ARTICLES

Simulated Annealing — New Developments in Combinatorial Optimization

John A. Kinsella

Introduction

Combinatorial Optimization (CO) may be defined as the problem of maximizing (or minimizing) some measure of utility (the objective function) of a system with a large but finite number of states. There may be constraints which limit the domain which must be searched for the optimal solution. Perhaps the best known example is Integer Linear Programming, optimization of a linear function of a set of independent variables subject to a set of linear constraints and the requirement that the solution vector have integer components. Special techniques have been developed for the solution of particular CO problems, such as the above example, but there are many important problems (timetabling, scheduling) for which an exact solution in a reasonable time is not always possible. More precisely, there is a large class of CO problems which are known as NP-complete. NP-completeness means that if any algorithm exists which solves a problem in a time which grows as a polynomial in N (a measure of the size of the problem), then all of the set are also soluble in polynomial time. Examples of NP-complete problems include Integer Linear programming, partitioning a set of integers into 2 sets whose sums are equal and the well-known Travelling Salesman problem (TSP) (find the shortest tour or Hamiltonian circuit for a set of coplanar points). The significance of the idea of NP-completeness is that many important problems are NP-complete and therefore, in a sense, equally difficult. In particular, all known algorithms for the TSP run in a time that grows as an exponential in the number of cities in the tour. For this reason, rather than from any

intrinsic interest, the TSP is often used as a benchmark for comparing heuristics, techniques for the approximate solution of CO problems. An important class of heuristics are known collectively as "iterative improvement" methods. The heuristic begins with the system in a particular, often randomly chosen, state. A rearrangement operation (such as 2-OPT for the TSP [1]) is applied randomly until a new state which reduces the objective function is found. The new configuration is adopted and the rearrangements continue until no further improvements can be achieved. Often the heuristic will get 'stuck' in a local, rather than the global, minimum so it is neccessary to use several different initial states and adopt the best final result. In fact, the problem of sticking in a local minimum is not confined to CO, gradient descent and Newton-type methods used in continuous optimization suffer from the same defect.

Simulated Annealing

A more systematic approach to this problem was proposed by S. Kirkpatrick et al.[2] in 1983 based on the Metropolis algorithm. N. Metropolis et al. [3], in 1953, proposed an simple algorithm for the efficient simulation of a collection of atoms in equilibrium at a given temperature. The following exposition of the Metropolis algorithm follows that of S. Geman and D. Geman [4]. Let Ω denote the possible configurations of the system in question; for example ω in Ω could be the molecular positions. If the system is in thermal equilibrium with its surroundings, then the probability of ω is given by

$$P(\omega) = \frac{e^{-\beta E(\omega)}}{\sum_{\omega} e^{-\beta E(\omega)}}, \qquad \omega \in \Omega \tag{1}$$

where $E(\omega)$ is the energy of the configuration and $\beta = 1/kT$ where k is Boltzmann's constant and T is temperature in degrees Kelvin. The quantities to be calculated are usually ensemble averages of the form

$$\langle Y \rangle = \int_{\Omega} Y(\omega)\, d\pi(\omega) = \frac{\sum_{\omega} Y(\omega) e^{-\beta E(\omega)}}{\sum_{\omega} e^{-\beta E(\omega)}} \tag{2}$$

where Y is some variable of interest. This expression is analytically intractable. In the standard Monte Carlo approach, one restricts the sums above to a sample of ω's drawn uniformly from Ω. This approach fails in the present case due to the exponential factor, as most of the mass of the distribution is concentrated in a very small part of Ω. In other words, for satisfactory

accuracy, excessively large samples are needed. The technique introduced in [3] was to chose the samples from P instead of uniformly and then weight the samples uniformly instead of by dP. In other words, one obtains $\omega_1, \ldots, \omega_R$ from P and the ensemble average for Y is approximated by

