Algorithmica (1991) 6: 346-366

Algorithmica

Analysis of Finite Length Annealing Schedules

Philip N. Strenski[1] and Scott Kirkpatrick[1]

Abstract. By constructing a master equation for the distribution of outcomes from simulated annealing, we are able to characterize this process exactly for arbitrary annealing schedules on extremely small problems. Two sorts of numerical experiments are reported, using this formalism. First, annealing schedules are found which minimize the cut cost of partitioning a highly symmetric weighted graph, using a fixed number of Monte Carlo search steps. The experiments yield some surprising results, which sharpen our understanding of the problems inherent in trying to optimize a stochastic search. For example, optimal annealing schedules are not monotone decreasing in temperature. Second, we construct configuration spaces of random energies and varying connectivity. These are used to compare different annealing schedules which are common in the literature. The experiments also provide an occasion to contrast annealing schedules derived from asymptotic, worst-case bounds on convergence to the global optimum with adaptive schedules which attempt to maintain the system close to equilibrium throughout the annealing process.

Key Words. Simulated annealing, Schedules.

1. Introduction. Since its introduction [1], simulated annealing has become a recognized tool for combinatorial optimization. In particular, the technique has been used successfully for chip placement [2], [3], image processing [4]–[7], and a variety of other applications [8], [9]. This algorithm searches a configuration space using random moves which are accepted or rejected based on their relative cost. The probabilistic criterion (the "Metropolis rule" [1]) for accepting some uphill moves mimics the thermal excitations of atoms and involves a parameter, T, analogous to an external temperature. The "annealing schedule" of these changing temperatures as functions of time must be provided by the user. The final result may be sensitive to the schedule chosen, so specification of an annealing schedule has become an important practical issue in the application of simulated annealing.

Using the theory of Markov processes, numerous authors [4], [10]–[12] have demonstrated the existence of schedules which are guaranteed to converge to the optimal solution in infinite time. These schedules have the form

$$T = \frac{a}{b + \log(t)}, \tag{1}$$

where a, b are positive constants depending on the particular problem and t, the elapsed time, is the interpretation given to the number of search iterations which have taken place. This result applies to any problem for which annealing can be formulated.

[1] IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA.

Received May, 1989; revised December, 1989. Communicated by Alberto Sangiovanni-Vincentelli.

In practical applications, the time (or the amount of computation) available is not infinite. Instead one trades off increasing computational cost in order to obtain better solutions or increased robustness, the probability that a given computer run will lead to an acceptable solution. The geometric sequence

$$T = ab^t, \tag{2}$$

where $a > 0$ and $0 < b < 1$, converges to zero more rapidly than (1) and has been used in applications since the first papers [1]. Although there is little theoretical basis for the form of (2), for the most common classes of optimization problems occurring in computer-aided design this schedule usually gives satisfactory results in limited amounts of time. It is normally employed with a host of subsidiary conditions, such as criteria to determine automatically an initial "melting" temperature, to estimate how many iterations are needed at each temperature, and to stop the annealing process when insufficient improvements are being found.

Derivations of the logarithmic schedule (1) equate the constant, a, to the height of a critical energy barrier separating regions which contains only local optima from the region in which the global optimum is found. Hajek's paper [12] gives a clear example of this argument. On the other hand, the geometric annealing schedule is at least consistent with the observation that many difficult optimization problems possess an implicit hierarchical structure, in which large energy barriers separate large valleys in the solution space and small barriers separate the many small valleys within each large valley [13]–[16]. This second view suggests that an efficient annealing schedule will spread its efforts across a wide range of energy barriers, rather than concentrating on one critical barrier size.

Recently several workers have proposed adaptive annealing schedules, which adjust the rate of decrease of the temperature by measuring the rate of change of the objective function or other, more local, measures of the rate at which the solution space is being searched, in order to cool slowly when important changes are taking place, and more rapidly when little additional improvement is obtained [17]–[20]. The most elaborate adaptive schedule has been introduced by Lam and Delosme [19]. They vary the inverse temperature at step i, β_i, using

$$\beta_{i+1} = \beta_i + \left(\frac{\lambda \beta^{1/2}}{C^{3/2}}\right)\left(\frac{4a_i(1 - a_i)^2}{(2 - a_i)^2}\right), \tag{3}$$

where a_i is the actual fraction of moves accepted at step i, C is the specific heat in equilibrium at this temperature, and λ is a control parameter. Their approach relies upon adjusting the move generation strategy to keep a_i close to 0.44, the value which maximizes its contribution to (3), and knowing or measuring C. The first step is not possible for every problem, and measuring C in the course of an experiment is often very difficult. While the approach is intuitively reasonable, there are several unproven assumptions. Lam and Delosme assume that the system begins in equilibrium at some initial temperature, and that the best schedule keeps the system as close to equilibrium while cooling as possible. However, a system may not start in equilibrium, and it is not clear whether it is better to bring it into

equilibrium quickly at a high temperature or more slowly at a low temperature. Finally, in the late stages of annealing, equilibration may be unnecessary, if the region of solution space that remains accessible has only one minimum.

The purpose of this paper is to test some of the assumptions and intuitions behind these annealing schedules. We explore the behavior of simulated annealing with finite-time resources. In order to get a more precise handle on the character of simulated annealing, we take a different approach than most authors. Rather than use the simple algorithm of simulated annealing to give us information about complicated state spaces, we use trivial state spaces to probe the complex character of simulated annealing.

The central idea in this analysis is the restriction to state spaces of small size ($<$about 200 states) whose structure (energies and connectivity) is known beforehand. It is then a simple matter to calculate exactly many properties of the Markov chains associated with simulated annealing on this state space. We average over all annealing runs with the same temperature schedule. This average then provides a probability density of states after each annealing step. This probability density can then be used to calculate various quantities such as average energy, probability of being in the ground state, or average transition rate.

Note that for real problems the size of the state space is at least exponential in the problem size, so that even a simple problem such as the equal bipartioning of a graph of 20 nodes has a space of nearly 200,000 states. The simulation of such a large problem is outside our computational ability. However, many of the features of larger spaces are found in smaller ones and the exact evaluation of annealing schedules on these smaller cases can be used to test ideas about annealing on their larger counterparts. In particular from the results of these comparisons we can test our assumptions about the efficacy of various schedules on state spaces of differing structure, about the behavior of optimal schedules, and about the importance of thermodynamic quantities such as the specific heat.

The exact evaluation of an annealing schedule as applied to a specific problem is straightforward. Let the number of states be N. The configuration of the system after any number of annealing steps i is represented by a column vector $P(i)$ of length N whose entries are the relative probabilities of finding the system in any given state at that step. The evolution of this vector by one annealing step at temperature T is simply a multiplication by the $N \times N$ transition matrix $A(T)$. The relation $P(i) = A(T(i))P(i-1)$ is conventionally referred to in statistical mechanics as a master equation. The final configuration of states (after L multiplications for a schedule of length L) is evaluated by applying some measure to produce a single number.

