Proceedings of 24th Conference
on Decision and Control
Ft. Lauderdale, FL • December 1985

A TUTORIAL SURVEY OF THEORY AND APPLICATIONS OF SIMULATED ANNEALING

By Bruce Hajek

Coordinated Science Laboratory
Dept. of Electrical & Computer Engineering
University of Illinois
1101 W. Springfield, Urbana, IL 61801

ABSTRACT

Annealing is the process of slowly cooling a physical system in order to obtain states with globally minimum energy. By simulating such a process, near globally-minimum-cost solutions can be found for very large optimization problems. The purpose of this paper is to review the basic theory of simulated annealing, to survey its recent applications, and to survey the theoretical approaches that have been used to study the technique.

The applications include image restoration, combinatorial optimization (eg VLSI routing and placement), code design for communication systems and certain aspects of artificial intelligence. The theoretical tools for analysis include the theory of nonstationary Markov chains, statistical physics analysis techniques, large deviation theory and singular perturbation theory.

1. THE ANNEALING ALGORITHMS

1.1. Finite State–Space, Discrete Time

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

$$T_1 \geq T_2 \geq \cdots \tag{1.1}$$

and

$$\lim_{k\to\infty} T_k = 0. \tag{1.2}$$

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_k is chosen from N(s) with probability distribution

$$P[Y_k = s' | X_k = s] = R(s,s').$$

Then we set

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

where

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

This specifies how the sequence $X_1, X_2, \cdots$ is chosen. The random process $X=(X_k: k \geq 0)$ produced by the algorithm is a discrete time Markov chain.

We will give an explanation for this algorithm. Let S* denote the set of states in S at which V attains its minimum value. We are interested in determining whether or not

$$\lim_{k\to\infty} P[X_k \in S^*] = 1. \tag{1.3}$$

We say that state i is reachable from state j if there is a sequence of states $j=i_0, i^1, \ldots, i_p=i$ such that $R(i_k,i_{k+1})>0$ for $0 \leq k < p$. We will assume that (S, V, R) has the following property:

Property SI (Strong irreducibility): Given any two states i and j, i is reachable from j.

Consider the annealing algorithm in the special case that the temperature is identically equal to a constant T. Then the chain X has stationary transition probabilities, and a unique (if $0 < T \leq +\infty$) equilibrium probability distribution π_T. It is easy to check that π_T for $0 < T < \infty$ is related to π_∞ by

$$\pi_T(s) = \pi_\infty(s)\exp\left\{-\frac{V(s)}{T}\right\}/Z_T \tag{1.4}$$

where

$$Z_T = \sum_s \pi_\infty(s)\exp\left\{-\frac{V(s)}{T}\right\}$$

By the assumption that R is strongly irreducible (and the fact that P is aperiodic if $S \neq S^*$ since $P(i,i) > 0$ for some i in that case), the Markov ergodic convergence theorem [Sen81] implies that

$$\lim_{k\to\infty} P[X_k \in S^*] = \sum_{s\in S^*} \pi_T(s). \tag{1.5}$$

Examination of (1.4) soon yields that the right hand side of (1.5) can be made arbitrarily close to one by choosing T small. Thus,

$$\lim_{T\to 0}\left\{\lim_{k\to\infty, T_k \equiv T} P[X_k \in S^*]\right\} = 1.$$

The idea of the simulated annealing algorithm is to try to achieve (1.3) by letting T_k tend to zero as k tends to infinity. The annealing algorithm with T constant was first suggested in [MRRTT53], and recent interest in it was sparked by the articles [KiGV83] and [Cer82].

1.2. Finite State–Space, Continuous Time

We assume that a cooling schedule $(T_t: t \geq 0)$ is given such that

$$T_t \text{ is strictly positive and nonincreasing in } t, \tag{1.6}$$

and

$$\lim_{t\to\infty} T_t = 0, \tag{1.7}$$

and we assume that (S, V, R) is given as before. Now let

$$Q_t(s,s') = R(s,s')\exp(-[V(s')-V(s)]^+/T_t) \tag{1.8}$$

for s, s' with $s \neq s'$. We consider the pure jump Markov process (X_t) with transition rate matrix (Q_t). Thus,

$$P[X_{t+h} = s' | X_t = s] = Q_t(s,s')h + o(h).$$

If we denote $\alpha_t(i) = P[X_t = i]$, then α satisfies the differential equation

$$\dot{\alpha}_t = \alpha_t Q_t.$$

The explanation we gave for the discrete time annealing algorithm -- including Eq. (1.4) -- remains valid for this continuous time algorithm.

