of the placement of electronic components. The method consists in distributing the components to be placed in K independent groups, respectively assigned to K processors. Each processor applies the simulated annealing technique to seek the optimal site for the components that belong to its group. The processors function in parallel, and in an asynchronous manner to each other. All of them have access to a common memory, which contains the current state of the circuit plan. When a processor plans to exchange the position of a component of its group with that of an affected component in another group pertaining to another processor, it temporarily blocks the activity of this processor. It is clear that the asynchronous working of the processors involves errors, in particular in the calculation of the overlapping between the blocks, and thus in the evaluation of the cost function. In fact, when a given processor needs to evaluate the cost of a move (translation or permutation), it will search, in the memory, the current position of all the components of the circuit . However the information collected is partly erroneous, since certain components are in the course of displacement, because of activities of the other processors. In order to limit these errors, the method is supplemented by the two following provisions. On one hand, the distribution of the components between the processors is in itself an object of optimization by simulated annealing technique, which is performed simultaneously with the optimization process already described: in this manner, the membership of the components geographically close to the same group can be favored. In addition, the maximum amplitude of the moves carried out by the components is reduced as the temperature

decreases. Consequently, when the temperature decreases, the moves mainly relate to nearby components, thus generally belonging to the same group . In this process the interactions between the groups can be reduced, thus reducing the frequency of the errors mentioned above. This technique of parallelization of simulated annealing was validated using several examples of real circuits: the algorithm functions approximately six times faster with eight processors than with only one, the results being of comparable quality with those of the sequential algorithm.

1.5 Some applications

The majority of the preceding theoretical approaches are based on asymptotic behaviors which impose several restrictive assumptions, very often causing excessive enhancements in computing times. This is why, to solve real industrial problems under reasonable conditions, it is often essential to adopt an experimental approach, which may frequently result in crossing the barriers recommended by the theory. The simulated annealing method proved to be interesting in solving many optimization problems, NP-hard or not. Some examples of these problems are presented here.

1.5.1 Benchmark problems of combinatorial optimization

The effectiveness of the method was initially tested on the benchmark problem instances of combinatorial optimization. In this type of problem, the practical purpose is secondary: the initial objective is to develop the optimization method, and to compare its performances with those of the other methods. We will detail only one example: that of the traveling salesman problem.

The reason for the choice of this problem is that it is very simple to formulate, and, at the same time, very difficult to solve: the larger problems for which the optimum was found, and proved, comprise of a few thousands of cities. To illustrate the disorder-order transformation, carried out by the simulated annealing technique, as the temperature goes down, we present, in the figure 1.2, three intermediate configurations obtained on 13206 nodes of the Swiss road network..

Bonomi and Lutton considered very high dimensional examples: between 1000 and 10000 cities [Bonomi and Lutton, 1984]. They showed that, to prevent a prohibitive computing time, the domain containing the cities can be deconstructed into areas, and the moves for a round of the traveler can be so forced that they are limited between the cities located in contiguous areas. Bonomi and Lutton compared simulated annealing with the traditional techniques of optimization, for the traveling salesman problem: simulated annealing was slower for small dimensional problems (N lower than 100); on the other hand, it was, by far, more powerful for higher dimensional problems (N higher than 800). The traveling salesman problem has been extensively studied

to illustrate and establish several experimental and theoretical developments on the simulated annealing method [Siarry and Dreyfus, 1989].

Many other benchmark problems of combinatorial optimization were also solved using simulated annealing [Siarry and Dreyfus, 1989, Pirlot, 1992]: in particular the problems of the "partitioning of graph", of the "minimal coupling of points", of the "quadratic assignment" ... The comparison with the best known algorithms leads to different results, varying according to the problems and ... the authors. Thus the works by Johnson et al. [Johnson et al., 1989, Johnson et al., 1991, Johnson et al., 1992], which were devoted to a systematic comparison of several benchmark problems, conclude that the only benchmark problem which can find favor with simulated annealing is that of the partitioning of graph. For some problems, promising results with the simulated annealing method could only be observed for high dimensional examples (a few hundreds of variables), and that too at the cost of a high computing time. Therefore, if simulated annealing has the merit to adapt simply to a great diversity of problems, it cannot claim as much to supplement those specific algorithms that already exist for these problems.