The starting vector $P(0)$ is typically the uniform vector P with $P_i = 1/N$ for each state i, since the initial state is usually chosen at random. The transition matrix $A(T)$ is typically specified by

$$(4) \qquad A(T)_{ij} = \begin{cases} 0 & \text{if the states } i, j \text{ are not connected,} \\ \text{MIN}(1, e^{(E_j - E_i)/T})/C & \text{otherwise,} \end{cases}$$

where E_i is the energy of state i and C is the number of connecting states. The final

state measure is typically the average final state energy, which can be found by multiplying the final column vector by the row vector of energies E. For example, an annealing schedule of length L with temperatures $T_1 \cdots T_L$ has a value given by $EA(T_L) \cdots A(T_1)P$ for the usual choices. Note also that since the configuration of relative probabilities is known at each step, it is also possible to measure other quantities which depend on this, such as the specific heat and acceptance ratio, as functions of time.

In the following sections we present a number of experiments involving a variety of small state spaces and parametrized annealing schedules. In the first experiment the bipartition of a particular small graph is the test case and the schedule of temperatures is completely arbitrary. We find the optimal schedule of a particular length for this problem.

In the latter sections we restrict our attention to one-parameter families of annealing schedules applied to state spaces of uniformly distributed random energies. We investigate how the characteristics of each family influence the success of annealing. We adjust the connectivity of the state space while leaving the energy levels fixed. This separates the effects of connectivity from other quantities such as specific heat which depend only on the density of states.

Note that in our analysis we restrict ourselves to static annealing schedules, that is, an average is performed over many annealing runs with the same temperature schedule. This can be contrasted with the idea of adaptive schedules in which the temperature schedule varies from run to run. However, an underlying assumption in most adaptive schemes is that the system is near equilibrium at the current temperature and that quantities such as specific heat measured during the run are near to equilibrium values at the current temperature. Since the equilibrium values are known for our test cases we can test these assumptions and comment on the merits of adaptive schedules, although we cannot analyze them directly.

Finally, we summarize our conclusions and suggest areas of further research.

2. A Partitioning Experiment. The first case we study involves the bipartition of a special complete graph of 2^K nodes motivated and described in more detail in [21]. Although each pair of nodes in the graph is connected the weights on the edges will vary in a hierarchical fashion. In particular, consider a binary tree of depth K whose 2^K leaf nodes correspond with the nodes of the complete graph (Figure 1). The weight of the edge between any two nodes will be α^k where k is the depth of the least common ancestor of the two nodes in the binary tree, and α is a constant greater than one. Thus any two nodes whose least common ancestor is the root have the smallest edge weight of 1, and two nodes with the same parent have the largest edge weight of α^{K-1}. A correspondence with the Rent exponent suggests the restriction $2 \leq \alpha \leq 4$ [21], and we observe this further restriction in this paper as well.

The partitioning problem will be to split this graph into two sets of 2^{K-1} nodes each, such that the weight of the cut edges is minimized. By construction this minimum is clearly obtained by dividing the nodes into the left and right branches

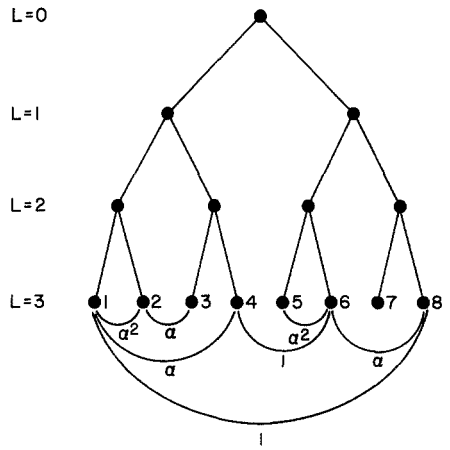

Fig. 1. Hierarchical weights for a graph with eight nodes.

of the tree, since then all cut edges will have the minimum weight of 1. The set of valid partitions will constitute the state space, and the set of pairwise swaps will serve as the move set.

We restrict ourselves to the simple case $K = 3$ involving a graph of eight nodes. A valid partition is an assignment of four nodes to the first partition and the remaining four nodes to the second. A valid move is the interchange of one of the four nodes in the first partition with one of the four in the second. There are thus 70 valid partitions in the state space and 16 possible moves from each state. Because of the symmetries of the problem, the states can be grouped into five equivalence classes. The connectivity and energies for this set of equivalence classes are indicated in Figure 2. Most of the experiments use the value $\alpha = 3$ for simplicity, although some of the theoretical results apply more generally. Note that the system contains local as well as global minima, and that most of the states are at higher energies initially.

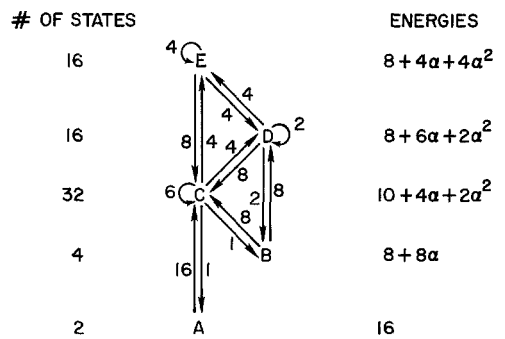

Fig. 2. Transition diagram for equivalence classes for a specific bipartioning problem. Arrows indicate number of moves between states (total 16 per state). Energy and number of states per class are indicated.

We use the average final state energy as the measure of the effectiveness of a given annealing schedule. The final state energy is given by

$$F(T_1, \ldots, T_L) = EA(T_L) \cdots A(T_1)P, \tag{5}$$

where E is the row vector of energies, $A(T)$ is the transition matrix at temperature T, and P is the column vector of initial probabilities (not uniform because of the merging into equivalence classes). We wish to minimize F with respect to the variables $T_1, \ldots, T_L$ subject to the constraints $T_i \geq 0$. Expressed in this fashion the task of determining the optimal schedule of a fixed length for this problem is a simple application of constrained nonlinear optimization. Although $A(T)$ involves some exponentials, it is fairly inexpensive to compute $F(T_1, T_2, \ldots, T_L)$ and its derivatives for use with an optimization program. There are some fine points which are discussed in the next section, but basically this procedure allows us to construct optimal annealing schedules of length $L \leq 70$ for this problem using a standard optimization package [22].