1.3. Continuous State −Space, Small Jumps

Consider a function V on $\mathbb{R}^n$ to be minimized and let $(T_t : t \geq 0)$ denote a temperature schedule as in Section 1.2. The Langevin Algorithm (notation suggested in [Gid85]) is to generate the solution of the following stochastic differential equation:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2T_t}dw_t; \quad X_0 = x_0$$

where $(w_t : t \geq 0)$ is a standard n-dimensional Wiener process. In the special case that $T_t \equiv 0$, this is the (deterministic) negative gradient algorithm, which cannot escape from local minima of V. The stochastic term dw_t is added in order to allow the process X to escape local minima of V.

2. CONVERGENCE IN PROBABILITY TO THE GLOBAL MINIMUM

We will give a convergence result for the finite state-space annealing algorithms. We say that state 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 (S, V, R) has the following two property:

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 $V(s) + E$.

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,

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

Given a cup C, define $\underline{V}(C) = \min\{V(s) : s \in C\}$ and

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

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 $d(C) = \bar{V}(C) - \underline{V}(C)$ the *depth* of the cup. Note that a local minimum of depth d is an element of the bottom of some cup of depth d.

Theorem1. [Haj85] Assume that SI, WR, (1.1) and (1.2) hold.

(a) For any state s 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

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

if and only if

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

(c) (Consequence of (a) and (b)) Let d* be the maximum of the depths of all states which are local but not global minima. Let S* denote the set of global minima. Then

$$\lim_{k \to \infty} P[X_k \in S^*] = 1 \tag{2.1}$$

if and only if

$$\sum_{k=1}^{\infty} \exp(-d^*/T_k) = +\infty. \tag{2.2}$$

Remarks. If T_k assumes the parametric form

$$T_k = \frac{c}{\log(k+1)} \tag{2.3}$$

then condition (2.2), and hence also condition (2.1), is true if and only if $c \geq d^*$. This result is consistent with the work of Geman and Geman [GeGe84]. They considered a model which is nearly a special case of the model used here, and they proved that condition (2.1) holds if (T_k) satisfies equation (2.3) for a sufficiently large constant c. Tsitsiklis [Tsi85] recently proved a result more general than Theorem 1 in which weak reversibility is not assumed.

Gidas [Gid84] also addressed the convergence properties of the annealing algorithm. He gave a value of c (actually, c here corresponds to $1/C_o$ in Gidas' notation) which he conjectured is the smallest such that Eq. (2.3) leads to Eq. (2.1). 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 [GeHw84] showed for the Langevin Algorithm that a schedule of the form (2.3) is sufficient for convergence to the global minima if c is no smaller than the difference 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. (Also, see Subsection 3.3).

3. THE CASE OF CONSTANT TEMPERATURE

3.1. Discussion

Most of the theoretical results on simulated annealing published to date can be better understood, if not even proved, by considering the annealing algorithm at a fixed temperature. The idea is that if the temperature varies with time like $c/\log(t)$, then the annealing process tends to reach quasi-equilibrium or escape local minima at a much faster rate than the temperature is varying. In particular, [GeGe84], [GeHw84], [Gid85] and [MiRS85] use this idea working with equilibrium distributions, where rate of convergence estimates such as those found in [Sen81] are used quite naturally. The recent papers [Kus85], [ChHS85] and [Tsi85] clearly point out this idea by working with escape time estimates. Here either large deviation theory such as that in [FrWe84] or the theory of Markov chains with a small parameter is applied. In this section we review some known escape-time estimates, and point out their significance for simulated annealing.