We now present the applications of simulated annealing for practical problems. The first significant application of industrial interest was developed in the field of the electronic circuit design; this industrial sector still remains the biggest domain in which maximum number of application works using simulated annealing have been published. Two applications in the area of electronics are discussed in detail in two following paragraphs. This is followed by discussions regarding other applications in some other fields.

1.5.2 Layout of electronic circuits

The first application of the simulated annealing method for practical problems was developed in the field of the layout-routing of the electronic circuits [Kirkpatrick et al., 1983, Vecchi and Kirkpatrick, 1983, Siarry and Dreyfus, 1984]. Till now numerous works have been reported on this subject in several publications and, in particular, two books were completely devoted to this problem [Wong et al., 1988, Sechen, 1988]. A detailed bibliography, concerning the works carried out in the initial period 1982-1988, can be found in the books [Siarry and Dreyfus, 1989, Van Laarhoven and Aarts, 1987, Wong et al., 1988, Sechen, 1988].

The search for an optimal layout is generally carried out in two stages. The first consists in calculating an initial placement quickly, by a constructive method: the components are placed one after another, in order of decreasing connectivity. Then an algorithm for iterative improvement is employed that gradually transforms, by elementary moves (e.g. exchange of components, operations of rotation or symmetry etc.), the initial layout configuration. The algorithms for iterative improvement of the layout differ according to the rule adopted for the succession of elementary moves. Simulated annealing can be used in this second stage.