3. Results. The transition matrix for annealing schedules is highly nonlinear in the temperature parameter. It is often useful to scale to another variable of the form $e^{C/T}$ where C is some constant. This change of variables is necessary to make many optimization packages operate correctly. For example, in our case, the transition array (4) may be simplified by rewriting it in terms of the variable $U = \exp[(2 - 2\alpha)/T]$, namely

$$A(U) = \begin{bmatrix} 1 - U^{\alpha+3} & 0 & \frac{1}{16} & 0 & 0 \\ 0 & 1 - \frac{1}{2}U^{\alpha-1} - \frac{1}{2}U^{\alpha} & \frac{1}{16} & \frac{1}{8} & 0 \\ U^{\alpha+3} & \frac{1}{2}U^{\alpha-1} & \frac{7}{8} - \frac{1}{4}U - \frac{1}{4}U^{\alpha+1} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2}U^{\alpha} & \frac{1}{4}U & \frac{3}{8} - \frac{1}{4}U^{\alpha} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4}U^{\alpha+1} & \frac{1}{4}U^{\alpha} & \frac{1}{4} \end{bmatrix} \tag{6}$$

For $\alpha > 2$, the lowest power of U present is 1 (other than constants) so the first derivative of $A(U)$ with respect to U is defined and nonsingular for all $U \geq 0$. This assures that the optimization routine is well behaved. Note that values of $U > 1$ correspond to unphysical negative temperatures. Although these were permitted to the optimization package they were not utilized and may be ignored. Also note that the value $U = 0$ corresponds to the value $T = 0$.

Since the objective function for schedules is quite nonlinear, it is not surprising that there may be more than one locally optimal solution. In fact a schedule of all zeros (iterative improvement) is locally optimal among schedules of any length for this problem (for $\alpha > 2$). This statement is equivalent to showing that, for $1 \leq i \leq L$,

$$\frac{\partial F}{\partial U_i}(0, 0, \ldots, 0) = [EA(0)^{L-i}]A'(0)[A(0)^{i-1}P] > 0. \tag{7}$$

The last bracketed expression is just the vector of state probabilities after $i-1$ steps. From the transition diagram of Figure 2, every state has a nonzero probability of going nowhere at any temperature, and since each state is initially occupied with some probability, this vector has strictly positive entries. In particular the third element is positive. $A'(0)$ is zero except for $A'(0)(3,3) = -\frac{1}{4}$ and $A'(0)(4,3) = \frac{1}{4}$. Therefore if we define $X_p = EA(0)^p$, then (7) reduces to showing that $(X_p(4) - X_p(3))/4 > 0$ for $0 \le p \le L$. Since $A(0)$ has a simple form this can be proved by induction. The recursion formulas are

$$X_{p+1}(1) = X_p(1) = E(1),$$

$$X_{p+1}(2) = X_p(2) = E(2) > E(1),$$

$$X_{p+1}(3) = \frac{14X_p(3) + X_p(2) + X_p(1)}{16},$$

$$X_{p+1}(4) = \frac{3X_p(4) + 4X_p(3) + X_p(2)}{8},$$

so

$$X_{p+1}(4) - X_{p+1}(3) = \frac{6(X_p(4) - X_p(3)) + (X_p(2) - X_p(1))}{16} > 0, \tag{8}$$

where the final inequality follows inductively from $E(4) > E(3)$.

In a similar fashion we can also prove that any locally optimal schedule must terminate with a number of zero temperatures. In particular it is simple to check that if, for some k, $X_k(1) < X_k(3) < X_k(4) < X_k(5)$ and $X_k(2) < X_k(4)$, then the entries of $X_k A'(U)$ are nonnegative for all $U \ge 0$ and the third entry is always positive. It follows that $\partial F/\partial U_{L-k} > 0$ for all $U_{L-k} > 0$ regardless of $U_1, \ldots, U_{L-k-1}$ provided that $U_{L-k+1}, \ldots, U_L$ are all zero. In this case for a locally optimal schedule $U_{L-k} = 0$ is the only possible solution. In particular the elements of $X_0 = E$ are so ordered for all $2 < \alpha < 4$, namely $16 < 10 + 4\alpha + 2\alpha^2 < 8 + 6\alpha + 2\alpha^2 < 8 + 4\alpha + 4\alpha^2$ and $8 + 8\alpha < 8 + 6\alpha + 2\alpha^2$, and thus optimality requires $U_L = 0$. Knowing $U_L = 0$ for a locally optimal schedule and that the elements of $X_1 = EA(0)$ are again ordered for all $2 < \alpha < 4$, shows that for a locally optimal schedule $U_{L-1} = 0$. Continuing this process we can show that at least the last five temperatures must be zero for any locally optimal schedule with $2 < \alpha < 4$. Also $2.4119 < \alpha < 4$ implies at least the last six temperatures are zero, $2.8993 < \alpha < 4$ implies at least the last seven temperatures are zero, and $3.4564 < \alpha < 4$ implies at least the last eight temperatures are zero. Note that these bounds are not tight, weaker conditions may in fact suffice.

For the remainder of the results we fix $\alpha = 3$. The optimal schedules for lengths L from 15 to 70 are shown in Figure 3. For schedules of length less than 14 the only locally optimal schedules consist of all zeros. For $L = 14$ other locally optimal schedules are possible but the all-zero schedule still has the best average final state

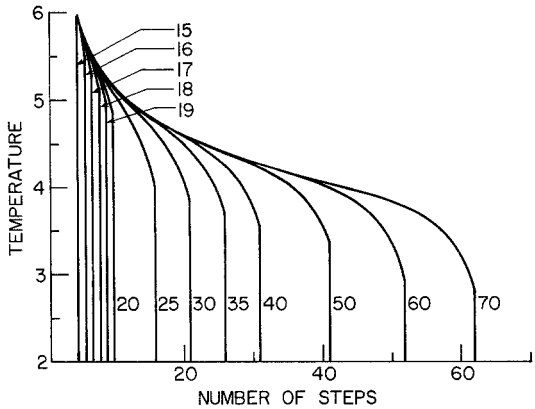

Fig. 3. Optimal annealing schedules for $\alpha = 3.0$ partitioning problem. The total numbers of steps are 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 50, 60, 70. The first four and the final eight to ten temperatures are always zero.

energy. For $L > 14$ there are typically about ten locally optimal schedules, and the one with the best average final state energy is shown in Figure 3. Some reasons for believing we have located all the optimal schedules will be given later.

The most striking feature of these results is that every locally optimal schedule begins with some number of zero temperatures. The global optima all begin with four zero temperatures and local optima begin with from two to five. After this initial segment the schedules closely follow a decreasing envelope which is relatively independent of the length. The envelope is consistent with the theoretical infinite-time schedule (3) but the evidence is certainly not conclusive. However, since the derivation of this schedule is based on tunneling across a single barrier, and this is effectively the situation in our problem, we would expect this result to apply here. When the schedules near the end, the temperature rapidly falls to zero and the last set of temperatures are all zero. There are from eight to ten final zeros for the global optima (recall that for $\alpha = 3$ we proved a lower bound of seven). The energies and standard deviations for locally optimal schedules of various lengths are shown in Figure 4 along with those for iterative improvement (schedule of all zeros). Note that the nonzero schedules of each length have energies which are visually indistinguishable, but all are well below the value for iterative improvement. The probability that the final state is a ground state is presented in Figure 5 for the same conditions. Note that the high probability of finishing in the ground state again supports the conclusion that the system is dominated by single-barrier tunneling for longer schedules.