$$\langle Y \rangle \approx \frac{1}{R} \sum_{r=1}^{R} Y(\omega_r) \tag{3}$$

The sampling algorithm in [3] can be summarized as follows. Given the state of the system at time t, say $X(t)$, one randomly choses another configuration X' and computes the energy change $\Delta E = E(X') - E(X(t))$ and the quantity

$$q = P(X')/P(X(t)) = e^{-\beta \Delta E} \tag{4}$$

If $q > 1$, the move to X' is allowed and $X(t+1) = X'$, while if $q \leq 1$, the transition is made with probability q. Thus, one chooses $0 \leq r \leq 1$ uniformly and sets $X(t+1) = X'$ if $r \leq q$ and $X(t+1) = X(t)$ if $r > q$. Metropolis et al. prove that starting from an arbitrary state, repeated application of this algorithm produces, in the limit of arbitrarily many applications, a sequence of samples from a Boltzmann distribution as stated above. In [2], Kirkpatrick et al. proposed applying the Metropolis algorithm to CO as follows. First select a technique for randomly selecting new states from the current state. For the TSP a widely used technique is Lin's 2-OPT, essentially taking a chain (of a given length) from the current tour and inserting it (possibly reversed in orientation) between two successive points in the tour. Again for TSP, the appropriate 'energy function' is the length of the tour under consideration, for a timetabling problem the energy might be the number of clashes or irreconcilable assignments. The transformation rule (4) is then applied repeatedly until approximate equilibrium is reached at the temperature chosen. The combinatorial system is first 'melted' by being allowed to reach equilibrium at a large value of T. The temperature is then reduced gradually, allowing the system to reach a steady state at each discrete value of T chosen. This decreasing sequence of temperatures is called an annealing schedule by analogy with the slow cooling- annealing- of a melt of a physical substance and the technique itself is called Simulated Annealing (SA) for the same reason. Note that for large values of T (small values of β) new states which increase the energy are likely to be accepted, while for small values of T such uphill moves will be rare. This capacity to escape from local minima is what distinguishes SA from

simple iterative improvement. Physical intuition suggests that cooling must be sufficiently slow to avoid (persisting with the analogy from physics) the formation of a defective crystal or glass, with only locally optimal structures. In fact, it can be proved (see below) that, under certain conditions on the annealing schedule, this procedure will converge to the state corresponding to the minimum of the energy function. In practice, these conditions are not satisfied, but SA still provides 'good' solutions in many cases.

Applications of Simulated Annealing

Much of the early published work on SA has consisted of reports on the results of numerical experimentation and, because of its simplicity and convenience, authors have frequently used TSP to evaluate the technique. Reports on the efficiency of the heuristic vary. C. Skiskim and B. Golden [5] found SA to be inferior to the CCAO procedure and moreover found the performance of SA to be highly sensitive to the details of the annealing schedule. However Skiscim and Golden considered N-city TSP with $N \leq 100$. In a later paper [6] E. Bonomi and J.-L. Lutton found that, for $N \geq 250$, SA outperformed Lin's [1] 2-OPT and the convex hull algorithm. More generally, the technique has proved useful for a wide variety of optimization problems in computer design and other areas; [7,8,9]. I.O. Bohachevsky et al.[10], W. Jeffrey and R. Rosner [11] and others have used Simulated Annealing successfully for optimization of continuous functions of many variables. In work in progress, the present author has applied SA to timetabling and to the Vehicle Routing problem. The latter problem can be posed as a TSP in a natural way with the use of dummy locations corresponding to the vehicular resources. As noted above, SA can be implemented for timetabling using the number of violations of the constraints as the objective function. The major difficulty is choice of a suitable data structure to allow the objective function to be evaluated efficiently. In an influential paper published in 1985, D.H. Ackley, G.E. Hinton and T.E. Sejnowski [12] introduced the idea of the 'Boltzmann machine', a domain-independent learning algorithm which modifies the connection strengths between units of a network in such a way that the whole network developes an internal model which captures the underlying structure of the environment. While space does not permit a full treatment here the following general points may be made. The Boltzmann machine is a neural network or 'parallel distributed processor' as developed initially by J.J. Hop-

