What is Simulated Annealing?

Michael W. Trosset*

February 26, 2000

Abstract

Beginning in 1983, simulated annealing was marketed as a global optimization methodology that mimics the physical annealing process by which molten substances cool to crystalline lattices of minimal energy. This marketing strategy had a polarizing effect, attracting those who delighted in metaphor and alienating others who found metaphor insufficient at best and facile at worst. In fact, the emotional outbursts that accompany many discussions of simulated annealing are an unfortunate distraction. Whatever its pros and cons, simulated annealing can be grounded in rigorous mathematics. Here we provide an elementary, self-contained introduction to simulated annealing in terms of Markov chains.

Contents

*Associate Professor, Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (e-mail: `trosset@math.wm.edu`); and Adjunct Associate Professor, Department of Computational & Applied Mathematics, Rice University, 6100 Main Street, Houston, TX 77005-1892 (e-mail: `trosset@caam.rice.edu`).

1 Introduction

The publication of "Optimization by simulated annealing" (Kirkpatrick, Gelatt, and Vecchi, 1983), in which was advertised a global optimization strategy that mimics the physical annealing process by which molten substances cool to crystalline lattices of minimal energy, had a remarkable impact. The annealing metaphor was enthusiastically embraced by large numbers of scientists and engineers—seventeen years later, it is one of the few optimization strategies that are widely known and used outside the optimization community. Another is the genetic algorithm(s), which also boasts an appealing metaphor. Some of the tone of the campaigns to market these strategies can be glimpsed in a passage from a recent book on the latter. Writing under the heading "Natural Optimization Methods," Haupt and Haupt (1998, p. 16) enthused:

> "All hope is not lost! Some outstanding algorithms have surfaced in the last 30 years. Two relatively new ones are the genetic algorithm and simulated annealing. The genetic algorithm models natural selection and evolution, while simulated annealing models the annealing process. ... They represent processes in nature that are remarkably successful at optimizing natural phenomena."

The optimization community has not been amused. Baffled by the success of a marketing campaign based on metaphor, perhaps vaguely threatened by the popularity of methods devised outside their discipline, many researchers have dismissed "natural optimization methods"—possibly with good reason. Do simulated annealing and/or genetic algorithms *really* represent nature? And what if they do? Just how successful is nature anyhow? In the minds of these researchers, what has really surfaced are some outstanding bedtime stories.

And yet, if one should refrain from embracing a method because it has a pleasing metaphor, neither should one reject it for the same reason. A story may please and yet be a great work of literature. Those with open minds will endeavor to penetrate beyond metaphor and understand the method itself. At least in the case of simulated annealing, one discovers rigorous mathematics if only one looks. Once one comprehends it, one can begin a dispassionate assessment of simulated annealing's pros and cons.

Let us be clear about what this article is not. It is neither a survey of research on simulated annealing nor a tutorial on how to use it. We will not attempt to assess the efficacy of simulated annealing, although we will eventually make some remarks that reveal some of our own prejudices. Our goal is simply to provide an elementary, self-contained introduction to simulated annealing that is grounded in mathematics rather than metaphor. Once that has been accomplished, an informed reader can objectively use, test, discard, or borrow ideas from simulated annealing as he or she sees fit.

We also emphasize that the perspective that we have adopted is not new. This description of simulated annealing is often presented in the simulation literature, as in the books by Ripley (1987), Fishman (1996), and Robert and Casella (1999). However, these books have broad agendas that are outside the mainstreams of the optimization and engineering communities. Accordingly, it was thought that a succinct summary of the salient issues might help to demystify the subject for those audiences.