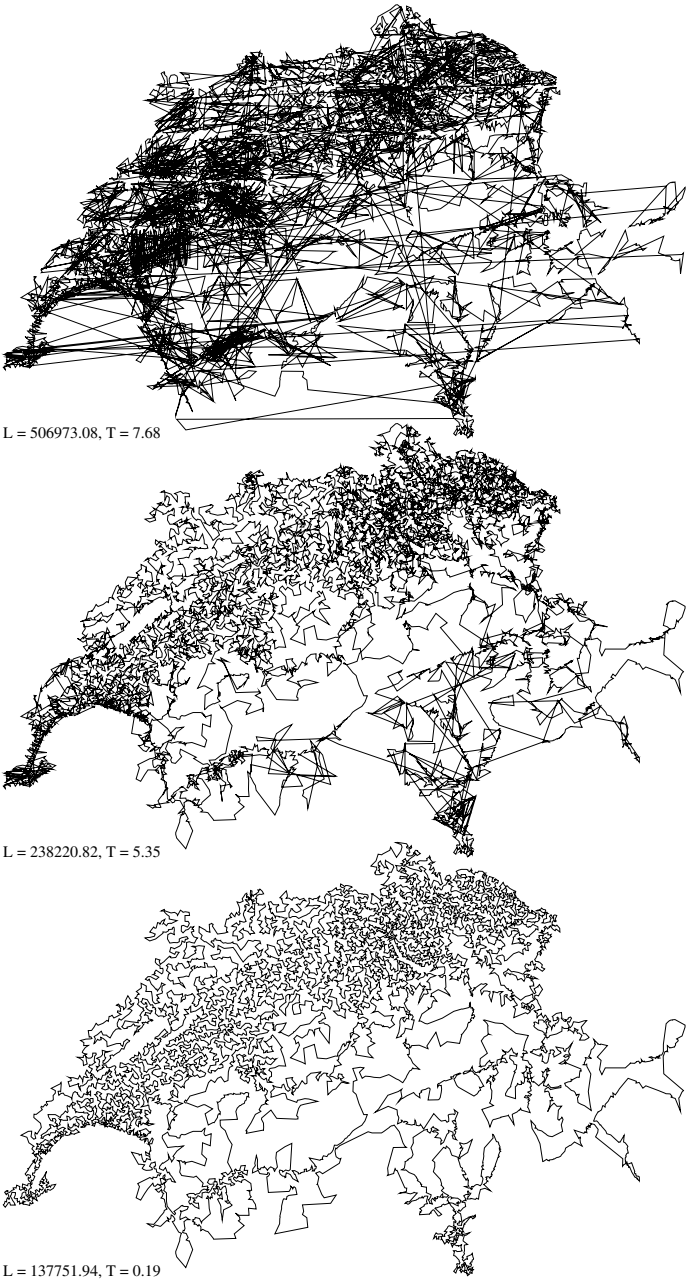

Fig. 1.2. The traveling salesman problem (13206 nodes of the Swiss road network): best known configurations (length: L) at the end of 3 temperature stages (T).

Our interest was concerned with a unit of 25 identical blocks to be placed on predetermined sites, which are the nodes of a planar square network. The list of connections is so that, in the optimal configurations, each block is connected only to its closer neighbors (see figure 0.6 of the foreword): an a priori knowledge about the global minima of the problem then facilitates to study the influence of the principal parameters of the method on its convergence speed. The cost function is the overall Manhattan length (i.e. the L length) of the connections. The only authorized elementary move is the permutation of two blocks. A detailed explanation for this benchmark problem on layout design — which is a form of " quadratic assignment" problem — can be found in the references [Siarry, 1986] and [Siarry et al., 1987]. Here, the discussions will be kept limited to the presentation of two examples of applications. First of all, to appreciate the effectiveness of the method, we start with a completely disordered initial configuration (see figure 0.6 of the foreword), and an initial "elevated" temperature (in the sense that at this temperature 90% of the moves are accepted): the figure 0.6 of the foreword represents the best configurations observed at the end of a few temperature stages. In this example, the temperature profile is that of a geometrical decrease, of ratio 0.9. A global optimum of the problem could be obtained after 12000 moves, whereas the total number of possible configurations is about 10^{25}.

To illustrate the advantages of the simulated annealing technique, we applied the traditional method of iterative improvement (simulated annealing at zero temperature), for the same initial configuration (see figure 1.3b), and by authorizing the same number of permutations as during the preceding test. It was observed that the traditional method got trapped in a local minimum (see figure 1.3c); it is clear that the shifting from this configuration to the optimal configuration as shown in the figure 1.3a would require several stages (at least five), majority of which correspond to an increase in energy, inadmissible by the traditional method. This problem of placement in particular made it possible to empirically develop a program of "adaptive" annealing, which could achieve gain in computing time by a factor of 2; the lowering of the temperature is carried out according to the law $T_{k+1} = D_k \cdot T_k$, with:

$$D_k = \min\left(D_0, \frac{E_k}{\langle E_k \rangle}\right)$$

that includes:

$D_0 = 0.5$ to 0.9

E_k is the minimal energy of the configurations accepted during the stage k

$\langle E_k \rangle$ is the average energy of the configurations accepted during the stage k

(at high temperature, $D_k = \frac{E_k}{\langle E_k \rangle}$ is small: hence the temperature is lowered quickly; at low temperature, $D_k = D_0$, this corresponds to a slow cooling).

Then we considered a more complex problem consisting of positioning components of different sizes, with an objective of simultaneous minimization of the length of the necessary connections and the surface area of the circuit used. In this case, the translation of a block is a new means of iterative transformation of the layout. Here we can observe that the blocks are overlapping with each other, what is authorized temporarily, but must be generally excluded from the final layout. This new constraint can be accommodated within the cost function of the problem, by introducing a new factor called the overlapping surface between the blocks. Calculating this surface area can become very cumbersome when the circuit comprises of many blocks. This is why the circuit was divided into several planar areas, whose size is such that a block can overlap only with those blocks located in the same area, or with one of the immediately close areas. The lists of the blocks belonging to each area are updated after each move, using a chaining method. Moreover, to avoid leading to a circuit congestion such as routing is impossible, a fictitious increase in the dimensions of each block is introduced. The calculation for the length of the connections consists in determining, for each equipotential, the barycentre of the terminations, and then to add the L distances of the barycentre with each termination. Lastly, the topology of the problem is adaptive, which can be described in the following manner: when the temperature decreases, the maximum amplitude of the translations decreases, and the exchanges are considered between the neighboring blocks only.

