OPTIMIZATION BY SIMULATED ANNEALING:

A NECESSARY AND SUFFICIENT CONDITION FOR CONVERGENCE

Bruce Hajek*

University of Illinois at Champaign-Urbana

A Monte Carlo optimization technique called "simulated annealing" is a descent algorithm modified by random ascent moves in order to escape local minima which are not global minima. The level of randomization is determined by a control parameter T, called temperature, which tends to zero according to a deterministic "cooling schedule". We give a simple necessary and sufficient condition on the cooling schedule for the algorithm state to converge in probability to the set of globally minimum cost states. In the special case that the cooling schedule has parametric form $T_k = c/\log(1+k)$, the condition for convergence is that c be greater than or equal to the depth, suitably defined, of the deepest local minimum which is not a global minimum state.

"Annealing" in the physics literature refers to the process of slowly cooling a substance in order to reach globally minimum energy states. Cerny (1982) and Kirkpatrick, Gelatt and Vecchi (1983) suggested simulating such a process in order to solve large-scale minimization problems on a computer. Randomization is introduced using the classical method of Metropolis et al. (1953) for generating sample realizations of random fields. See Geman and Geman (1984), Kirkpatrick et al. (1983) and the references therein for more background information.

* This work was supported by the Office of Naval Research under Contract N000-14-82-K-0359 and the National Science Foundation under contract NSF-CS-83-52030.

AMS 1980 subject classification. Primary 60J05, Secondary 93E03, 68Q75.

Key words and phrases: stochastic optimization, probabalistic hill-climbing, simulated annealing.

Suppose that a function V defined on some finite set $\underline{S}$ is to be minimized. We assume that for each state s in $\underline{S}$ that there is a set N(s), with $N(s) \subset \underline{S}$, which we call the set of neighbors of s. Typically the sets N(s) are small subsets of $\underline{S}$. In addition, we suppose that there is a transition probability matrix R over $\underline{S}$ such that $R(s,s') > 0$ if and only if s' is in N(s).

Let $T_1, T_2, \ldots$ be a sequence (called a temperature schedule) of strictly positive numbers such that

$$T_1 \geq T_2 \geq \ldots \tag{1}$$

and

$$\lim_{k \to \infty} T_k = 0 \tag{2}$$

Consider the following sequential algorithm for constructing a sequence of states $X_0, X_1, \ldots$. An initial state X_0 is chosen. Given that $X_k = s$, a potential next state Y is chosen from N(s) with probability distribution

$$P[Y=s' \mid X_k=s] = R(s,s').$$

Then we set

$$X_{k+1} = \begin{cases} Y & \text{with probability } P_k \\ X_k & \text{otherwise} \end{cases}$$

where

$$p_k = \exp\left\{\frac{-[V(Y)-V(x)]^+}{T_k}\right\}.$$

This specifies how the sequence $X_1, X_2, \ldots$ is chosen. Let $\underline{S}^*$ denote the set of states in $\underline{S}$ at which V attains its minimum value. We are interested in

determining whether or not

$$\lim_{k\to\infty} P[X_k \in S^*] = 1.$$

We say that i is reachable at height E from state j if there is a sequence of states $j=i_0,i_1,\ldots,i_p=i$ such that

$$R(i_k,i_{k+1}) > 0 \text{ for } 0 \leq k < p$$

and

$$V(i_k) \leq E \text{ for } 0 \leq k \leq p.$$

We will assume that $(\underline{S},V,R)$ has the following two properties:

Property SI (strong irreducibility): Given any two states i and j is reachable (at some height) from j.

Property WR (weak reversibility): For any real number E and any two states i and j, i is reachable at height E from j if and only if j is reachable at height E from i.

State s is said to be a local minimum if no state s' with $V(s') < V(s)$ is reachable from s at height V(s). We define the depth of a local minimum s to be plus infinity if s is a global minimum. Otherwise, the depth of s is the smallest number E, $E > 0$, such that some state s' with $V(s') < V(s)$ can be reached from s at height $V(s) + E$. These definitions are illustrated in Fig. 1.