3.2. Escape Time Estimates—Discrete State Space

Consider the (discrete or continuous time) annealing algorithm run at a fixed temperature T. Suppose the underlying system (S,V,R) is strongly irreducible and weakly reversible. Let C be a cup with depth d(C), and let

$$\tau_C = \min\{t \geq 0 : X_t \notin C\}$$

Then the following in true (consequence of [Haj85, Thms. 3 and 4])

$$\lim_{t \to 0} T \log E_i \tau_C = d(C) \quad i \in C, \tag{3.1}$$

where the subscript i on E_i denotes that $X_0 = i$. Results of this sort can be found in the literature on singular perturbation of Markov chains (see [DeMQ85, CoWS83]) which is based on the theory of singular pertubation of linear operators [Kat76]. I suspect that much more is true and can be proved in a straight-forward way using singular perturbation techniques. For example, I conjecture that $\tau_C \exp(-d/T)$ converges in distribution to an exponentially distributed random variable with parameter

$$\Gamma_C = \left(\sum_{i \in C} \sum_{j \notin F} q_i R(i,j) \right) / \sum_{i \in B} q_i$$

where B denotes the set of states at the "bottom" of cup C, F denotes the set of states of smallest cost reachable directly from some states of C, and q is a positive solution to

$$q_j = \sum_{i \in C} q_i R(i,j) \quad j \in C$$

An even stronger conjecture is that

$$\lim_{t \to o} -\frac{1}{t} \ln P[\tau_C \geq t e^{d/T}] = \Gamma_C \quad \text{uniformly in } t \geq 0.$$

Now, let us examine a consequence of (3.1) for the annealing algorithms with time varying temperature of the form

$$T_t = \frac{c}{\log t} \qquad t \geq 1$$

We will also use the change of variable $\lambda_t = \exp(-1/T_t)$.

Suppose that $X_{t_0} = i$ for some time t_0 and some state i in a cup C with depth d(C). Suppose that the annealing algorithm is then run for $t \geq t_0$ at the constant temperature T_{t_0}. Then, by (3.1), a typical time t' that the process will escape from C has the form

$$t' = t_0 + L \exp(-d/T_{t_0}),$$

where L is typically O(1). At time t' the actual time-varying annealing schedule (3.1) gives a temperature

$$T_{t'} = \frac{c}{\log (t')} .$$

Writing $\lambda_{t'} = \exp(-1/T_{t'})$, we see that

$$\lambda_{t'}/\lambda_{t_0} = [1+L\lambda_{t_0}^{c-d}]^{1/c}$$

Now, if c > d, this shows that

$$\lambda_{t'} = \lambda_{t_0}(1-\frac{1}{c}L\lambda_{t_0}^{c-d} + o(\lambda^{t_0^{c-d}}))$$

Hence, $\lambda_{t'}$ is very nearly equal to λ_{t_0} so that approximating the schedule by a constant during the interval $[t_0,t']$ should be accurate, and the process should escape C as predicted by the constant-temperature escape time result.

On the other hand, if c < d then (3.1) shows that

$$\lambda_{t_0}/\lambda_{t'} \approx (L\lambda^{t_0^{c-d}})^{1/c}$$

For moderate L and very small λ this ratio is very large. Thus, by the time the process could typically escape the cup (assuming constant temperature T_{t_0}), the actual temperature will be much smaller. This is strong evidence that if c < d, then the cup will, with positive probability, never be exited.

3.3. Escape Time Estimates for the Langevin Algorithm

Consider the Langevin Algorithm run at a fixed temperature T. Suppose that V is thrice continuously differentiable and that a subset D of R^n has the form

$$D = \{x \in R^n : V(x) < k\}$$

and is bounded, connected and has a smooth boundary ∂D. We define d(D), the "depth" of D, by

$$d(D) = k - \min\{V(x) : x \in D\}$$

Proposition Suppose the set D contains at most finitely many zeros of ∇V (or more generally, at most finitely many critical sets of ∇V in the sense of Assumption A, p. 169 of [FrWe85].) Let

$$\tau_D = \inf\{t \geq 0 : X_t \in \partial D\}$$

Then for any x in D

$$\lim_{T \to 0} T \log E_x \tau_D = d(D) \qquad (3.2)$$

Proof This is a special case of Theorem 5.3, pp. 196 of [FrWe84] with $\epsilon = \sqrt{2T}$, which allows non-gradient type vector fields (see [FrWe85, Th. 3.1, p. 118] to help in the specialization).

