SIMULATED ANNEALING FOR MULTI OBJECTIVE OPTIMIZATION PROBLEMS

Paolo Serafini
Department of Mathematics and Computer Science, Via delle Scienze 206, University of Udine, 33100 Udine, Italy.

1. INTRODUCTION

In the last decade some large scale combinatorial optimization problems have been tackled by way of a stochastic technique called 'simulated annealing' first proposed by Kirkpatrick et al. (1983). This technique has proved to be a valid tool to find acceptable solutions for problems whose size makes impossible any exact solution method.

The simulated annealing technique lends itself to a setting with multiple objectives so that the decision maker is eventually offered a large set of nondominated solutions. Of course, since the method is heuristic only suboptimal solutions may be found. The larger the number of iterations are allowed the higher the chance will be of finding 'true' non dominated solutions.

The idea is to control the iteration so that the nondominated solutions have higher stationary probability. In turn, among the nondominated solutions, some of them could have higher stationary probability according to the preferences of the decision maker which could be stated either by means of possibly varying weights or by some domination structure.

In this paper we examine several alternative criteria for the probability of accepting a new solution. We shall see that a special rule given by the combination of different criteria makes the probability distribution to be concentrated almost exclusively on the set of nondominated solutions.

2. MATHEMATICAL BACKGROUND

We first give a brief account of how the simulated annealing technique works for single objective combinatorial problems. Let $f : X \to R$ be a function to be minimized over X, where X is a finite (but very large) set. To each element $x \in X$ a neighbourhood $N(x) \subset X$ is associated. Typically $N(x)$ is very small with respect to X.

Iterations can be defined by first choosing a starting point x and then repetitively selecting $y \in N(x)$ and assigning $x := y$. Local search methods select y so that $f(y) < f(x)$. In case there is no such y in $N(x)$ the local search stops yielding a local minimum x.

The simulated annealing technique differs from a pure local search by letting the choice of y be governed by the following stochastic rule: first $y \in N(x)$ is chosen

with probability q_{xy}, then y is accepted (i.e. the assignment $x := y$ is executed) with probability

$$p_{xy}(T) := \min\left\{1; e^{(f(x)-f(y))/T}\right\} \tag{1}$$

where T is a parameter called temperature. Clearly for $T = 0$ only improving choices are accepted and the method reduces to a pure local search. On the opposite side, for very large values of T, any y chosen in the neighbourhood is accepted, no matter how worse it is with respect to x. Any value $T > 0$ allows the iteration to escape from a local minimum sooner or later.

In order to understand the properties of the iteration defined above it is useful to model it as a Markov chain with state space X and transition probabilities $p_{xy}q_{xy}$. If q_{xy} is symmetric and the neighbourhood structure is such that each state can be reached by any other state (i.e. the transition matrix is irreducible) the equilibrium vector $\pi_x(T)$ can be computed as:

$$\pi_x(T) = Ke^{-f(x)/T} = \frac{e^{-f(x)/T}}{\sum_y e^{-f(y)/T}} = \frac{e^{(f(x^*)-f(x)/T}}{1+\sum_{y\neq x^*} e^{(f(x^*)-f(y))/T}} \tag{2}$$

where x^* is the global minimum of the problem. From (2) it is seen that the global optimum has the highest probability and that $\pi_{x^*}(T) \to 1$ as $T \to 0$ and $\pi_x(T) \to 0$ as $T \to 0$, for $x \neq x^*$. Unfortunately for $T = 0$ the transition matrix is no longer irreducible, so that the behaviour of the Markov chain for $T > 0$ is quite different from the one for $T = 0$. Furthermore the second largest eigenvalue of the transition matrix, which is responsible for the speed of convergence toward the stationary probability vector tends to 1 as $T \to 0$. These facts suggest controlling the Markov chain by decreasing T at decreasing speed. The way T is varied during the iteration is called 'annealing schedule'.

For a more comprehensive understanding of simulated annealing the reader is referred to the literature, for instance, Černy (1985), Kirkpatrick and Swendsen (1985), Lundy and Mees (1986), Laarhoven and Aarts (1987).