field [13,14] and others. The machine is composed of elements called units that are connected by symmetric links. A unit is always either on or off, and it adopts these states as a function of the states of the neighbouring units and the weights on its links to them. A unit being on or off is taken to mean that the system either accepts or rejects some elemental hypothesis about the input data (environment). The weight on a link represents a weak constraint between two hypotheses. A variant of the transformation rule (4) is used to modify the state of the individual units of the network so as to bring the network to equilibrium at a given temperature. As always for SA, the temperature is gradually lowered, resulting in (eventually) convergence to a configuration which minimises the objective function. Here the objective function or energy of a configuration is a measure of the extent to which that combination of hypotheses violates the constraints implicit in the input data. The reader is referred to [12,15,16] for further details.

Theoretical Results

The major contribution to the (very small) body of exact results about SA is due to S. Geman and D.Geman [4] (November 1984).In a paper on Bayesian restoration of noisy 2-D images they proved three significant theorems about SA. Here it will suffice to state the three theorems; A, B and C and to discuss their significance. First, some notation is neccessary. (Some changes have been made from that of [4] in the interests of clarity.) Let the state of the system be specified by a vector $x(t)$ with N components x_s. The state-generation process, (without loss of generality), can be required to alter only one component of the state-vector x per update. Let $\{n_t,\ t = 1, 2, \ldots\}$ be the sequence in which the components of x are chosen for updating. Then $\{X(t),\ t = 0, 1, 2, \ldots\}$ is a random process which describes the evolution of the system being studied, where X is a random vector with components X_s, and the evolution $X(t-1) \to X(t)$ of the system is given by

$$P(X_s(t) = x_s, s = 1, \ldots, N) = \Pi(X_{n_t} = x_{n_t} \mid X_s = x_s, s \neq n_t)\, P(X_s(t-1) = x_s, s \neq n_t) \quad (5)$$

where $\Pi = e^{-\beta U} / \sum e^{-\beta U}$ is the Boltzmann factor correspomding to (4) (U corresponds to the energy E to be minimised.) Let the initial configuration of the system be $X(0)$, i.e. the initial distribution $P(X_s(0) = x_s,\ s = 1, \ldots, N)$ is specified for the range of possible values of x_s.

Theorem A (Relaxation) *Assume that for each s, $1 \leq s \leq N$, the sequence $\{n_t,\ t \geq 1\}$ contains s infinitely often. Then for every starting configuration η and every possible state ω,*

$$\lim_{t \to \infty} P(X(t) = \omega \mid X(0) = \eta) = \Pi(\omega) \quad (6)$$

In other words, the distribution of $X(t)$ converges to Π (the Boltzmann or Gibbs distribution) as $t \to \infty$ regardless of $X(0)$. This is essentially a rewording of the result in Metropolis' paper [3].

Some further notation is needed for Theorem B. Rewrite (5) as

$$P(X_s(t) = x_s,\ s = 1, \ldots, N) = \Pi_{T(t)}(X_{n_t} = x_{n_t} \mid X_s = x_s,\ s \neq n_t)\, P(X_{s(t-1)} = x_s,\ s \neq n_t) \quad (7)$$

to indicate the dependence of Π on T, the temperature. The annealing procedure generates a different random process $\{X(t), t = 1, 2, \ldots\}$ for each successive temperature value such that (6) holds. Let $\Omega_0 = \{\omega \in \Omega : U(\omega) = \min_\Omega U(\omega)\}$, that is, the minimum energy configurations of the system and let Π_0 be the uniform distribution on Ω_0. Finally, define $U^* = \max_\Omega U(\omega)$, $U_* = \min_\Omega U(\omega)$ and $\Delta = U^* - U_*$.