2 Motivation

We begin by considering a directed graph with three nodes. Let p_{ij} denote the length of the edge that begins at node i and ends at node j. If no such edge exists, then we set $p_{ij} = 0$. Thus, the

edge lengths can be collected in a matrix $P = [p_{ij}]$, e.g.

$$P = \begin{bmatrix} 0 & .50 & .50 \\ 0 & .25 & .75 \\ 1 & 0 & 0 \end{bmatrix}. \tag{1}$$

The P that we have constructed has the interesting property that the sum of each row (the sum of the lengths of the edges emanating from each node) is unity. Thus, each row of P is a vector of proportions or probabilities. This leads to two instructive interpretations of the directed graph, one deterministic and one stochastic:

1. Imagine that each node of the directed graph is a bucket. Together, the buckets contain one unit of water. How this unit of water is distributed among the buckets varies in time. Let $\pi_t(i)$ denote the proportion of the water in bucket i at time $t = 0, 1, 2, 3, \ldots$. To obtain π_{t+1}, the vector of proportions at time $t+1$, from π_t, we simultaneously pour $p_{ij}\pi_t(i)$ units of water from bucket i into bucket j. Thus, the edge lengths specify the proportion of bucket i's water that will be poured into bucket j at each time step.

   There are many interesting questions that can be asked about this experiment. One question in which we will be especially interested is the following:

   > Does $\lim_{t\to\infty} \pi_t$ exist, i.e. do the water levels eventually stabilize?

2. Imagine that each node of the directed graph is a lilypad, on one of which sits a frog. If the frog occupies lilypad i at time t, then the frog leaps to lilypad j with probability p_{ij}. In this interpretation, $\pi_t(i)$ represents the probability that the frog occupies lilypad i at time t. Again we may ask:

   > Does $\lim_{t\to\infty} \pi_t$ exist, i.e. do the proportions of time that the frog spends on each lilypad eventually stabilize?

The questions that we have posed turn out to be crucial to our understanding of simulated annealing. Two further questions are the following:

> If $\lim_{t\to\infty} \pi_t$ exists, then does it depend on π_0, the initial distribution of the water levels or the frog's initial lilypad?
>
> If $\lim_{t\to\infty} \pi_t$ exists, then how can we compute it?

The experiments that we have described are popular interpretations of *Markov chains*, the study of which is devoted to answering questions such as those we have posed. In the language of Markov chains, nodes are *states* and edge lengths are *transition probabilities*. A visual display of the directed graph is a *state diagram*. We now present a more detailed introduction to Markov chains.

3 Markov Chains

Let us remark on several distinctive features of the experiments described in Section 2:

1. The number of states (buckets, lilypads) is finite.
2. Time is discrete ($t = 0, 1, 2, \ldots$), rather than continuous ($t \in [0, \infty)$).

3. The future depends only on the present, not on the past. For example, where the frog jumps next depends on where the frog is now, but not on where the frog was previously.

4. The transition probabilities (that describe, for example, the frog's behavior) do not change with time.

More generally, let D denote a finite set of n states and let $\{X_t : t = 0, 1, 2, \ldots\}$ denote a sequence of random variables with values in D. Assume the *Markov property*, that

$$P\left(X_{t+1} = i_{t+1} | X_0 = i_0, \ldots, X_t = i_t\right) = P\left(X_{t+1} = i_{t+1} | X_t = i_t\right).$$

Let

$$p_{ij}(t) = P\left(X_{t+1} = j | X_t = i\right)$$

denote the transition probabilities and write $P(t) = [p_{ij}(t)]$. In this section, we will assume that $P(t) \equiv P$ does not depend on t. Then $\{X_t\}$ is a *discrete-parameter finite Markov chain with stationary transition probabilities.*

Returning to our example, let us suppose that $\pi_0 = (1, 0, 0)'$, which means that the frog begins on lilypad 1 or that all of the water begins in bucket 1. Then half of the water is poured into bucket 2 and the other half is poured into bucket 3, resulting in $\pi_1 = (0, .5, .5)'$. Next, the water in bucket 3 is returned to bucket 1 and three fourths of the water in bucket 2 is poured into bucket 3, resulting in $\pi_2 = (.500, .125, .375)'$.