In view of the similarity of (3.1) and (3.2), the discussion at the end of subsection 3.2 also applies to the Langevin Algorithm, giving strong evidence that the natural analogue of Theorem 1 is true.

Remark We suspect that Assumption A in the proposition is unnecessary and that, moreover the following convergence result is also true: Uniformly over initial states x in compact subsets of D,

$$\tau_D \exp(-d(D)/T)$$

converges in distribution to an exponentially distributed random variable (not necessarily with mean one) as $T \to 0$. Assumption A can likely be lifted through the use of PDE methods (see [Gid85]).

4. MEAN TIME TO FINISH VS. PROBLEM SIZE -- TWO EXAMPLES

4.1. Annealing for a Decoupled System

In this section we consider an annealing system which is a "direct product" of independent systems. Our motivation is that, on the one hand, the system is relatively easy to analyze, while on the other hand, it my be well reflect the behavior of annealing for large systems where coupling is present but weak.

Let S = {a,b,c} with corresponding selection matrix

$$R = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ .5 & .5 & 0 \end{bmatrix}$$

and cost function V' satisfying V'(a)<V'(b)<V'(c). Let u = V'(b)−V'(a) and d = V'(c)−V'(b). Note that state b is a local minimum of V' with depth d. For $N \geq 1$ let S^N denote the set of N-tuples of elements from S, define V^N on S^N by

$$V^N(\underline{s}) = \sum_{i=1}^{N} V'(s_i)$$

and let R^N denote the selection matrix over S^N defined by

$$R^N(\underline{s}, \underline{s'}) = \begin{cases} \frac{1}{N} R(s_i, s'_i) \text{ if } s_j = s'_j \text{ for } j \neq i \\ 0 \text{ if no such i exists} \end{cases}$$

Let $X(t) = (X_1(t), \ldots, X_N(t))$ denote the state of the continuous-time annealing algorithm for (S^N, V^N, R^N) operating at a fixed temperature T. The component processes speeded up by a factor of N, $(X_i(Nt))$, are independent of each other and each is an annealing process for (S', V', R') for fixed temperature T. The equilibrium distribution π of X is thus given by a product of equilibrium probabilities of the individual components. For example, using $\lambda = \exp(-1/T)$, we have

$$\pi((a,a,\ldots,a)) = (1+\lambda^u+2\lambda^{u+d})^{-N}$$

By the theory of Markov chains, this probability is equal to the mean holding time of state (a,a,...,a), $\lambda^{-(d+u)}$, divided by the mean recurrence time of state (a,a,...,a). Thus, the mean recurrence time of state (a,a,...,a) is

$$\frac{1}{\lambda^{u+d}}(1+\lambda^u+2\lambda^{u+d})^N .$$

Subtracting the mean holding time of state (a,a,...,a) from this we obtain the mean time to return to state (a,a,...,a) given that the state has just been exited. By symmetry, this is the expected time,

$$E_1\tau = \frac{1}{\lambda^{u+d}}\{(1+\lambda^u+2\lambda^{u+d})^N - 1\} ,$$

needed to reach state (a,a,...,a) from any of its neighboring states. This is clearly also a lower bound to the mean time needed to reach (a,a,...,a) starting from *any* initial distribution on the set $S^N - \{(a,a,\ldots,a)\}$. This shows that if λ is bounded away from zero as $N \to \infty$, then the mean time to hit the global minimum grows exponentially with N. Thus, at least when annealing at a constant temperature, a lower temperature is needed for larger problems. Now, consider $E_1\tau$ for $\lambda = (\alpha/N)^{\frac{1}{u}}$ in the limit as $N \to \infty$. We get

$$E_1\tau \sim N^{\frac{u+d}{u}} f(\alpha) \quad \text{as} \quad N \to \infty \qquad (4.1)$$

where

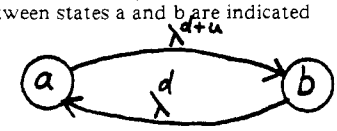
$$f(\alpha) = (e^\alpha - 1)/\alpha^{1+\frac{d}{u}}$$

The fact that

$$f(\alpha) \to \infty \text{ as } \alpha \to 0 \text{ or } \alpha \to \infty$$

indicates that (4.1) represents the smallest asymptotic growth rate possible for $E_1\tau$. This rate is polynomial, rather than exponential, in N.