Theorem B (Annealing) *Assume that there exists an integer $\tau \geq N$ such that for every $t = 0, 1, 2, \ldots$ we have $\{s_1, \ldots, s_N\} \subset \{n_{t+1}, n_{t+2}, \ldots, n_{t+\tau}\}$. Let $T(t)$ be any decreasing sequence of temperatures for which*

(a) $T(t) \to 0$ *as* $t \to \infty$;

(b) $T(t) \geq N\Delta / \ln(t)$ *for all* $t \geq t_0$, *for some* $t_0 \geq 2$.

Then for any starting configuration η in Ω and for every ω in Ω,

$$\lim_{t \to \infty} P(X(t) = \omega \mid X(0) = \eta) = \Pi_0(w)\,. \quad (8)$$

The first condition merely requires that the update procedure does not slow to an arbitrarily low frequency as the system evolves, and imposes no limitations in practice. Condition (a) is trivially satisfied by any reasonable annealing schedule. However condition (b) is a major problem. For the image restoration problem studied in [4], for example, of the order of e^{40000} updates would be needed to reach $T = 0.5$. Some points can be made. First, condition (b) is a sufficient condition for convergence, and may not be a neccessary one.

(However, the physical process of annealing requires very slow cooling, especially near the freezing point.) Moreover, the modification of SA mentioned in Section 5 below due to H. Szu [17], will, if successful, greatly improve on the performance predicted by Theorem B.

Theorem C (Ergodicity) *As in Theorem B, assume that there exists an integer $\tau \geq N$ such that for every $t = 0, 1, 2, \ldots$ we have $\{s_1, \ldots, s_N\} \subset \{n_{t+1}, n_{t+2}, \ldots, n_{t+\tau}\}$. Then for every function Y on Ω and for every starting configuration η in Ω, the ergodic hypothesis*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Y(X(t)) = \int_{\Omega} Y(\omega)\, d\pi(\omega) \tag{9}$$

holds with probability one.

The significance of this result is that time averages rather than phase averages, which are computationally intractable, can be used to compute expected values.

Modifications of the Heuristic

Several modifications to the basic SA heuristic have been suggested. J.W. Greene and K.J. Supowit [18], have proposed a 'rejectionless' form of annealing. They observe that, at low temperatures, the run time is quite high because many candidates are rejected before each move to a different state. This follows for two related reasons. First of all, for T small, the exponential transition probability to states of higher energy is very small. Moreover, at low temperatures, the system is likely to be in a state with only a small proportion of the accessible states having lower energy. Greene and Supowit propose the following alternative strategy. Let x_i, $i = 1, \ldots, N$, be the states accessible (in one move) from the current state of the system. Store $w_i = \min 1, q$, $i = 1, \ldots, N$, where q is given by (4). Then choose state x_i with probability $w_i / \sum w_i$, make the change of state and re-calculate the w_i's. The sequence of states generated by this method is probabilistically equivalent to the corresponding sequence generated by SA, if the repetitions of the current state each time a move is rejected are omitted. This can be demonstrated as follows using the notation of [18]. Let α_{xT} be the probability that SA accepts the

chosen move at temperature T, i.e.

$$\alpha_{xT} = \frac{1}{N} \sum w_i \tag{10}$$

Then the probability that SA makes the move from state x to state x_i (say) after some number of rejections is

$$\sum_{k=0}^{\infty} (1 - \alpha_{xT})^k \frac{1}{N} w_i = \frac{w_i}{N\alpha_{xT}} \tag{11}$$