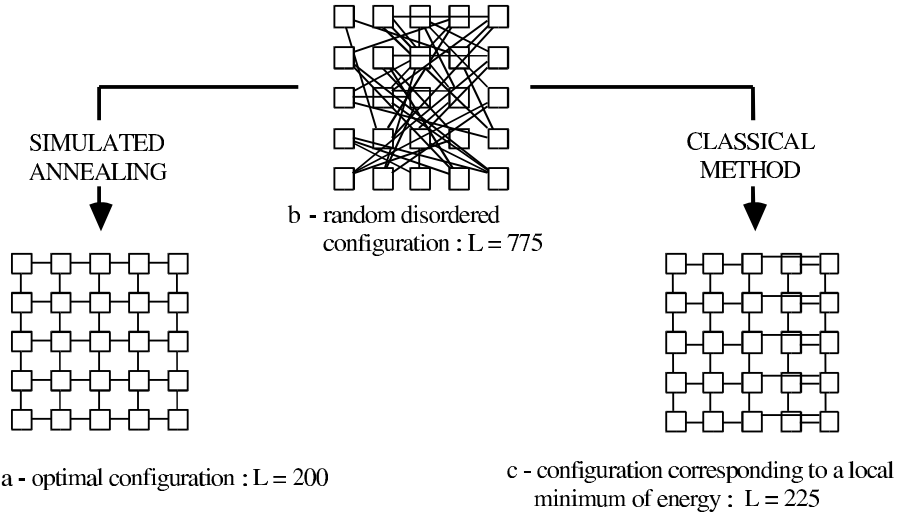

Fig. 1.3. The traditional method getting trapped in a local minimum of energy.

With the simulated annealing algorithm, it was possible to optimize industrial circuits, in particular in hybrid technology, in collaboration with the

Thomson D.C.H. (Department of the Hybrid Circuits) company. As an example, we present in the figure 1.4, the result of the optimization of a circuit layout comprising of 41 components and 27 equipotentials: the automated layout design procedure causes a gain of 18 % in the connection lengths, compared to the initial manual layout.

This study showed that the flexibility of the method enables it to take into account not only the rules of drawing, which translate the standards of technology, but also the rules of know-how, which are intended to facilitate the routing. Indeed, the rules of drawing impose in particular a minimal distance between two components, whereas the rules of know-how recommend a larger distance, allowing the passage of connections. To balance these two types of constraints, the calculation of the area of overlapping between the blocks, on a two by two basis, is undertaken according to the formula:

$S = S_r + a \cdot S_v$, where the notations indicate:

S_r	the "real" overlapping surface
S_v	the "virtual" overlapping surface
a	a weight factor (typically: 0.1)

Surfaces S_r and S_v are calculated by increasing dimensions of the components fictitiously, with a larger increase in S_v. Here, this induces some kind of an "intelligent" behavior, similar to that of an expert system. We notice, from the figure 1.4, a characteristic of the hybrid technology, which was easily incorporated in the program: the resistances, shown by a conducting link, can be placed under the diodes or the integrated circuits.

The observations noted by the majority of the authors concerning the application of the simulated annealing technique for the layout design problem conform to our observations: the method is very simple to implement, it adapts easily to various and evolving technological standards, and the final result is of good quality, but it is sometimes obtained at the cost of a significant computing time.