A chain related to this one is obtained by making state c instantaneous and then taking the N-fold product. Then, the transition rates between states a and b are indicated



For this new model, we have

$$\tilde{E}_1\tau_1 = \frac{2}{\lambda^{u+d}}\{(1+\lambda^u)^N-1\}$$

(which is not significantly different from $E_1\tau$). Moreover, we can show that for any initial distribution μ

$$\tilde{E}_\mu\tau_1 \leq 2\tilde{E}_1\tau + \left\lfloor \frac{2}{\lambda^u} \right\rfloor \log N$$

uniformly in N (proof omitted). We conjecture that a similar bound is true for the original model.

An alternative optimization strategy, called "multistart descent", is to select a state in S^N at random and then run the annealing algorithm at temperature T=0 until a local minimum is reached. This procedure is repeated several times, and the lowest-cost minimum is saved. For this problem a local minimum is a state with each coordinate equal to a or b. Thus, V^N evaluated at a local minimum found by a cycle of the multistart descent algorithm can be expressed as

$$(N-Z)V'(a)+ZV'(b)$$

where Z is the number of coordinates equal to b. Assuming that each state in S^N has the same chances of being the initial state of a cycle, Z is a binomial random variable with parameters $(N,\frac{1}{2})$. Thus, for example, if $V'(a) = 0$, $V'(b) = 1$ and $V'(c) = 2$, then V^N evaluated at the local minimum found by the cycle is equal to Z. By Sterling's approximation, given $0<\alpha<\frac{1}{2}$ and any $\epsilon>0$, there is a constant K_ϵ so that

$$P[Z \leq \alpha N] \leq K_\epsilon \exp(-N(h(\alpha)-\epsilon)) .$$

where

$$h(\alpha) = -\alpha\ln\alpha-(1-\alpha)\ln(1-\alpha)$$

Thus, the expected number of cycles needed to reach a state with cost at most αN is at least

$$K_\epsilon^{-1}\exp(N(h(\alpha)-\epsilon)) ,$$

which grows exponentially with N. We have an example where multistart descent is considerably slower than the annealing algorithm run at a properly chosen constant (in time) temperature. See [LuMe84] for a quite different example.

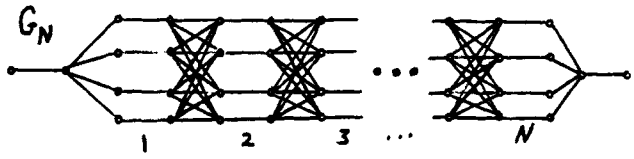
4.2. Maximum Matching by Annealing

A somewhat more complex class of minimization problems is provided by the maximum matching problem in a graph. Let G be an undirected graph with N vertices, and let E denote the set of edges of the graph. A matching is a set of edges, no two of which have a common vertex. Let S denote the set of all matchings for G. We let matchings M and M' be neighbors if their symmetric differences $M\Delta M'$ contains at most one edge. We define

$$R(M,M') = \begin{cases} \dfrac{1}{|E|} \text{ if } |M\Delta M'| = 1 \\ 0 \text{ for other } M' \text{ with } M \neq M' \end{cases}$$

Equivalently, if the current state of the algorithm is a matching M, the potential next matching is chosen by choosing an edge in E at random and either adding or deleting the edge from M.

We define V on S by $V(M) = -|M|$, so that minimizing V is the problem of finding a maximum cardinality matching. This problem is relatively simple in the sense that there are algorithms for solving it using at most $O(N^3)$ computations. Therefore, the following result of my student G. Sasaki is discouraging.