The problem we are dealing with in this paper is concerned with the minimization in a multi objective sense of m functions $f_1(x), f_2(x), \ldots, f_m(x)$ over the finite set X. The m objective functions define a preference structure over X. A preference structure (Yu (1989)) is a partition of $X \times X$ based on the binary relations $\{\prec, \succ, \sim, ?\}$ such that for any ordered pair $(x, y) \in X \times X$ exactly one of the following alternatives holds:

$$i)\ x \prec y, \quad ii)\ x \succ y, \quad iii)\ x \sim y, \quad iv)\ x\,?\,y.$$

Here $x \prec y$ means that x is preferred to y and holds if and only if $y \succ x$, $x \sim y$ that x and y are indifferent, and $x\,?\,y$ that no preference can be stated between x and y. Optimal points are those x such that there does not exist any y such that $y \prec x$. They are also called non dominated points.

Simple ways to deduce a preference structure from the objective functions are for instance the followings:

i) *Scalar ordering*

$$\begin{aligned} x \prec y &\iff F(f(x)) < F(f(y)), \\ x \sim y &\iff F(f(x)) = F(f(y)). \end{aligned} \tag{3}$$

with $F : R^m \to R$ a suitable scalar function like for instance $F(f) = \sum_i w_i f_i$, with nonnegative weights w_i (in this case the scalarization is called 'convex combination') or, alternatively $F(f) = \max_i w_i(f_i - r_i)$, with nonnegative weights w_i and reference points r_i (this scalarization is also called 'Čebišev norm'). Both scalarizations have been extensively dealt with in the literature.

ii) *Pareto ordering*

$$\begin{aligned} x \prec y &\iff f(x) \leq f(y) \text{ and } f(x) \neq f(y), \\ x \sim y &\iff f(x) = f(y), \\ x \,?\, y &\iff \exists i, j : f_i(x) < f_i(y),\ f_j(y) < f_j(x). \end{aligned} \tag{4}$$

iii) *Cone ordering*

$$\begin{aligned} x \prec y &\iff f(x) - f(y) \in C \setminus \{0\}, \\ x \sim y &\iff f(x) = f(y), \\ x \,?\, y &\iff f(x) - f(y) \notin C \cup (-C). \end{aligned} \tag{5}$$

with C a suitable cone. In case $C = R^m_-$ we have the Pareto ordering. In case C is a halfspace with normal vector w we have a scalar ordering with $F = wf$.

3. RULES FOR TRANSITION PROBABILITIES

Two alternative approaches can be taken by considering which points in $N(x)$ should be accepted with probability 1. In one approach we may consider the criterion that only dominating point should be accepted with probability 1. This approach may be called *strong criterion*. In the other approach (*weak criterion*) we may reverse this attitude by deciding that only dominated points are accepted with probability strictly less than 1. The two approaches lead to quite different Markov chains. As we shall see a combination of the two approaches can provide good results. Throughout the paper we shall assume that $y \in N(x) \iff x \in N(y)$, $q_{xy} > 0 \iff y \in N(x)$, $q_{xy} = q_{yx}$ and $N(x)$ is sufficiently large so that the transition matrix is irreducible.

3.1 Scalar ordering

For scalar functions we have the standard acceptance criterion which can be stated as

Rule SL :

$$p_{xy}(T) := \min\left\{1\,;\, e^{\sum_i w_i(f_i(x) - f_i(y))/T}\right\}$$

if $F(f) = wf$ ('SL' for Scalar Linear), or, alternatively, as

Rule SC :

$$p_{xy}(T) := \min\left\{1\,;\; e^{(\max_i w_i(f_i(x)-r_i)-\max_i w_i(f_i(y)-r_i))/T}\right\}$$

if $F(f) = \max_i w_i(f_i - r_i)$ ('SC' for Scalar Čebišev).

In these two cases the results of the previous section can be applied. The stationary probability vector is the one with highest probability for the minimum of the corresponding scalar function. Of course, for the scalar ordering the strong criterion coincides with the weak criterion.