1.5.3 Search for an equivalent schema in electronics

We now present an application which mixes the combinatorial and the continuous aspects: automatic identification of the "optimal" structure of a linear circuit pattern. The objective was to automatically determine a model which comprises of the least possible number of elementary components, while ensuring a " faithful" reproduction of experimental data. This activity, in collaboration with the Institute of Fundamental Electronics (IEF, CNRS URA 22, in Orsay), began with the integration, in a single software, of a simulation program of the linear circuits (implemented in the IEF) and of a simulated annealing based optimization program developed by us. We initially validated this tool, by characterizing models of real components having a given structure (described using their distribution parameters S). A comparison with a

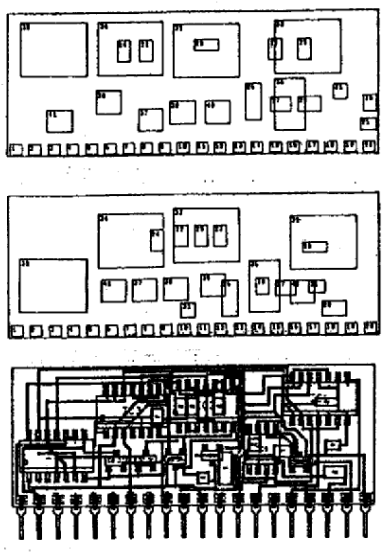

Fig. 1.4. Optimization by simulated annealing of the design of an electronic circuit layout comprising of 41 components.

- *drawing at the top: initial manual layout; length of connections: 9532;*
- *drawing at the middle: final layout, optimized by annealing; length of connections 7861;*
- *drawing at the bottom: manual routing using the optimized layout.*

commercial software (developed using the gradient method), at that moment in use in the IEF, showed that simulated annealing is particularly useful if the orders of magnitude of the parameters of the model are completely unknown: obviously the models under consideration are of this nature, since their structure even is to be determined. We developed an alternative simulated annealing, called logarithmic simulated annealing [Courat et al., 1994], which allows an effective exploration of the space of variations of the parameters, when this space is very wide (more than 10 decades per parameter). Then the problem of structure optimization was approached by the examination — in the case of a passive circuit — of progressive simplification of a general "exhaustive" model: we proposed a method which could be successfully employed to automate all the simplification stages [Courat et al., 1995]. This technique rests on the progressive elimination of the parameters, according to their statistical behavior during the process of optimization by simulated annealing.

We present, with the help of illustrations, the example of the search for an equivalent schema for an MMIC inductance, in the frequency range of 100 MHz to 20 GHz. On the basis of the initial "exhaustive" model with 12

parameters, as shown in the figure 1.5, and allowing each parameter to move over 16 decades, we obtained the equivalent schema shown in the figure 1.6 (the final values of the 6 remaining parameters are beyond the scope of our present interest: they are specified in [Courat et al., 1995]). The layouts in the Nyquist plane of the four S parameters of the quadripole of the figure 1.6 coincide nearly perfectly with the experimental results of MMIC inductance, and this is true for the entire frequency range [Courat et al., 1995].

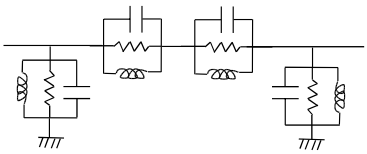

Fig. 1.5. Initial structure with 12 elements.

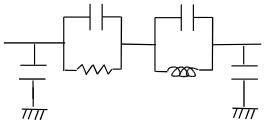

Fig. 1.6. Optimal structure with 6 elements.