To aid in the convergence of the optimization, a number of initial and final temperatures were fixed at zero. Then the remaining nonzero temperatures were obtained using a double precision constrained optimization program from the Harwell package [22]. The user provides the function to be evaluated (the average final energy $F(0, \ldots, 0, U_i, \ldots, U_j, 0, \ldots, 0)$) and its derivatives with respect to the variables $U_i, \ldots, U_j$ as well as any linear constraints ($U_k \geq 0$). These functions are

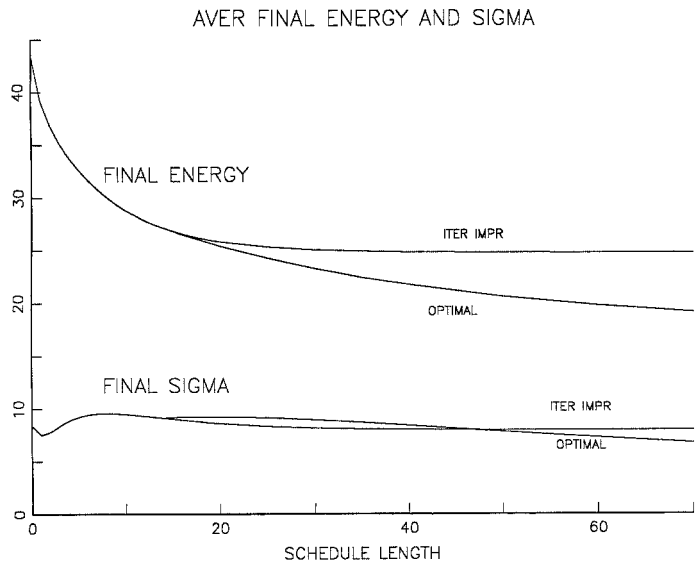

Fig. 4. Average final state energy and standard deviation as a function of length of schedule for locally optimal schedules for the $\alpha = 3$ partitioning problem. Curves are provided for iterative improvement as well as other locally optimal schedules (where they exist) with four initial zeros and eight, nine or ten final zeros. The three curves for nonzero locally optimal schedules are visually indistinguishable but different. The curve for iterative improvement is noticeably poorer for longer schedules.

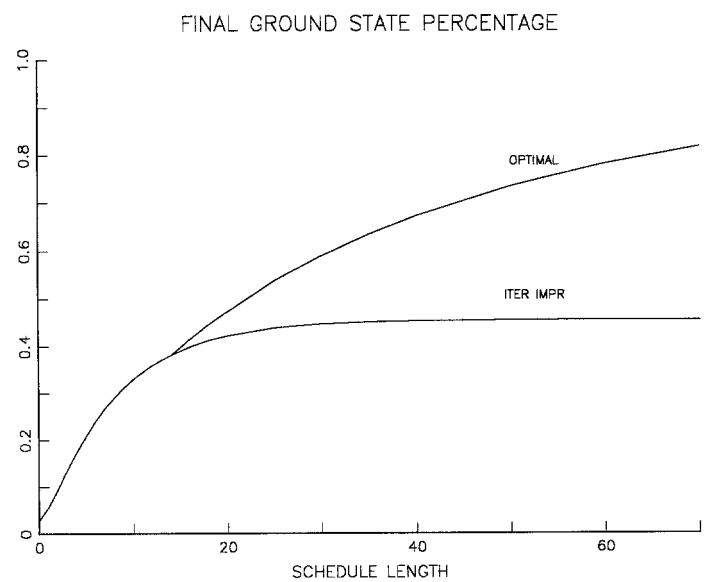

Fig. 5. Probability that the final state is the ground state as a function of length of schedule for both optimal schedule and iterative improvement. The problem is $\alpha = 3$ partitioning.

simple matrix multiplications involving the arrays E, P, $A(U)$, and $A'(U)$. Note that for $\alpha = 3$ the array elements are simple polynomials in U. The initial guess was typically a decreasing linear sequence of U values. The convergence to the solution was quite rapid and independent of the initial guess (although poor guesses resulted in convergence to a different local optimum having more zeros). The number of initial and final zeros was varied until no further locally optimal schedules were found and the one with the best value of the average final state energy was selected as the optimal annealing schedule of that length. There were typically about ten schedules from which to choose. Accuracy was sufficient to distinguish between local optima which sometimes had energies differing by 1 part in 10^6.

Although there is always the possibility that more locally optimal schedules exist, there is good reason to believe that the set we have found is complete. First it is unlikely that zero temperatures would appear other than at the extremes of the schedules. Initial guesses with middle zeros were tried and either resulted in the forcing of other (typically all) temperatures to zero, or else the temperatures simply moved away from zero toward their preferred value. Multiple optimal schedules with the same number of initial and final zeros are possible and were tested for by trying different starting configurations. Again the results involved convergence to the same solution or forcing of other temperatures to zero. Some numbers of initial and final zeros resulted in the forcing of the other temperatures to zero regardless of the starting schedule. Although no locally optimal schedules were found for these values, it is possible that some were missed due to sensitivity to initial conditions. However the energies of the optimal schedules that were found formed a simple parabolic surface as a function of the number of initial and final zeros, which changed slowly and smoothly with the length of the schedule. Any locally optimal schedules with small basins that were missed would presumably be at higher energy values anyway.

4. Interpretation. A number of the features observed in these schedules need further explanation. It has already been noted that the schedule of all zeros (iterative improvement) is always a local minimum in the space of schedules. It is also apparent that at some schedule length there is a transition after which this is no longer the global optimum. This transition simply reflects the idea that traditional annealing should not be used unless a reasonable amount of time is allocated.

The presence of the initial segment of zeros deserves more attention since virtually everyone (including the authors) had assumed that a correct annealing schedule should be monotonically decreasing. Unfortunately no proof is yet available for this phenomenon as there is for the section of final zeros. However with some hindsight this segment of the schedule becomes understandable. First notice that the length of this segment is basically independent of the length of the total schedule. Next note that although the initial configuration is random (infinite temperature), any real problem has a finite temperature scale. Finally, the temperature that appears in the annealing schedule is not an internal temperature

of the system unless the annealing temperature remains constant for long enough. From these points it is reasonable that when studying a system whose internal energies are bounded but whose initial temperature is infinite, the first order of business is to bring the nonequilibrium internal temperature down to a reasonable value. The most efficient way to do this is to use the lowest bath temperature, namely zero. Once the system has cooled down to the order of its internal energies, the intuitive picture of gradually decreasing schedule is valid. Also note that the segment of initial zero temperatures is closely tied to the random starting condition. A different starting configuration (for example, one obtained via some heuristic) could easily lead to a different initial segment. In particular, if only the local minima are initially populated, it is simple to show that a schedule of length two must not have an initial zero temperature.

The final temperatures in the schedule are also interesting. Although the number of zeros at the end decreases from ten to eight with increasing length of schedule this is probably due to the discrete nature of the schedule. When more temperatures are available a smoother approach to zero is possible, so fewer finite temperatures near the end are forced to zero. This viewpoint is consistent with the observation that the form of the decrease for differing lengths is quite similar.