3.2 Pareto ordering

Let us first consider the strong acceptance criterion. In order to compute the transition probability we introduce a quantitative criterion simply based on the objective functions. One possibility is given by the following:

Rule SP :

$$p_{xy}(T) := \prod_{i=1}^{m} \min\left\{1\,;\; e^{(f_i(x)-f_i(y))/T}\right\}$$

Rule SP ('SP' for Simple Product) is clearly 'local'. Given a point x we try to improve the situation with respect to x. Also it is separable in the various objectives. Nonetheless the computation of the stationary probability indicates a result which is perhaps counterintuitive. The probability stationary vector for Rule P is given by

$$\pi_x = K \prod_{i=1}^{m} e^{-f_i(x)/T} = K\, e^{-\sum_i f_i(x)/T}$$

as can be seen by verifying that the detailed balance equations $\pi_x\, q_{xy}\, p_{xy} = \pi_y\, q_{yx}\, p_{yx}$ are satisfied.

This result shows that the Markov chain given by Rule SP has the same stationary distribution as the one given by Rule SL with unit weights. This is rather counterintuitive, since the local rule of transition does not make any trade-off between different objectives. Nonetheless the stationary distribution looks like if a scalar minimization is being performed.

Note that the Markov chain given by Rule SP could be controlled by using different temperatures, one for each objective. It is therefore quite clear that there is an equivalence between possible weights and different temperatures for the objective functions.

In other words Rule SP could be modified into:

Rule P :

$$p_{xy}(T) := \prod_{i=1}^{m} \min\left\{1\,;\; e^{w_i(f_i(x)-f_i(y))/T}\right\}$$

('P' for Product) where the quantity T/w_i plays the role of a temperature T_i. This is like having a different annealing schedule for each objective function. An immediate application of this idea consists in varying slowly the weights, while keeping the temperature at a very low level, so to have the possibility to explore the entire Pareto set.

We may introduce a new rule based on the local version of the Čebyšev norm scalarization, that is we take the current values $f_i(x)$ as the reference point, so that we are led to the following rule

Rule C :

$$p_{xy}(T) := \min\left\{1\,;\; \min_{i=1,\ldots,m}\left\{e^{w_i(f_i(x)-f_i(y))/T}\right\}\right\}$$

Rule C ('C' for Čebišev) has transition probabilities not less (and most of time larger) than the ones of Rule P. As $T \to 0$ the transition matrices tend to the same matrix . For Rule C it is not possible to derive an analytical expression for the stationary probability.

With the weak criterion any solution which is not strictly dominated is accepted with probability 1. For continuity reasons solutions which are dominated in some objectives and are indifferent in other objectives must also be accepted with probability 1. So the only practical possibility is the following transition rule ('W' for Weak):

Rule W :

$$p_{xy}(T) = \min\left\{1\,;\; \max_{i=1,\ldots,m}\left\{e^{w_i(f_i(x)-f_i(y))/T}\right\}\right\}$$

For $T = 0$ transitions are not allowed only from a point to another point dominated by it. This 'permissive' behaviour of the Markov chain makes it possible in most cases to have an irreducible chain even for $T = 0$. Therefore the stationary vector for $T > 0$ tends to the stationary vector for $T = 0$ as $T \to 0$. It is not possible to derive an analytical expression for the stationary vector given by Rule W.

3.3 Cone ordering

In case we have a cone ordering based on a polyhedral cone C with polar cone $C^* := \{u : uv \geq 0 \quad \forall v \in C\}$ and generators c^j, $j = 1, \ldots, h$, of C^* we may define functions $\tilde{f}_j(x) := c^j\, f(x)$, $j = 1, \ldots, h$, and apply the rules previously defined to the new functions:

Rule CP :

$$p_{xy}(T) := \prod_{j=1}^{h} \min\left\{1\,;\, e^{(\tilde{f}_j(x)-\tilde{f}_j(y))/T}\right\} = \prod_{j=1}^{h} \min\left\{1\,;\, e^{c^j(f(x)-f(y))/T}\right\}$$

('CP' for Cone Product). The stationary probability vector of this chain is

$$\pi_x = K\, e^{-\sum_j \tilde{f}_j(x)/T}$$

so that the point minimizing

$$\sum_j c^j f(x) = \sum_{ij} c_i^j f_i(x) = \sum_i (\sum_j c_i^j) f_i(x)$$

is the one with highest probability and therefore Rule CP has the same stationary vector as Rule P and Rule SL with $w_i = \sum_j c_i^j$. The transition matrices are however different.