which is just $w_i / \sum w_j$, the probability of choosing x_i under the rejectionless method. The run time per change of state for rejectionless annealing clearly has a value independent of the acceptance ratio, while for SA the value is proportional to the reciprocal of the acceptance ratio. However the overheads in terms of memory requirements and CPU for the rejectionless method are large, so the method is only useful at very low temperatures. In numerical experiments undertaken by the present author and a student, temperatures sufficiently low to warrant the use of rejectionless annealing were never reached. Another variant on standard SA is due to I.O. Bohachevsky et al.[10]. In a recent paper they propose using a modified form for the transition probability q to states of higher energy (4). For problems where the minimum of the objective function Φ is known to be zero (if the value is non-zero just use Φ less the known minimum value as the energy) they suggest setting $q = \exp(-\beta \Phi^g \Delta\Phi)$ where g is a suitably chosen negative number. The purpose of the modification is to ensure that when close to the minimum, the heuristic is unlikely to move a large distance away. No theoretical analysis of this modified SA is offered but numerical experiments (on optimization of continuous functions of two variables) are quoted which suggest the technique might be useful when the value of the global minimum is known. For the more common situation, where the value of the global minimum is not known, the authors suggest an adaptive approach, starting with an estimate of $\Phi_{\min}$ and modifying it as necessary as the search proceeds. It is not clear how effective this proposal is in practice. Perhaps the most significant modification of SA is that proposed by H. Szu [17] in 1986. As noted in Section 4, the result (8) due to Geman and Geman [4] demands an unacceptably slow cooling rate for guaranteed convergence to the optimal solution. Szu suggests an alternative approach which he calls the 'Cauchy machine', in deference to the Boltzmann machine of Ackley et al. [12]. In standard SA the successive states of the system are generated

from a uniform (or, more generally [15], from a Gaussian) distribution. In all cases the distribution is of bounded variance. (The probability of accepting this new state is, of course, given by $\min\{1,q\}$, where q is given in (4).) In his paper, Szu claims that using the Cauchy distribution, which has unbounded variance, a cooling schedule reciprocal in t, rather than $\ln(t)$, can be used. Unfortunately only a rather unsatisfactory sketch proof is quoted and the reader is referred to an (as yet) unpublished paper for a rigourous derivation [19]. (Some numerical results are produced in support of his assertion.) However, if, as seems likely, Szu's result is valid, the consequences for SA are major. An exponentially faster cooling rate will be possible, making the method far more realistic as a general-purpose optimization technique than previously.

Summary

Simulated Annealing, in various guises, has been in existence for five years and has been applied to a steadily widening range of problems. With developments like those quoted in this review, continued interest in the topic seems assured.

References

[1] S. Lin and B.W. Kernighan, *An Effective Heuristic Algorithm for the Travelling Salesman Problem.* Operations Research, 21 (1973), 498-516.

[2] S. Kirkpatrick, C.D. Gelatt, Jr. and M.P. Vecchi, *Optimization by Simulated Annealing.* Science, 220 (1983), 671-680.

[3] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, *Equations of state calculations by fast computing machines.* J. Chem. Phys., vol. 21, 1087-1091, 1953.

[4] S. Geman and D. Geman, *Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images.* IEEE Trans. Patt. Anal. Mach. Int., PAMI-6, No.6, 721-741, 1984.

[5] C.C. Skiskim and B.L. Golden, *Optimization by Simulated Annealing: A Preliminary Computational Study for the TSP.* Proceedings of the Winter Simulation Conference. Arlington, VA USA, (1983), 523-535.

[6] E. Bonomi and J.-L. Lutton, *The N-City Travelling Salesman Problem: Statistical Mechanics and the Metropolis Algorithm.* SIAM Review, 26 (1984),551-568.

[7] F. Romeo, C. Sechen and A. Sangiovanni-Vincentelli, *Simulated Annealing Research at Berkeley.* Proc. Int. Conf. Computer Design, (Port Chester, NY), 652-657, 1984.

[8] M.P. Vecchi and S. Kirkpatrick, *Global Wiring by Simulated Annealing.*, IEEE Trans. Computer-Aided Des., vol. CAD-2, 4, 215-222, 1983.

[9] S.R. White, *Concepts of Scale in Simulated Annealing.*,Proc. Int. Conf. Computer Design, (Port Chester, NY), 1984, 646-651.