The observation of an envelope for the annealing curves may also be surprising, since, given more annealing time, besides extending the schedule to lower temperatures, it would also be possible to allow more time at higher temperatures. This is apparently not optimal, however. This property could easily be an artifact of the small size of the current problem and the presence of only a single barrier.

5. Random Energies and a Variety of Schedules. The second group of experiments tested the performance of the various types of annealing schedules in common use. In each case, we have left one free parameter to be optimized for a given model and available number of iterations, rather than adjusting each temperature individually. In the introduction we gave an example of each type of schedule as a continuous function of time, t. Our model used discrete schedules, in which T is a function of an index variable, i, which are specified here. Besides the logarithmic and geometric schedules already described, we also considered a linear cooling schedule for its simplicity. This takes the form

$$T_{\text{lin}}(t) = T_0(1 - t/t_{\max}) \tag{9}$$

or

$$T_{\text{lin}}(i) = T_0(N_{\text{iter}} - i)/N_{\text{iter}}, \tag{10}$$

where the annealing extends up to time $t_{\max}$ (or for N_{iter} iterations),

A logarithmic cooling schedule was implemented discretely as

$$T_{\log}(i) \begin{cases} = \dfrac{T_0}{1 + \log(i)}, & i \leq N_{\text{iter}} - 10, \\ = 0, & N_{\text{iter}} - 10 < i \leq N_{\text{iter}}. \end{cases} \tag{11}$$

The value of final steps at zero temperature was shown in the preceding section. They also proved important with the logarithmic cooling schedule because the first formula of (11) leaves the temperature at intermediate values as i approaches N_{iter}. This implies that there is a significant probability of the Markov chain being in an excited state. The small number of final steps at $T(i) = 0$ fixes this, driving the probability density down into the nearby local minima. Our linear annealing schedule decreases continuously to zero, and does not need this extra "cleanup" phase. The geometric schedule, described next, does not terminate precisely at $T(N_{\text{iter}}) = 0$. It does end, however, at a sufficiently low temperature that a few final steps at $T = 0$ gave only a small additional improvement in the expected energy. We could perhaps have improved or accelerated our annealing schedules by adding initial steps at very low temperatures, as was found in the previous section. However, we did not study this effect in this series of experiments.

A geometric annealing schedule of the form (2) was created by choosing a final temperature, T_{final}, and interpolating logarithmically between T_0 and T_{final} for N_{iter} steps:

$$T_{\text{geo}}(i) = T_0(T_{\text{final}}/T_0)^{(i/N_{\text{iter}})}. \tag{12}$$

Geometric annealing schedules have two parameters, T_0 and T_{final}. In order to compare with the other, one-parameter annealing schedules, we used the same T_{final} for all values of N_{iter}. We selected the value $T_{\text{final}} = 0.02$, which gave the best result for $N_{\text{iter}} = 1000$ with the models used. Note that the factor b used in (2) can be identified as $b = (T_{\text{final}}/T_0)^{1/N_{\text{iter}}}$. Thus $b \to 1$ as N_{iter} increases.

We consider several model configuration spaces in these experiments, but most of our observations can be related to one of these—a set of 200 random energies created by a standard random number generator which produces a sequence of numbers uniformly distributed between 0.0 and 1.0. If we assume that transitions occur between adjacent states in the sequence, the "potential surface" drawn in Figure 6 is obtained. We can study other connectivities, for example by allowing

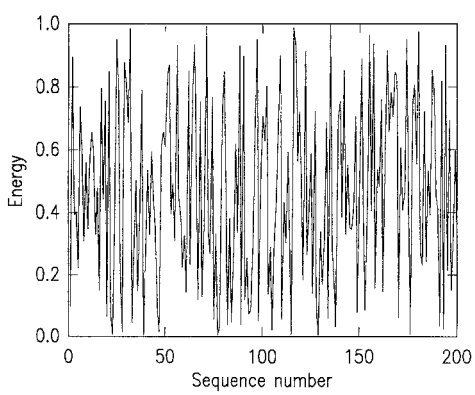

Fig. 6. Energy sequence used in the 200-random-states model. Used with nearest neighbor (200-2) and nearest plus next nearest neighbor (200-4) move sets.

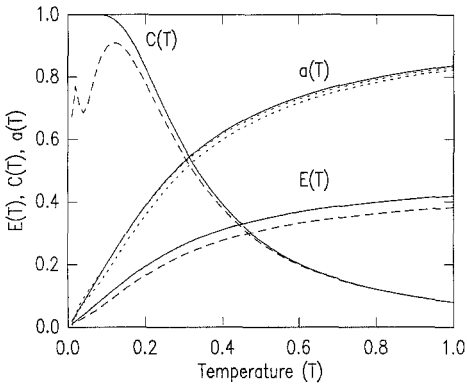

Fig. 7. Equilibrium energy, acceptance rate, and specific heat for the 200-state-random-energy model (dashed lines) are compared with the predictions of a uniform continuous distribution of states (solid lines). The acceptance rates for both the 2 and 4 neighbor versions of the random-state model are plotted (the 200–4 curve is higher). $E(T)$ and $C(T)$ are the same for the two models.

transitions between states separated by two (or more) steps along the sequence. Adding connections while keeping the state energies in the model unchanged will not change the equilibrium properties, such as the average energy or its derivative with respect to temperature, the specific heat

$$C(T) = dE(T)/dT, \tag{13}$$

since these quantities are determined by averages over all states, weighted by Boltzmann factors, which are functions of energy and temperature only. However, transport and relaxation properties tend to be dominated by "critical" barriers, the highest energies through which a Markov chain must pass to move from one part of the state space to another distant part. These barriers will be affected by changes in connectivity.

In Figure 7 we show the equilibrium properties of this model—energy, specific heat, and Metropolis move acceptance rate—evaluated by a Boltzmann-weighted average over the 200 energies used. As a check, and for comparison, we also evaluated and show the limiting expressions which a larger model would give. These are obtained by replacing the Boltzman sums with integrals over a continuous uniform distribution of energies. With only 200 energies, our model varies noticeably from the uniform distribution limit. The average energy is roughly 0.45, there are 15 energies less than 0.05 (ten are expected), and four states with energies less than 0.01 (two are expected). The smallest energy present is 0.004.

For a very large number of states, thermodynamic properties in equilibrium would take the "exact" limiting forms obtained by substituting the uniform density of states (from 0 to 1) into the expressions which define the partition function $Z(T)$.

$$Z(T) \equiv \int_0^1 dE\, e^{-\beta E} = \beta^{-1}(1 - e^{-\beta}), \tag{14}$$

where $\beta = 1/T$, and for the energy, $E(T)$,

$$E(T) \equiv Z^{-1} \int_0^1 dE\, E e^{-\beta E} = \frac{1 - e^{-\beta}(1+\beta)}{\beta(1-e^{-\beta})}. \tag{15}$$

Note that $E(T)$ in (15) tends to $\frac{1}{2}$ as $T \to \infty$, and has the limit $E(T) \sim T$ as $T \to 0$. The specific heat, $C(T)$, is given by

$$C(T) = Z^{-1}[2\beta^2 E(T) - \beta e^{-\beta E}] - Z^{-2}\beta^2 E(T)^2. \tag{16}$$

This complicated-looking expression is plotted in Figure 7. It reduces to 1 at $T = 0$, and decreases monotonically with increasing T. Boltzmann-weighted averaging over the 200 random energies gives an $E(T)$ and $C(T)$ which are in reasonable agreement with the "exact" limits, except at the lowest temperatures.