Similarly we may define the following

Rule CC :

$$p_{xy} := \min\left\{1\,;\, \min_{j=1,\ldots,h}\left\{e^{c^j(f(x)-f(y))/T}\right\}\right\}$$

Rule CW :

$$p_{xy} := \min\left\{1\,;\, \max_{j=1,\ldots,h}\left\{e^{c^j(f(x)-f(y))/T}\right\}\right\}$$

which have different stationary vectors than the corresponding Rule C and Rule W. All these rules boil down to the minimization of $wf(x)$ whenever $C^* = \{\alpha w : \alpha \geq 0\}$.

4. A COMPOSITE RULE

By examining Rule P (or Rule C) we see that essentially one solution gets the highest probability in the steady state. All other solutions, including Pareto optima, have a very low probability. This is not a desirable property of the Markov chain since we would like the Pareto set to have a relevant probability over all other points.

On the other hand we see that Rule W does give a prominent role to the Pareto set, since these are the solutions to which the chain most often goes. However, also other solutions which are not Pareto optima have a rather high probability, being, so to speak 'transit' points for the Pareto set. Again this is not an entirely satisfactory behaviour.

By combining together Rule P and Rule W as

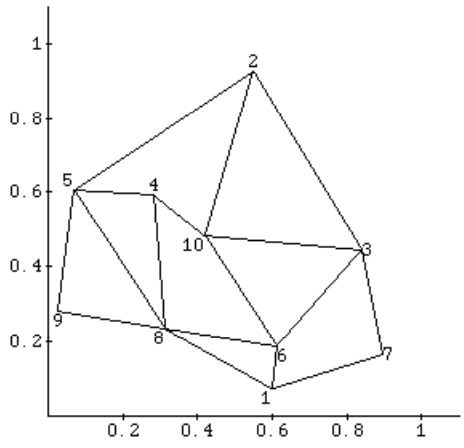

Figure 1

Rule M :

$$p_{xy}(T) := \alpha \prod_{i=1}^{m} \min\left\{1; e^{w_i(f_i(x)-f_i(y))/T}\right\} + (1-\alpha) \min\left\{1; \max_{i=1,\ldots,m}\left\{e^{w_i(f_i(x)-f_i(y))/T}\right\}\right\}$$

with $0 < \alpha < 1$ ('M' for Mixed), we may retain both the advantages of Rule P and Rule W. Furthermore the transition matrix of Rule M remains irreducible for $T = 0$ as long as Rule W does.

5. TWO EXAMPLES

In order to appreciate the effect of choosing one of the above rules a small example has been provided. In Figure 1 ten points are displayed in a bidimensional objective space. The segments connecting the points represent the neighbourhood structure. The probabilities q_{xy} are taken as $1/d_x$ with d_x the number of points adjacent to x. The points 1, 8 and 9 are Pareto optima, and the points 4 and 10 are local Pareto optima. For such an example it is easy to compute the transition matrix and the relative stationary probabilities. In Table 1 and 2 the stationary probabilities π_x are reported for temperatures ranging from $T = 10$ to $T = 0.01$ for Rules P,C and W with unit weights. Furthermore the values for $T = 0$ are also given for Rule W. Note that Rule P and Rule C give almost identical results. For small values of T the point 9 which minimizes the sum of the coordinates is clearly the highest probability solution.

By contrast note the quite different behaviour of Rule W. As $T \to 0$ the most likely solutions are the Pareto optima. Slightly less likely are other solutions which apparently constitute transit points between the optima or are themselves local Pareto optima. Note the positive effect of Rule M with parameters $\alpha = 0.9$ and $\alpha = 0.99$ (i.e. by giving a predominant role to Rule P) displayed in Table 3.