Proposition [Sas85]. There is a $\delta>1$ so that for all N, if G_N is the graph pictured above, if the annealing algorithm starts in a state which is not a maximal matching and if it is run at a constant temperature, then the mean time to find the global minimum is at least δ^N.

We conjecture that this result is true for time-varying temperature schedules. On the positive side, if we restrict attention to fairly sparse graphs, then the annealing algorithm can find almost maximum cardinality matchings in expected polynomial time by annealing at a fixed temperature:

Proposition [Sas85]. There exists a polynomial p with the following property. Let G be any undirected graph with N vertices, let M* denote the cardinality of a maximum cardinality matching of G and let τ denote the maximum degree of vertices in G. Then there exists a temperature (which is a decreasing function of the number of vertices of G) such that the annealing algorithm run at a fixed temperature T finds a matching of cardinality at least

$$M^*\left(1-\frac{2\log(5\tau)}{\log N}\right)$$

in average time at most p(N).

This result is strongest when $\tau << N$ (the case of "sparse graphs"). We feel a stronger result holds for general graphs.

5. THE ROLE OF STATISTICAL MECHANICS

The simulated annealing algorithm is a decedent of work in statistical mechanics. Can statistical mechanics help us further? Perhaps the main use for statistical mechanics concepts is to predict the "typical" behavior of the annealing algorithm when applied to "typical" large problems. Of course, its behavior on different problems may vary drastically -- but it may be desirable to identify several types of typical behavior. We first summarize some of the work of White [Whi84].

Consider a large system. We postulate that there is an approximate "density of states" w.

$$w(E)dE \approx \#\{\text{states with energy in } [E,E+dE]\}$$

We make the following postulate

Postulate A. Away from regions of extremely high or low energy, the density of states is approximately Gaussian;

$$w(E) \propto \exp(-(E-\bar{E})^2/2\Delta)$$

where $\bar{E}$ and Δ are constants. The density of states observed at temperature T is thus

$$\sim e^{-(E-\bar{E})^2/2\Delta} e^{-E/T} \qquad \sim e^{-(E-2E(\bar{E}-\frac{\Delta^2}{T}))/2\Delta}$$

so

$$< E(T) > \; = \bar{E}-\frac{\Delta^2}{T}, \; < E(T)^2 > - < E(T) >^2 = \Delta^2$$

Note that $< E(T) >$ approaches $\bar{E}$ as $T\to 0$. When $T = \Delta$ then $< E(T) >$ is within one standard deviation of $\bar{E}$, which suggests that the system "looks almost" like a $T = +\infty$ system at temperature $T = \Delta$.

A second postulate describes behavior at low energies.

Polstulate B. There is a minimum possible energy E_0 and the density of states near E_0 is given by

$$w(E) = \exp((E-E_0)\gamma) \quad E \geq E_0$$

for some constant $\gamma>0$.

Postulate B is roughly true, for example, if the smallest possible energies $E_0<E_1<E_2<\ldots$ are equally spaced with about M^j states with energy E_j. This would lead to

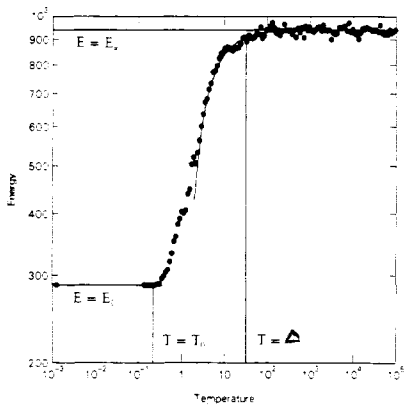
$$w(E) \propto M^{E/E_1-E_0} \text{ or } \gamma = \frac{\ln M}{E_1-E_0}$$

Under Postulate B, the probability density of states observed at temperature T in equilibrium,

$$(\text{const.})w(E)\exp(-E/T),$$

will be increasing as E tends down to E_0 if $T \lesssim 1/\gamma$. Thus, $T_0 = 1/\gamma$ is roughly the temperature at which equilibrium states will tend to be minimum energy states.