The average rate, $a(T)$, at which transitions are accepted under the Metropolis rule is an input to adaptive annealing schedules. In equilibrium, it depends only upon the initial and final energies and the temperature. Since the energies in our model are uncorrelated, $a(T)$ is obtained by averaging (4) over initial and final state energies. Summing over the number of final states which can be reached from a given initial state simply removes the factor $1/C$ in (4). Using the continuous distribution for both initial and final states, we obtain

$$a(T) = 2E(T), \tag{17}$$

for the uniform distribution. The actual rates of transitions predicted for the 200 random energies in equilibrium are shown by dotted lines in Figure 7 for the cases with connectivity 2 and 4. These cases are denoted models 200-2 and 200-4 for brevity. The differences between the two cases are small, and both agree fairly well with the estimate (17) from the uniform distribution.

The method of the previous section was now applied to the 200-energy models, using each of the annealing schedules and starting with the probability distribution uniform over the 200 states. Figures 8–10 show the expected energies achieved as functions of the adjustable parameter T_0 for the inverse logarithmic, geometric, and linear annealing schedules, respectively. None of the expected results approaches very closely the exact minimum energy in the model, 0.004. Inspection of the limiting probability distributions achieved showed that five to ten low-lying local minima participated in the result. The best value of T_0 is indicated on each curve by a data point. Results are shown for connectivities 2 and 4, and for values of N_{iter} increasing by factors of 5: 40, 200, 1000, and 5000. The curves with connectivity 4 roughly overlap those with connectivity 2 and 25× more search steps. That is, doubling the number of moves available in this particular structureless search space is equivalent to searching for 25× as long. Note that T_0 varies strongly with N_{iter} in both the inverse logarithmic and geometric schedules (Figure

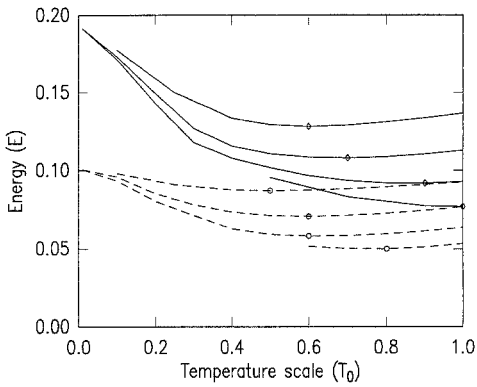

Fig. 8. Average final state energy as a function of temperature parameter T_0, using the inverse logarithmic cooling schedule defined in (11), for the 200–2 (solid lines) and 200–4 (dashed lines) models. Results for $N_{\text{iter}} = 40, 200, 1000, 5000$ are shown for each model. The longer schedules correspond to lower energies. Values of T_0 for which best results were obtained are marked with a diamond (for 200–2) or a circle (200–4).

8 and 9), but in the opposite direction. There is little dependence on total annealing time in the linear schedule results (Figure 10). In all three cases, the best value of T_0 for a given N_{iter} decreases when the connectivity is increased from 2 to 4.

The inverse logarithmic schedule obtains better results than the linear or geometric schedule when $N_{\text{iter}} = 40$, but is the poorest of the three for $N_{\text{iter}} \geq 200$. This schedule also requires the most careful tuning of the three, and thus would be the least robust in practice. As a measure of robustness, we compared the expected energy at twice the best T_0 with the best expected energy. For the inverse logarithmic schedule, the increase was 15% at $N_{\text{iter}} = 40$, increasing to about 40% at $N_{iter} = 5000$.

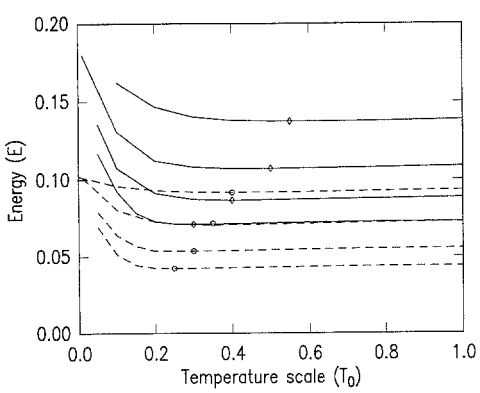

Fig. 9. Average final state energy as a function of temperature parameter T_0 for the geometric cooling schedule defined by (12). Conventions are as in Figure 8.

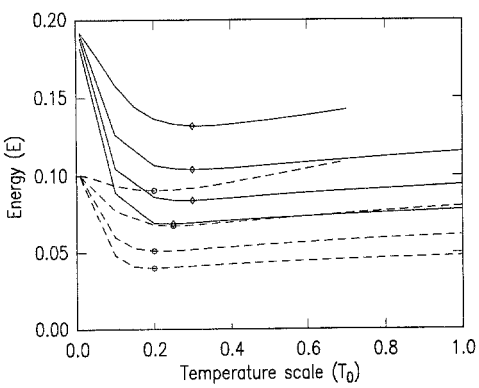

Fig. 10. Average final state energy as a function of temperature parameter T_0 with the linear cooling schedule defined by (10). Conventions are as in Figure 8.

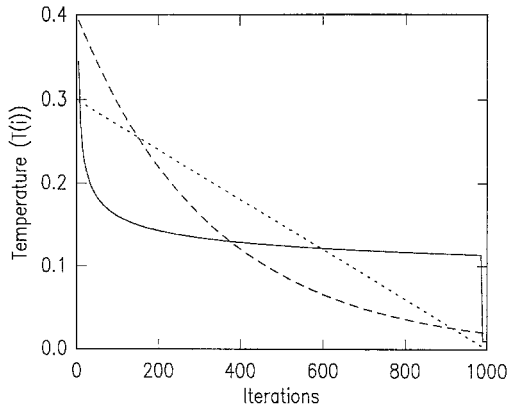

Fig. 11. Three optimized annealing schedules for $N_{\text{iter}} = 1000$: inverse logarithmic cooling (solid line), geometric (long-dashed line), and linear (short-dashed line).

It is hard to choose between the linear and geometric schedules for this model problem, since their results are consistently close. The geometric schedule is the least sensitive to the choice of T_0, since the energy expected at twice the best choice of T_0 is only about 1% higher than the minimum. For the linear schedule, this increase was approximately 5% at all values of N_{iter}. However, the best results obtained with the linear annealing schedule were about 3% better than with the geometric annealing. On the other hand, the geometric annealing schedule has two free parameters. If we vary T_{final} as well as T_0, for $N_{\text{iter}} = 5000$, the best expected energy obtained was 0.5% below the linear schedule's result.