As a second example we have applied the simulated annealing technique to a travelling salesman problem (TSP) with two objectives. We recall that the TSP is defined

Table 1

	T=10			T=1		
	Rule P	Rule C	Rule W	Rule P	Rule C	Rule W
π_1	0.0898	0.0894	0.0886	0.101	0.0984	0.0919
π_2	0.0829	0.0843	0.0867	0.0454	0.0552	0.0745
π_3	0.112	0.113	0.116	0.0735	0.0832	0.105
π_4	0.0880	0.0877	0.0885	0.0829	0.0814	0.0912
π_5	0.119	0.118	0.118	0.134	0.129	0.125
π_6	0.118	0.117	0.118	0.119	0.115	0.122
π_7	0.0576	0.0580	0.0583	0.0458	0.0500	0.0543
π_8	0.151	0.150	0.147	0.191	0.184	0.154
π_9	0.0621	0.0616	0.0593	0.0976	0.0917	0.0637
π_{10}	0.117	0.117	0.117	0.107	0.110	0.115

Table 2

	T=0.1			T=0.01			T=0
	Rule P	Rule C	Rule W	Rule P	Rule C	Rule W	Rule W
π_1	0.0283	0.0270	0.117	$4.23\ 10^{-13}$	$1.35\ 10^{-12}$	0.186	0.217
π_2	$9.22\ 10^{-6}$	$1.90\ 10^{-4}$	0.0225	$8.14\ 10^{-20}$	2.0810^{-20}	0.00741	0.00672
π_3	$8.51\ 10^{-5}$	$4.84\ 10^{-4}$	0.0536	$1.51\ 10^{-21}$	3.1610^{-20}	0.0299	0.0268
π_4	0.00381	0.00486	0.100	$3.12\ 10^{-21}$	7.9510^{-19}	0.0933	0.0919
π_5	0.0368	0.0537	0.141	$1.61\ 10^{-16}$	1.4410^{-14}	0.104	0.103
π_6	0.0105	0.0111	0.145	$1.98\ 10^{-18}$	2.2110^{-17}	0.119	0.0970
π_7	$3.88\ 10^{-4}$	$9.93\ 10^{-4}$	0.0292	$2.62\ 10^{-20}$	4.6510^{-20}	0.00748	0.00672
π_8	0.166	0.167	0.189	$7.79\ 10^{-11}$	7.7910^{-11}	0.218	0.223
π_9	0.749	0.728	0.0893	1.	1.	0.137	0.141
π_{10}	0.00377	0.00545	0.110	$3.12\ 10^{-15}$	1.9010^{-14}	0.0947	0.085

Table 3

	Rule M							
	$\alpha = 0.9$				$\alpha = 0.99$			
	T=1	T=0.1	T=0.01	T=0	T=1	T=0.1	T=0.01	T=0
π_1	0.100	0.0855	0.318	0.338	0.101	0.0351	0.316	0.368
π_2	0.0483	$4.35\ 10^{-4}$	$7.61\ 10^{-5}$	$1.11\ 10^{-4}$	0.0457	$2.42\ 10^{-5}$	$1.18\ 10^{-7}$	$1.05\ 10^{-6}$
π_3	0.0769	0.00176	0.00213	0.00313	0.0738	$1.48\ 10^{-4}$	3.1710^{-5}	$2.81\ 10^{-4}$
π_4	0.0841	0.0203	0.0369	0.0695	0.0831	0.00525	0.00509	0.0629
π_5	0.134	0.0946	0.0322	0.0255	0.134	0.0447	0.00389	0.00297
π_6	0.119	0.0406	0.0306	0.0236	0.119	0.0135	0.00311	0.00271
π_7	0.0469	0.00337	$1.02\ 10^{-4}$	$1.42\ 10^{-4}$	0.0459	$5.72\ 10^{-4}$	$4.51\ 10^{-7}$	$1.39\ 10^{-6}$
π_8	0.187	0.249	0.273	0.265	0.191	0.182	0.265	0.274
π_9	0.0931	0.482	0.277	0.233	0.0972	0.712	0.402	0.258
π_{10}	0.108	0.0212	0.0277	0.0403	0.107	0.00523	0.00333	0.0290