[10] I.O. Bohachevsky, M.E. Johnson and M.L. Stein, *Generalized Simulated Annealing for Function Optimization.*, Technometrics, vol.28, no.3, 209-217, 1986.

[11] W. Jeffrey and R. Rosner, *Optimization algorithms: Simulated Annealing and Neural Network Processing.*, The Astrophysical Journal, October 1, 1986.

[12] G.E. Hinton, T.J. Sejnowski and D.H. Ackley, *A learning algorithm for Boltzmann machines.*, Cognitive Science, 9, 147-169, 1985.

[13] J.J. Hopfield, *Neural Networks and physical systems with emergent collective computational abilities.*, Proceedings of the National Academy of Sciences USA, 79, 2554-2558.

[14] J.J. Hopfield, *Neurons with graded response have collective computational properties like those of two-state neurons.*, Proceedings of the National Academy of Sciences USA, 81, 3088-3092, 1984.

[15] G.E. Hinton, T.J. Sejnowsi and D.H. Ackley, *Boltzmann Machines: Constraint satisfaction networks that learn.*, Tech. Rep. No.CMU-CS-84-119, Carnegie-Mellon University, Pittsburgh, PA, USA.

[16] D.E. Rumelhart, J.L.McClelland and the PDP Research Group, *Parallel Distributed Processing*. MIT Press, Cambridge MA, 1986.

[17] H. Szu, *Fast Simulated Annealing*, Proc. of the Second Annual Conf. on Neural Networks for Computing, American Institute of Physics, 1986.

[18] J.W. Greene and K.J. Supowit, *Simulated Annealing without rejected moves*, IEEE Trans. Computer-Aided Des., vol. CAD-5, 1, 215-222, 1986.

[19] R.L. Hartley and H.H. Szu, *Generalized Simulated Annealing*, Optics Letter, (submitted for publication).

Department of Mathematics
National Institute for Higher Education
Limerick

On The Level

D.W. Lewis

We present a survey on the notion of the level of a field and its various generalizations. We describe a lot of results that are attractive from an algebraic viewpoint and also highlight the extremely interesting relations between algebra and topology that have been unearthed in the last decade in connection with the level. We hope to persuade the reader that this is an appealing area of mathematics and that it should be a fruitful area for future research. In Section 1, we look at levels of fields, in Section 2, we deal with commutative rings and the link with topology and in Section 3, we look at the non-commutative situation and generalisations of the idea of level.

1 Fields

Let F be a field. F is said to be *formally real* if -1 is not expressible as a sum of squares in F. If F is not formally real we define the *level* of F, denoted $s(F)$, to be the smallest natural number N such that -1 is a sum of N squares in F (We define $s(F) = \infty$ if F is formally real).

The Artin-Schreier theorem [35, p.227] says that a field F is formally real if and only if F admits an ordering (i.e. $s(F) = \infty$ if and only if F admits an ordering).

We look now at levels of some well-known fields

Example 1 $F = \mathbf{R}$, the real numbers, $s(\mathbf{R}) = \infty$.

Example 2 $F = \mathbf{C}$, the complex numbers, $s(\mathbf{C}) = 1$ since $-1 = i^2$ in $\mathbf{C}$.

Example 3 $F = \mathbf{F}_p$, a finite field with p elements, p an odd prime. It is a fairly easy exercise to show $s(\mathbf{F}_p) = 1$ if $p \equiv 1 \pmod 4$ and $s(\mathbf{F}_p) = 2$ if $p \equiv 3 \pmod 4$.

Example 4 $F = \mathbf{Q}_p$, the field of p-adic numbes. Then $s(f) = 1$ if $p \equiv 1 \pmod 4$, $s(F) = 2$ if $p \equiv 3 \pmod 4$. If F is the field of dyadic numbers then $s(F) = 4$. See [35, p.151] for a proof.