The three schedules we studied all spend most of their time at temperatures below 0.3, but that time is allocated quite differently. Figure 11 shows the best annealing schedule of each type, for $N_{\text{iter}} = 1000$. The inverse logarithmic schedule keeps its temperature above 0.12 until the final iterations at zero temperature. The optimal value of T_0 for the inverse logarithmic schedule increases with increasing N_{iter} and this has the effect of keeping the temperatures reached in most of the schedule from becoming small. The geometric schedule has a form intermediate between the linear and the inverse logarithmic. It also ends at nonzero temperature, and thus benefits from having a short period of zero temperature iterations added at the end to permit final local improvements. The average energies resulting as functions of time (iteration count), using the schedules shown in Figure 11 are shown in Figure 12. The energy curves in Figure 12 have the same shape at intermediate temperatures as the temperature curves, although there seems to be no reason to expect this in more general problems.

Figures 13–15 show the average energy at each temperature during the schedule for the best schedule of each type (logarithmic, geometric, and linear cooling), with connectivity 2 and several values of N_{iter}. The equilibrium value of the energy at that temperature is also given for comparison. In all cases the annealed energy stays strictly above the equilibrium value. In both the geometric and linear cases the energy rapidly closes with its equilibrium value, follows it until lower

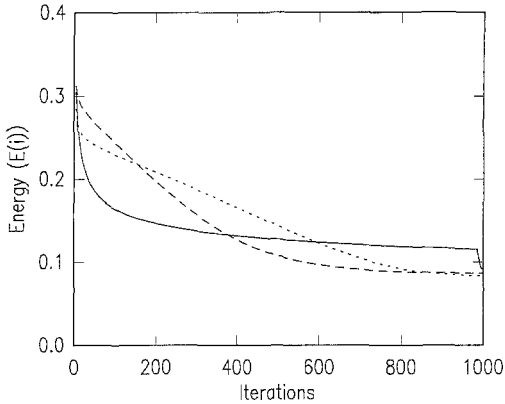

Fig. 12. Energy as a function of time (iteration number) for the three annealing schedules shown in Figure 11.

temperatures and then diverges. In contrast the logarithmic schedules do not seem to approach the equilibrium value until quite late in the schedule. The logarithmic schedules have the peculiarity that an appreciable reduction in energy occurs during the ten final iterations, which are taken at a very low temperature. This gives rise to the points which are isolated on the left-hand side of the graph.

We observe that none of the annealing procedures produces thermal equilibrium, even at intermediate temperatures. Furthermore, the inverse logarithmic schedule, which gave the best results when only a limited number of iterations is allowed, stays further from equilibrium during the course of its annealing than the less effective schedules, so it is not clear that equilibrium is desirable, except at the completion of the process. This challenges the basic assumption made in proposing and analysing adaptive annealing schedules, namely that equilibrium properties of

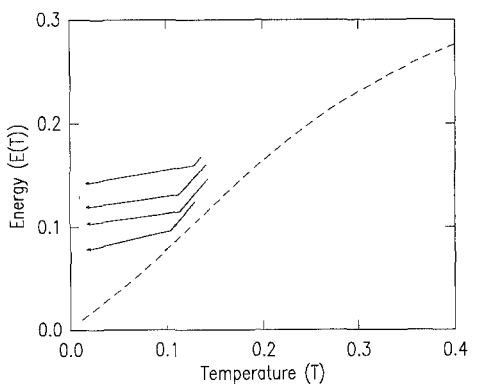

Fig. 13. Average energy at each temperature during the best logarithmic schedule for the 200–2 model for $N_{\text{iter}} = 40, 200, 1000$, and 5000, compared with the equilibrium energy (dashed line). Longer schedules correspond to lower energies. The cluster of points at $T = 0$ result from the multiple zero temperatures used at the ends of the schedules.

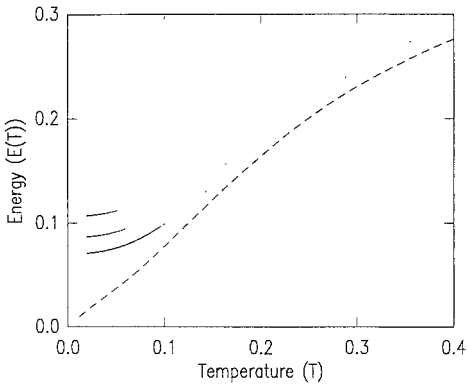

Fig. 14. Average energy at each temperature during the best geometric schedule for the 200–2 model for $N_{\text{iter}} = 200, 1000$, and 5000, compared with the equilibrium energy (dashed line). Longer schedules correspond to lower energies.

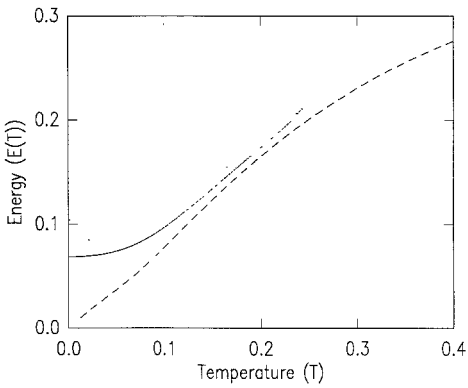

Fig. 15. Average energy at each temperature during the best linear schedule for the 200–2 model for $N_{\text{iter}} = 200, 1000$, and 5000, compared with the equilibrium energy (dashed line). Longer schedules correspond to lower energies.

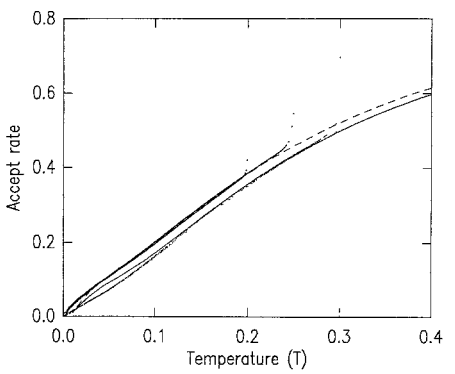

Fig. 16. Average acceptance rate at each temperature with the best linear schedules of 200 and 1000 steps (dots) compared with their equilibrium predictions for the 200–2 (solid) and 200–4 (dashed) models (also shown in Figure 7).

the system will be measured during simulated annealing and can be used in the analysis. We cannot measure the specific heats that would be observed during actual simulated annealing runs, adaptive or otherwise, because our method does not give access to averages restricted to a single Markov chain. However, we can observe the rate at which moves are accepted. Figure 16 shows the acceptance rates as a function of the temperature during linear annealing, plotting both connectivities and two values of N_{iter}. Although the annealed energy stays above the equilibrium value, the rate of acceptance closely follows the equilibrium value. Equilibrium acceptance rates, calculated with a Boltzmann distribution, are probably, therefore, a good estimate of the rates that will be seen with the actual out-of-equilibrium distribution that occurs when annealing.