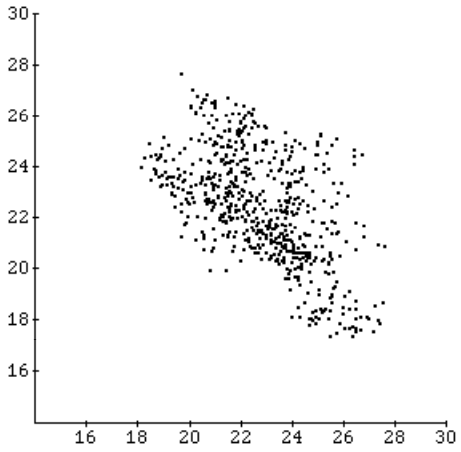

Figure 2

by assigning intercity distances d_{ij} for a set of n cities and asking for a tour which passes through each city exactly once and has minimal length. This is a notoriously difficult problem (see Lawler et al. 1985). Whereas exact solutions can be obtained for problems up to 1000 cities (Grötschel and Holland (1991)) and exact solutions are reported for even larger instances (2393 cities in Padberg, Rinaldi (1990)), the only practical approach for problems of more than 10,000 cities are heuristic methods among which the simulated annealing technique.

Although the TSP is a typical test problem for new techniques (and this is what we are also doing for multi objective problems), one should be aware of the fact that in the TSP feasible solutions are quite easily generated and this is not always the case for general problems. Finding a feasible solution can be a NP-hard problem by itself.

Here we use two sets of intercity distances d^1_{ij}, d^2_{ij} in order to generate a biobjective problem. In our test problem two random numbers a, b between 0 and 1 are generated and then they are used to produce distances in the following way

$$d^1_{ij} := a + \beta(b - 0.5) \qquad d^2_{ij} := b + \beta(a - 0.5)$$

with β a parameter expressing the correlation between the objectives. We have set $\beta := -0.2$. The number of cities is 50.

For space reasons it is not possible to report all computational tests. The rule producing the best results has been Rule M with parameter $\alpha = 0.9$. Moreover we have chosen variable weights for the two objectives during the iteration; more exactly starting from the values $w_1 = w_2 = 1$ at each iteration a random number in the range $[-0.05, +0.05]$ was added to each weight so as to have two different weights varying rather slowly. The starting temperature was $T = 10,000$ and the final one $T = 1$. The annealing schedule was chosen as $T = c/\log(k_0 + k)$ with k index of iteration and c, k_0 parameters set so to have $T = 10,000$ for $k = 0$ and $T = 1$ for $k = 1,000$.

Out of 1,000 iterations 641 solutions have been accepted and 19 of them were non-dominated among the generated solutions. In Figure 2 the 641 solutions are displayed in the objective space. Note that the picture does not reflect properly the stationary

distribution, since repeated solutions (and this happened 359 times) are drawn only once.

6. REFERENCES

Černy, V., 1985, "Thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm", *J. of Optimization Theory and Applications*, **45**, 41-51.

Grötschel, M. and O. Holland, 1991, "Solution of large-scale symmetric travelling salesman problems", *Mathematical Programming*, **51**, 141-202.

Kirkpatrick, S., C.D. Gelatt, Jr., and M.P. Vecchi, 1983, "Optimization by simulated annealing", *Science*, **220**, 671-680.

Kirkpatrick, S. and R. Swendsen, 1985, "Statistical mechanics and disordered systems", *Communications of ACM*, **28**, 363-373.

Laarhoven, P.J.M. van and E.H.L. Aarts, 1987, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht.

Lawler, E.L., J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys eds., 1985, *The Traveling Salesman Problem: a Guided Tour of Combinatorial Optimization*, Wiley, Chichester.

Lundy, M. and A. Mees, 1986, "Convergence of an annealing algorithm", *Mathematical Programming*, **34**, 111-124.

Padberg, M. and G. Rinaldi, 1990, "Facet identification for the symmetric travelling salesman problem", *Mathematical Programming*, **47**, 219-257.

Yu, P.L., 1989, "Multiple criteria decision kaking: five basic concepts", in in *Handbooks in OR & MS, Vol 1: Optimization*, G.L. Nemhauser et al., Eds., North Holland, Amsterdam.