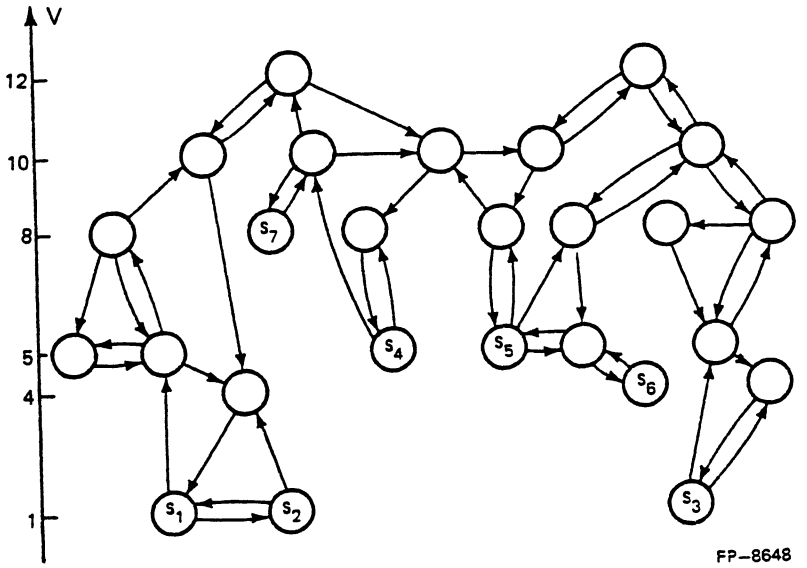

Fig. 1. The graph pictured arises from a triplet ($\underline{S}$,V,R). Nodes correspond to elements in $\underline{S}$. V(s) for s in $\underline{S}$ is indicated by the scale at left. Arcs in the graph represent ordered pairs of states (s,s') such that R(s,s') > 0. Properties SI and WR are satisfied for the example shown.

States s_1, s_2 and s_3 are global minimum. States s_4, s_6 and s_7 are local minima of depths 5.0, 6.0 and 2.0 respectively. State s_5 is not a local minima. State s_2 is reachable at height 1.0 from s_1 and state s_3 is reachable at height 12.0 from s_1.

We define a cup for (S,V,R) to be a set C of states such that for some number E, the following is true: For every s in C,

$$C = \{s': s' \text{ can be reached at height } E \text{ from } s\}.$$

Given a cup C, define

$$\underline{V}(C) = \min\{V(s): s \in C\}$$

and

$$\overline{V}(C) = \min\{V(s): s \notin C \text{ and } R(s',s) > 0 \text{ for some } s' \text{ in } C\}.$$

We call the subset B of C defined by

$$B = \{s \in C: V(s) = \underline{V}(C)\}$$

the bottom of the cup, and we call the number d(C) defined by

$$d(C) = \overline{V}(C) - \underline{V}(C)$$

the depth of the cup. These definitions are illustrated in Fig. 2. Note that a local minimum of depth d is an element of the bottom of some cup of depth d. Our main result is the following theorem.