6. Conclusions. This paper has provided the first results on optimal finite-time annealing schedules, employing a new technique to evaluate exactly the probability distribution of outcomes of the annealing process. Although this technique is limited to problems with very small solution spaces, for some nontrival problems we find interesting results. We find that several locally-optimal annealing schedules may exist, including iterative improvement. When a sufficiently long schedule is employed, annealing replaces iterative improvement as the globally best solution.

The initial segment of zeros found at the beginning of each schedule (leading to nonmonotonic behavior) was unexpected, but we suggest a physical motivation for the phenomenon. A final segment of zeros is expected and observed. A theoretical lower bound for the number of final zeros is provided. The curves for different length schedules seem to follow an envelope consistent with a theoretical infinite-time schedule, but this evidence is not conclusive.

We have also compared the performance of three types of annealing schedules on an artificial system—200 random energies with limited conectivity providing a "white noise" potential surface. Both the linear and geometric annealing schedules studied outperformed an inverse logarithmic schedule when sufficient computing effort was allowed, and both were reasonably robust in giving good results for a range of parameter settings. The inverse logarithmic schedule is known to maximize the probability that the system ends up on the better side of a given large "critical" barrier. This may be why it gives a good approximation to the schedule envelope found for our first model. In the second model, with random energies, multiple low-lying solutions are reached, and presumably there are multiple critical barriers. It might be expected that a heuristic which blends several barrier-dependent schedules would take a form more like the linear or geometric schedule.

Our experiments did not distinguish strongly between linear and geometric forms of annealing schedule, but they do show that the geometric schedules are not degraded seriously by setting the initial temperatures too high. Having a second parameter, T_{final}, may make the geometric more powerful in practice, since the two temperatures affect different portions of the schedule. They can be determined independently, one from the observed high-temperature rearrangements of the system (its "melting" behavior) and the other from the reduced transition rates seen at low temperatures (its "freezing").

Since our master equation technique does not permit simulating adaptive annealing schedules, we have concentrated on checking the underlying assumptions of these schedules. Adaptive schedules are tuned by continously measuring specific heat and move acceptance rate. We constructed two models (200-2 and 200-4) which have the same equilibrium specific heat and almost the same local transition rates, yet the best annealing schedules found were rather different for the two models. If an adaptive technique could really access equilibrium values, it would produce identical schedules for these two cases, so at least one of them would be quite suboptimal. We find that our best heuristic schedules produce annealing trajectories which are never in equilibrium. The rate at which moves are accepted, however, appears to be insensitive to the degree to which the system is not in an equilibrium distribution. Adaptive annealing schedules do seem effective in practice, though our results challenge their usual derivations. We suggest that more work is needed to determine when quantities such as specific heat, estimated out of equilibrium by averaging over only a portion of the system's phase space, will nonetheless contain the proper information to identify "interesting" temperature regions over which to anneal more slowly.

Acknowledgments. We would like to thank Greg Sorkin for suggesting the model partitioning problem and for a detailed review of the manuscript. Comments from Daphna Weinshall are also gratefully acknowledged.

References

[1] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Optimization by Simulated Annealing, *Science*, vol. 220, pp. 671-680, May 1983.

[2] D. Jepsen and C. D. Gelatt, Jr., Macro Placement by Monte Carlo Annealing, *Proc. ICCAD 83*, pp. 495-498, Oct. 1983.

[3] C. Sechen and A. Sangiovanni-Vincentelli, The Timberwolf Placement and Routing Package, *IEEE J. Solid-State Circuits*, vol. 20, no. 2, pp. 510-522, Apr. 1985.

[4] S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions, and Bayesian Restoration of Images, *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 6, pp. 721-741, Nov. 1984.

[5] P. Carnevali, L. Coletti, and S. Patarnello, Image Processing by Simulated Annealing, *IBM J. Res. Develop.* vol. 29, no. 6, pp. 569-579, Nov. 1985.

[6] E. D. Sontag and H. J. Sussman, Image Restoration and Segmentation Using the Annealing Algorithm, *Proc. 24th Conf. on Decision and Control*, pp. 768-773, Dec. 1985.

[7] F. C. Jeng and J. W. Woods, Image Estimation by Stochastic Relaxation in the Compound Gaussian Case, *Proc. ICASSP 88*, Apr. 1988.

[8] P. Molitor, Layer Assignment by Simulated Annealing, *Microprocess. Microprogramm. (Netherlands)*, vol. 16, no. 4,5, pp. 345-349, Nov./Dec. 1985.

[9] M. Vecchi and S. Kirkpatrick, Global Wiring by Simulated Annealing, *IEEE Trans. Computed-Aided Design*, vol. 2, no. 4, pp. 215-222, Oct. 1983.

[10] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli, Convergence and Finite-Time Behavior of Simulated Annealing, *Proc. 24th Conf. on Decision and Control*, pp. 761-767, Dec. 1985.

[11] S. B. Gelfand and S. K. Mitter, Analysis of Simulated Annealing for Optimization, *Proc. 24th Conf. on Decision and Control*, pp. 779-786, Dec. 1985.

[12] B. Hajek, A Tutorial Survey of Theory and Applications of Simulated Annealing, *Proc. 24th Conf. on Decision and Control*, pp. 755-760., Dec. 1985.

[13] S. A. Solla, G. B. Sorkin, and S. R. White, Configuration Space Analysis for Optimization Problems, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman, and G. Weisbuch (eds.), Springer-Verlag, New York, pp. 283, 1986.

[14] S. Kirkpatrick and G. Toulouse, Configuration Space Analysis of Travelling Salesman Problems, *J. Physique*, vol. 46, pp. 1277–1292, Aug. 1985.

[15] M. Mezard, On the Statistical Physics of Spin Glasses, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman, and G. Weisbuch (eds.), Springer-Verlag, New York, 1986.

[16] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glasses and Optimization*, World Scientific Press, London, 1988.

[17] E. Aarts and P. van Laarhoven, Statistical Cooling Algorithm: A General Approach to Combinatorial Optimization Problems, *Philips J. Res.*, vol. 40, no. 4, 193–226, 1985.

[18] M. Huang, F. Romeo, and A. Sangiovanni-Vincentelli, An Efficient General Cooling Schedule for Simulated Annealing, *Proc. ICCAD 86*, pp. 381–384, 1986.

[19] J. Lam and J.-M. Delosme, An Efficient Simulated Annealing Schedule: Derivation, Technical Report 8816, and An Efficient Simulated Annealing Schedule: Implementation and Evaluation, Technical Report 8817, Department of Electrical Engineering, Yale University, New Haven, CT, Sept. 1988.

[20] R. H. J. M. Otten, Floorplan Design Using Simulated Annealing, *Proc. ICCAD 84*, pp. 96–98, Nov. 1984.

[21] P. N. Strenski, Optimal Annealing Schedules, IBM Research Report RC12923, Yorktown Heights, NY, July 1987.

[22] Harwell Subroutine Library, compiled by M. J. Hopper, AERE-R9185, Computer Science and Systems Division, Atomic Energy Research Establishment, Harwell, Oxfordshire, Dec. 1978.