1.5.4 Practical applications in various fields

An important field of application for simulated annealing happens to be image processing : here the main problem is to restore the images, by using computer, mainly in three-dimensional forms, starting from incomplete or irregular data. There are numerous practical applications in several domains like robotics, medicine (e.g. tomography), geology (e.g. prospections)... The restoration of an image using an iterative method involves, under normal circumstances, the treatment of a large number of variables. Hence it calls for development of a suitable method, which can limit the computing time of the operation. Based on the local features of the information contained in an image, several authors proposed numerous structures and algorithms specifically addressed to carry out calculations in parallel. Empirically, it appears that the simulated annealing method should be particularly well suited for this task. A rigorous theoretical justification of this property can be obtained starting from the concept of *Markovian field* [Geman and Geman, 1984], which provides a convenient and coherent model of the local structure of information in an image. This concept has been explained in detail in the reference

[Siarry and Dreyfus, 1989]. The "Bayesian approach" for the problem of optimal restoration of an image, starting from a scrambled image, consists in determining the image which presents "the maximum of a posteriori likelihood". It has been shown that this problem is ultimately configured as a well known minimization problem of an objective function, comprising of a very large number of parameters, e.g. light intensities of all the "pixels" of an image, in case of an image in black and white. Consequently, the problem can be considered as a typical problem for simulated annealing. The iterative application of this technique consists in updating the image by modifying the intensity of all the pixels in turn, in a pre-specified order. This procedure leads to a significant consumption of computing time: indeed, the number of complete sweepings of the image necessary to obtain a good restoration is, typically, about 300 to 1000. But as the calculation of the energy variation is purely local in nature, several methods were proposed to update the image by simultaneously treating a large number of pixels, using specialized elementary processors. The formalism of the Markovian fields made it possible to treat, by simulated annealing, several crucial tasks in automated analysis of the images: the restoration of scrambled images, the image segmentation, the image identification ... Apart from this formalism, other problems in the image processing domain were also solved by annealing: for example, the method was utilized to determine the geological structure of the basement, starting from results of seismic experiments.

To finish, we will mention some specific problems, in very diverse fields, where simulated annealing was employed successfully: organization of the data-processing network for the French Loto (it required ten thousand playing machines to be connected to host computers), optimization of the collection of the household garbage in Grenoble, timetable problems (the problem was, for example, to determine the optimal planning of the rest days in a hospital), optimization in architecture (in a project on constructing a 17 floor building for an insurance company, it was necessary to distribute the activities among the various parts, so that the work output from 2000 employees can be maximized)... Several applications of simulated annealing for the scheduling problems can be found (particularly, in the references [Van Laarhoven et al., 1992, Brandimarte, 1992, Musser et al., 1993, Jeffcoat and Bulfin, 1993]). The adequacy of the method for this type of problem has been discussed. For instance, Lenstra et al. [Van Laarhoven et al., 1992] showed that the computing time involved was unsatisfactory. Moreover, in [Fleury, 1995], Fleury underlines several characteristics of the scheduling problems which make them unsuitable for simulated annealing and he recommends a different stochastic method, inspired by simulated annealing and tabu search: the "kangaroo method", for this problem.

1.6 Advantages and disadvantages of the method

From the preceding discussion, the principal characteristics of the method can be established. Firstly, the *advantages*: it is observed that the simulated annealing technique generally achieves a good quality solution (i.e. absolute minimum or good relative minimum for the objective function). Moreover, it is a general method: it is applicable, and easy to implement, for all the problems which can potentially employ the iterative optimization techniques, under the condition that after each transformation the corresponding change in the objective function can be evaluated directly and quickly (often the computing time becomes excessive if complete re-computation of the objective function cannot be avoided, after each transformation). Lastly, it offers great flexibility, as one can add new constraints easily afterwards in the program.

Now, let us discuss the disadvantages. The users are sometimes repelled by the involvement of great many parameters (initial temperature, rate of decrease of the temperature, length of the temperature stages, termination criterion for the program...). Although the standard values published for these parameters generally allow an effective operation of the method, the essential empirical nature of them can never guarantee suitability for a large variety of problems. The second defect of the method — which depends on the preceding one — is the computing time involved, which is excessive in certain applications.