Combining these descriptions for middle and low range energies, we find that the mean observed energy versus temperature should follow the curve shown.



"Typical" log-log plot of mean observed energy in equilibrium vs. temperature T.

From a theoretical point of view, a way to give meaning to statements about "typical behavior" is to study large random problems. For example, spin glass theory [ThAP77] considers the random cost function

$$V(\underline{s}) = \underline{s}^T W \underline{s} \quad \underline{s} \in \{1,-1\}^N$$

where the components of W are independent mean-zero Gaussian random variables with $EW_{ij}^2 = J/2N$ for $i \neq j$ and $W_{ii} \equiv 0$. Since the cost function is random, so are the sets of local and global minima and their costs. It is not hard to show, for example, that with high probability V has roughly $2^{.2N}$ local minima [BrMo80]. It is generally agreed that the minimum of V is typically cN, but there is still some controversy regarding the value of the constant c.

Combinatorial problems on a random graph (see [Pal85]) are closely related (they often correspond to different distributions on W) and provide more examples in which some progress has been made regarding the distribution of minimum cost states. I also think analysis of distributed algorithms such as that of [Lub85] is relevant. Some work in statistical mechanics and mathematical statistical mechanics is concerned with rates of convergence to equilibrium [Hol85a,Hol85b] and with large deviations of large regular systems of locally interacting particles [Isr79]. This may prove useful in research on simulated annealing. Finally, the "mean field" approximation, as used in [ThAP77] for example, is itself closely related to a different optimization technique proposed in [Hop84].

6. APPLICATIONS

Simulated annealing has been applied to many large optimization problems. It has been used to

- find estimates of noisy images which maximize a Bayes cost [GeGe84]
- find placements of devices and wire routings for VLSI chips [KiGV83,VeKi83,Whi84]
- discover new combinatorial constants, such as sphere-packing bounds for binary sequences (a topic in coding theory) [HeWe84]
- solve instances of difficult combinatorial optimization problems [Cer82,GiGV83,Bol84,GrSu84]
- generate samples distributed according to an equilibrium distribution for low temperatures [HiSA84] (The purpose here is not to find a global minimum.)

Researchers find that the annealing algorithm can find near-minimum solutions, but it can be very slow. Even though it can sometimes be speeded up by parallel implementation, no serious real-time applications have emerged. It is clear in numerous cases that simulated annealing has many worthy competitors in the form of other heuristics for hard problems [AJMS84].

ACKNOWLEDGEMENTS

I'm grateful to Stuart and Donald Geman, Basilis Gidas, Petar Kokotovic, Harold Kushner, Galen Sasaki and Steve White for useful discussions regarding simulated annealing. This work was supported by the Office of Naval Research under contract NOOO14-82-K-0359 and the National Science Foundation under contract NSF-ECS-83-52030.

REFERENCES

[AJMS84] C. R. Aragon, D. S. Johnson, L. A. McGeoch, C. Schevon "Optimization by simulated annealing: an experimental evaluation" Draft and viewgraphs, November 1984.

[BoL84] E. Bonomi, and J.-L. Lutton "The N-city travelling salesman problem: statistical mechanics and the Metropolis algorithm," SIAM Review, 26, pp. 551-568, 1984.

[BrMo80] A. J. Bray and M. A. Moore "Metastable states in spin glasses," J. Phys. C, 13, pp. L469-476, 1980.

[Cer82] V. Cerny, "A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," preprint, Inst. of Phys. and Biophysics, Comenius Univ., Bratislava, 1982.

[ChHS85] T.-S. Chiang, C.-R. Hwang and S.-J. Sheu "Diffusion for global optimization in $\mathbb{R}^n$," Preprint, Institute of Mathematics, Academia Sinica, Taipei, Taiwan, Received August 1985.

[CoWS83] M. Coderch, A. S. Willski and S. S. Sastry "Hierarchical aggregation of singularly perturbed finite state Markov chains," Stochastics, 1983.

[DeMQ85] F. Delebeque, O. Muron and J.P. Quadrat "Singular perturbation of Markov chains," To appear in a Springer book as a chapter (1985).