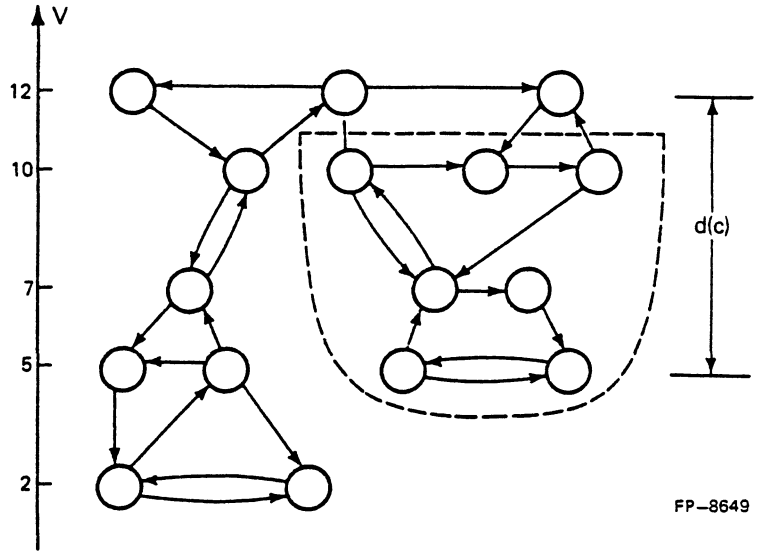

Fig. 2. A cup C is enclosed with dashed lines. $\underline{V}(C) = 5$, $\overline{V}(C) = 12$, and $d(C) = 7$ and the bottom B of C contains two states.

THEOREM 1. Assume that SI, WR, (1) and (2) hold.

(a) For any state that is not a local minimum,

$$\lim_{k\to\infty} P[X_k = s] = 0.$$

(b) Suppose that the set of states B is the bottom of a cup of depth d and that the states in B are local minima of depth d. Then

$$\lim_{k\to\infty} P[X_k \in B] = 0 \tag{3}$$

if and only if

$$\sum_{k=1}^{\infty} \exp(-d^*/T_k) = +\infty \ . \tag{4}$$

Remarks. If T_k assumes the parametric form

$$T_k = \frac{c}{\log(k+1)} \tag{5}$$

then condition (4), and hence also condition (3), is true if and only if $c \geq d^*$. This result is consistent with the work of Geman and Geman (1984). They considered a model which is nearly a special case of the model used here, and they proved that condition (3) holds if (T_k) satisfies equation (5) for a sufficiently large constant c. They gave a value of c which is sufficient for convergence. Their value is substantially larger, although somewhat related, to d^*.

Gidas (1984) also addressed the convergence properties of the annealing algorithm. The Markov chains that he considered are more general than those that we consider. He required little more than the condition that the one-step transition probability matrices P_k converge as k tends to infinity. In the special case of annealing processes, he gave a value of c (actually, c here corresponds to $1/C_0$ in Gidas' notation) which he conjectured is the smallest such that Eq. (5) leads to Eq. (3). His constant is different from the constant d^* defined here. Gidas also considered interesting convergence questions for functionals of the Markov chains.

Geman and Hwang (1984) showed that in the analogous case of non-stationary diffusion processes that a schedule of the form (5) is sufficient for convergence to the global minima if c is no smaller than the diference between the maximum and minimum value of V. We conjecture that the smallest constant is given by the obvious analogue of the constant d^* that we defined here.

Sketch of the proof: It is best to consider an example to get an idea of what is involved. Consider ($\underline{S}$,V,R) giving rise to Fig. 3. Note that d* = 3.

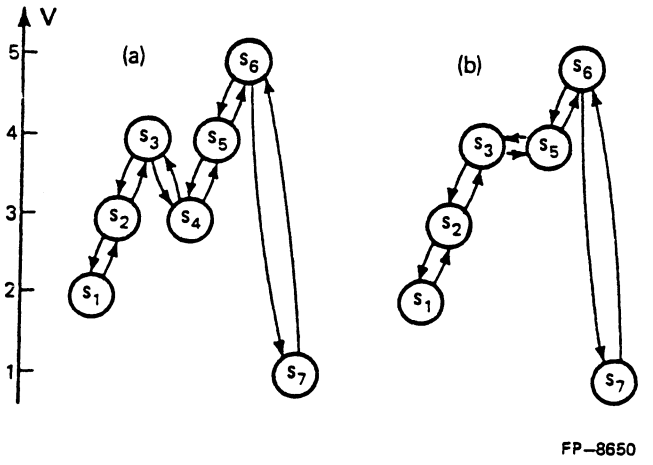

Fig. 3. Diagram (a) arises from a triple ($\underline{S}$,V,R) and diagram (b) is obtained by "filling-in" the cup $\{s_4\}$.