In order to reduce this computing time, it still requires an extensive research effort to determine the best values of the parameters of the method [Siarry, 1994], particularly for the law of decrease of the temperature. Any progress in the effectiveness of the technique and the computing time involved can be obtained by continuing the analysis of the method in three specific directions: utilization of interactive parameter setting, parallelization of the algorithm and incorporation of statistical physics based approaches to analyze and study disordered mediums.

1.7 Simple practical suggestions for the beginners

- *Definition of the objective function* : some constraints are integrated here, others constitute a limitation in allowed disturbances for the problem.
- *Choice of the disturbance mechanisms* for a "current configuration": the calculation of the corresponding ΔE variation of the objective function must be *direct* and rapid.
- *Initial temperature* T_0: it may be calculated as a preliminary step using the following algorithm:
  - initiate 100 disturbances at random; evaluate the average $\langle \Delta E \rangle$ of the corresponding ΔE variations;

  - choose an initial rate of acceptance τ_0 of the "degrading perturbations", according to the assumed "quality" of the initial configuration; for example:
    - "poor" quality: $\tau_0 = 50\%$ (starting at high temperature),
    - "good" quality: $\tau_0 = 20\%$ (starting at low temperature),
  - deduce T_0 from the relation: $e^{\frac{-\langle \Delta E \rangle}{T_0}} = \tau_0$.
- *Acceptance rule of Metropolis.* it is practically utilized in the following manner: if $\Delta E > 0$, a number r in $[0, 1]$ is drawn randomly, and accept the disturbance if $r < e^{\frac{-\Delta E}{T}}$, where T indicates the current temperature.
- *Change in temperature stage*: can take place as soon as one of the 2 following conditions is satisfied during the temperature stages:
  - $12 \cdot N$ perturbations accepted;
  - $100 \cdot N$ perturbations attempted,
    N indicating the number of degrees of freedom (or parameters) of the problem
- *Decrease of the temperature*: can be carried out according to the geometrical law: $T_{k+1} = 0.9 \cdot T_k$.
- *Program termination*: can be activated after 3 successive temperature stages without any acceptance.
- *Essential verifications during the first executions of the algorithm*:
  - the generation of the real random numbers (in $[0, 1]$) must be well *uniform;*
  - the "quality" of the result should not vary significantly when the algorithm is executed *several times*:
    - with different "seeds" for the generation of the random numbers,
    - with different initial configurations,
  - for each initial configuration used, the result of simulated annealing can be favorably compared, theoretically, with that of the *quenching* ("disconnected" Metropolis rule).
- *An alternative for the algorithm in order to achieve less computation time*: simulated annealing is greedy and not very effective at low temperature; hence the interest may lie in utilizing the simulated annealing technique, prematurely terminated, in cascade with an algorithm of local type, for specific optimization of the problem, of which role is "to refine" the optimum.

1.8 Annotated bibliography

[Siarry and Dreyfus, 1989]: This book describes the principal theoretical approaches and the applications of the simulated annealing in the early years of formation of the method (1982-1988), when the majority of the theoretical bases were established.

[Reeves, 1995]: The principal metaheuristics are explained in great detail in this work. An elaborate presentation of simulated annealing is proposed in the chapter 2. Some applications are presented: in particular, design of electronic circuits and treatment of the scheduling problems.

[Saït and Youssef, 1999]: In this book several metaheuristics have been extensively explained, which includes simulated annealing (in chapter 2). The theoretical elements relating to the convergence of the method are clearly put in detail. The book comprises also the study of an application in an industrial context (that of the TimberWolf software, which is a reference tool for the layout-routing problem). This should be cited as an invaluable contribution for the teachers: each chapter is supplemented by suitable exercises.

[Pham and Karaboga, 2000]: The principal metaheuristics are also explained in this book. Here, chapter 4 is completely devoted to simulated annealing which concludes with an application in the field of the industrial production.

[Teghem and Pirlot, 2002]: This recent book is a collection of the contributions of a dozen authors. Simulated annealing is however not treated in detail.