[FrWe84] M. I. Freidlin and A. D. Wentzell "Random perturbations of dynamical systems," Springer-Verlag, New York, 1984.

[GeGe84] S. Geman and D. Geman "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 6, pp. 721-741, November 1984.

[GeHw84] S. Geman and C.-R. Hwang "Diffusions for global optimization," Preprint, Div. of Applied Math, Brown University, Received December 5, 1984.

[Gid84] B. Gidas "Non-stationary Markov chains and convergence of the annealing algorithm," J. Stat. Phys., 39, pp. 73-131, 1985.

[Gid85] B. Gidas "Global minimization via the Langevin equation," Proc. IEEE Conf. on Decision and Control, 1985.

[GrSu84] J.W. Greene and K. J. Supowit "Simulated annealing without rejected moves", Proc. IEEE Int. Conf. on Computer Design, 6 pages, October 1984.

[Haj85] B. Hajek, "Cooling schedules for optimal annealing" Preprint, January 1985. Submitted to Mathematics of Operations Research.

[HeWe84] L. A. Hemachandra and V.K. Wei "Simulated annealing and error correcting codes" Preprint, Bell Communications Research, Murray Hill, April 1984.

[HiSA84] G. E. Hinton, T.J. Sejnowski and D. H. Ackley, "Boltzman machines: constraint satisfaction networks that learn," Carnegie-Mellon Technical Report CMU-CS-84-119, 1984.

[Hol85a] R. Holley "Rapid convergence to equilibrium in one dimensional stochastic Ising models," Annals of Probability, 13, pp. 72-89, 1985.

[Hol85b] R. Holley "Possible rates of convergence in finite range attractive systems," Preprint.

[Hop84] J. J. Hopfield "Neurons with graded response have collective computational properties like those of two-state neurons," Proc. Natl. Acad. Sci., USA, 81, pp. 3088-3092, May 1984.

[Isr79] R. B. Israel, *Convexity in the theory of lattice gases*, Princeton University Press, 1979.

[Kat76] T. Kato *Perturbation theory for Linear Operators*, Springer Verlag, 1976.

[KiGV83] S. Kirkpatrick, C. D. Gelett and M. P. Vecchi "Optimization by simulated annealing," Science 220, May 13, pp. 621-680, 1983.

[Kus85] H. J. Kushner, "Asymptotic global behavior for stochastic approximations and diffusions with slowly decreasing noise effects," Lefschetz Center for Dynamical Systems Report No. 85-7, Brown University, April 1985.

[Pal85] E. Palmer, *Graphical Evolution*, Wiley Interscience, 1985.

[Lub85] M. Luby, "A simple parallel algorithm for the maximal independent set problem."

[LuMe84] M. Lundy and A. Mees "Convergence of the annealing algorithms," Preprint, Cambridge University

[MRRTT53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller "Equations of state calculations by fast computing machines," J. Chem. Phys., 21, pp. 1087-1091, 1953.

[MiRS85] D. Mitra, F. Romeo and A. Sangiovanni-Vancentelli "Convergence and finite-time behavior of simulated annealing," Proc.IEEE Conf. on Decision and Control, 1985.

[Sas85] G. H. Sasaki, Personal communication.

E. Seneta "Non-negative matrices and Markov chains," Springer-Verlag, New York, 1981.

[ThAP77] D. J. Thouless, P. W. Anderson and R. G. Palmer "Solution of 'Solvable model of a spin glass'," Philosophical Mag., 35, pp. 593-601, 1977.

[Tsi85] J.N. Tsitsiklis, "Markov chains with rare transitions and simulated annealing," Preprint, MIT Laboratory for Information and Decision Systems, August 1985.

[VeKi83] M. P. Vecchi and S. Kirkpatrick "Global wiring by simulated annealing," IEEE Trans. on Computer-Aided Design, 2, pp. 215-222, October 1983.

[Whi84] S. White "Concepts of scale in simulated annealing," Proc. IEEE Int. Conf. on Computer Design, pp. 646-651, October 1984.