The next transition is more complicated—it will help us to discern the general pattern of the calculations if we express the new amount in each bucket as the sum of the contributions from the three buckets:

$$\begin{aligned}
\pi_3(1) &= p_{11} \cdot \pi_2(1) + p_{21} \cdot \pi_2(2) + p_{31} \cdot \pi_2(3) \\
&= 0 \cdot .500 + 0 \cdot .125 + 1 \cdot .375 \\
&= .375, \\
\pi_3(2) &= p_{12} \cdot \pi_2(1) + p_{22} \cdot \pi_2(2) + p_{32} \cdot \pi_2(3) \\
&= .50 \cdot .500 + .25 \cdot .125 + 0 \cdot .375 \\
&= .28125, \\
\pi_3(3) &= p_{13} \cdot \pi_2(1) + p_{23} \cdot \pi_2(2) + p_{33} \cdot \pi_2(3) \\
&= .50 \cdot .500 + .75 \cdot .125 + 0 \cdot .375 \\
&= .34375.
\end{aligned}$$

This can be expressed more succinctly in matrix notation as $\pi_3 = P'\pi_2$.

More generally, let π_t denote the distribution of X_t; then

$$\pi_t(j) = P\left(X_t = j\right) = \sum_{i=1}^{n} P\left(X_t = j | X_{t-1} = i\right) P\left(X_{t-1} = i\right) = \sum_{i=1}^{n} p_{ij} \pi_{t-1}(i).$$

In matrix notation, we have $\pi_t = P'\pi_{t-1}$, and by recursion we obtain

$$\pi_t = \left(P'\right)^t \pi_0.$$

Thus, the distribution of any X_t is completely determined by the distribution of X_0.

Now it may be that the water levels in our buckets are such that pouring water in the prescribed proportions does not change how much each bucket contains. This observation motivates

Definition 1 *A distribution π is stationary if it satisfies the Chapman-Kolmogorov equations, $\pi = P'\pi$.*

Returning to our example, we can identify a stationary distribution by rewriting the Chapman-Kolmogorov equations as the linear system $(P' - I)\pi = 0$. We seek a solution π_* with nonnegative components that sum to unity. We find such a solution by augmenting the Chapman-Kolmogorov equations with the summation constraint, then solving the augmented linear system and verifying that the solution is nonnegative. Thus, we solve $A\pi = b$, where

$$A = \begin{bmatrix} 0-1 & 0 & 1 \\ .5 & .25-1 & 0 \\ .5 & .75 & 0-1 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

and obtain the stationary distribution $\pi_* = (.375, .250, .375)'$.

Notice that, although the initial distribution $\pi_0 = (1, 0, 0)'$ does not resemble the stationary distribution π_*, the distribution $\pi_3 = (.37500, .28125, .34375)'$ does. This observation recalls the questions that we posed in Section 2, about the behavior of π_t as $t \to \infty$. In fact, we have the following result:

Theorem 1 *Suppose that a discrete-parameter finite Markov chain with stationary transition probabilities has the following properties:*

1. *It is possible to get from any state i to any state j in a finite number of transitions.*
2. *Choose any state i and consider the possible numbers of transitions required to pass from i back to i. Then the greatest common factor of these numbers is* 1.

Then there exists a unique stationary distribution π_∞ and, from any initial distribution π_0, $\pi_t \to \pi_\infty$ as $t \to \infty$.

The first property in Theorem 1 is often described by saying that the states *communicate*. In our example, it is obvious that the frog can reach any lilypad from any lilypad. The second property in Theorem 1 is often described by saying that the Markov chain is *aperiodic*. In our example, suppose that the frog is sitting on lilypad $i = 1$. Then the frog could jump to lilypad 3 and back to lilypad 1, a total of 2 transitions. Alternatively, the frog could jump to lilypad 2, then to lilypad 3, and then to lilypad 1, a total of 3 transitions. Because the greatest common factor of 2 and 3 is 1, this Markov chain is aperiodic.

For our purposes, it will suffice to consider the case of strictly positive transition probabilities, i.e. the case that each $p_{ij} > 0$. The theory for this case is especially elegant.