Now suppose that (T_k) satisfies (1), (2),

$$\sum_{k=1}^{\infty} \exp(-2/T_k) = +\infty, \tag{6}$$

and

$$\sum_{k=1}^{\infty} \exp(-3/T_k) < +\infty. \tag{7}$$

Since s_4 is reachable at height $V(s_1)+2$ and since (6) is assumed, one might (correctly) guess that if the process starts at s_1 then it will eventually reach s_4 with probability one. By similar reasoning, one might then (incorrectly) guess that the process must eventually reach s_7. However, by Theorem 1 and (7),

$$\lim_{k\to\infty} P[X_k = s_7] \neq 1.$$

What happens is that if k is large so that T_k is small and if the process is in state s_4 at time k, then it is much more likely that the process hits state s_1 before it hits (if ever) state s_6. We think of the cup consisting of state s_4 alone as being "filled-in" (see Fig. 3), so that to get from s_1 to s_6, the process has to climb up three levels. Roughly speaking, the small depression in V at s_4 does not allow the process to always make it up three levels by going up two at a time and "resting" in between. This would not be true if condition SR was violated by, for example, setting the probability of jumping from s_4 to s_3 to zero.

A cornerstone of our proof of Theorem 1 given in Hajek (1985) is the lemma stated beow, which allows us to "fill-in" cups in a precise sense. Let $E_0, E_1, \ldots, E_k$ denote the possible values of V(s) as s varies over $\underline{S}$, ordered so that $E_0 < E_1 < \ldots < E_K$. By an embedding argument, we can suppose without loss of generality that $(\underline{S},V,R)$ has the following property:

Continuous Increase Property: Given any two states s and s', if $V(s) = E_i$ and $V(s') = E_j$ and $j \geq i+2$, then $R(s,s') = 0$.

The $(\underline{S},V,R)$ giving rise to Fig. 1 does not have the continuous increase property, while the one giving rise to Fig. 2 does.

Suppose that C is a cup, let F denote the set of states in the complement of C which can be reached in one jump from states in C, and let d denote the depth of C.

LEMMA (How cups runneth over): There exists an $\varepsilon > 0$, depending only on $(\underline{S},V,R)$ and C, so that for any time $t_0 \geq 0$, any i_0 in C, any j_0 in F, and any T

satisfying (1), the following conditions hold:

(a) (Exponential expulsion rate)

$$P[\text{not exit } C \text{ during } [t_0, t_0+r] \mid X_{t_0} = i_0] \leq \frac{1}{-} \exp(-\varepsilon \sum_{j=t_0}^{t_0+r} \exp(-d/T_j))$$

for all $r \geq 0$.

(b) (Quasi-uniform exit distribution)

(3.1) $P[\text{never exit } C \text{ or hit } j_0 \text{ upon first jump out of } C \mid X_{t_0} = i_0] \geq \varepsilon$.

Remark. It is useful to interpret the integral in the exponent on the right hand side of the inequality in part (a) to be the time escaped, as measured on the "d-th time scale", between actual times t_0 and t_0+r. Then part (a) means that the time required to exit from a cup of depth d is exponentially bounded on the d-th time scale.

The lemma can be proved essentially by induction on the depth of the cup C.

Acknowledgments.

I benefited greatly from my enchanting experience working for Herbert Robbins one summer at Brookhaven National Laboratories while I was a highly impressionable undergraduate student. I will always be thankful to him.

REFERENCES

Cerny, V. (1982). A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. Preprint, Inst. of Phys. and Biophysics, Comenium Univ., Bratislava.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Trans. Pattern Analysis and Machine Intelligence. 6 721-741.

Geman, S. and Hwang, C.-R. (1984). Diffusions for global optimization. Preprint. Div. of Applied Math, Brown University, Received December 5, 1984.

Gidas, B. (1985). Non-stationary Markov chains and convergence of the annealing algorithm. J. Stat. Phy. **39** 73-131.

Hajek, B. (1985). Cooling schedules for optimal annealing. Submitted to Mathematics of Operations Research.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. Science **220** 621-680.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines. J. Chem. Phys. **21** 1087-1091.