First, we note that the states communicate and the Markov chain is aperiodic. Hence, there exists a unique stationary distribution π_∞ and, for any initial distribution π_0,

$$\lim_{t\to\infty} \pi_t = \lim_{t\to\infty} \left(P'\right)^t \pi_0 = \pi_\infty.$$

Furthermore, π_∞ is an eigenvector of P' that corresponds to an eigenvalue of 1. Finally, if $\delta = \min\{p_{ij} : i, j \in D\}$, then $\pi_\infty(j) \geq \delta$ and

$$\left| \left(P^t\right)_{ij} - \pi_\infty(j) \right| \leq (1 - n\delta)^{t-1}.$$

For any initial distribution π_0, it follows that

$$\begin{aligned}
|\pi_t(j) - \pi_\infty(j)| &= \left| \sum_{i=1}^{n} \left(P^t\right)_{ij} \pi_0(i) - \pi_\infty(j) \right| \\
&= \left| \sum_{i=1}^{n} \left(P^t\right)_{ij} \pi_0(i) - \sum_{i=1}^{n} \pi_\infty(j)\pi_0(i) \right| \\
&= \left| \sum_{i=1}^{n} \left[\left(P^t\right)_{ij} - \pi_\infty(j) \right] \pi_0(i) \right| \\
&\leq \sum_{i=1}^{n} \left| \left(P^t\right)_{ij} - \pi_\infty(j) \right| \pi_0(i) \\
&\leq \sum_{i=1}^{n} (1 - n\delta)^{t-1} \pi_0(i) \\
&= (1 - n\delta)^{t-1}.
\end{aligned}$$

Thus, we can actually estimate how rapidly π_t converges to π_∞.

Now we consider a real-valued function, $g : D \to \Re$. Let

$$I = \sum_{x \in D} g(x)\pi_\infty(x) = \int_D g(x)\pi_\infty(dx) = E_\infty g(X)$$

and

$$\hat{I}_N = \frac{1}{N} \sum_{t=1}^{N} g\left(X_t\right).$$

We emphasize that I is a real number and that $\{\hat{I}_N\}$ is a sequence of random variables. The two are related by the following results, adapted from Doob (1953, Chapter V):

Theorem 2 *(Strong Law of Large Numbers) If a discrete-parameter finite Markov chain has strictly positive stationary transition probabilities, then $\hat{I}_N$ converges almost surely to I as $N \to \infty$.*

Theorem 3 *(Central Limit Theorem) If a discrete-parameter finite Markov chain has strictly positive stationary transition probabilities, then there exists $\sigma^2 \in \Re$ such that*

$$\lim_{N \to \infty} E_\infty \left[\sqrt{N} \left(\hat{I}_N - I \right) \right]^2 = \sigma^2.$$

Furthermore, if $\sigma^2 > 0$, then $\sqrt{N}(\hat{I}_N - I)$ converges in distribution to a normal random variable with mean 0 and variance σ^2.

Thus $\hat{I}_N$ is a consistent estimator of I. It is also possible to estimate σ^2 and thereby construct asymptotic confidence intervals for I. Because I is a definite integral, these results suggest a relation between integration and the asymptotic behavior of Markov chains. We will explore that relation in greater detail in Section 4.

4 Integration

Suppose that we are given the following:

- a finite set D;
- a probability measure π on D such that $\pi(x) > 0$ for every $x \in D$; and
- a function $g : D \to \Re$.

We would like to integrate g with respect to π, i.e. we want to evaluate

$$I = \sum_{x \in D} g(x)\pi(x) = \int_D g(x)\pi(dx) = E_\pi g(X).$$

If we could construct a Markov chain $\{X_t\}$ on D in such a way that $\pi_t \to \pi$ as $t \to \infty$, then—applying Theorems 2 and 3—we could estimate I by $\hat{I}_N$. Even better, we could discard m initial observations, allowing the Markov chain some time to approach equilibrium, and estimate I by

$$\hat{I}_{N-m} = \frac{1}{N-m} \sum_{t=m}^{N} g\left(X_t\right).$$

In fact, as discussed by Hastings (1970), various such constructions *are* possible. The best known is due to Metropolis et al. (1953).

The Metropolis Algorithm To construct a suitable Markov chain, let Q be *any* symmetric matrix of strictly positive transition probabilities on D. Let

$$\alpha_{ij} = \min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\},$$

let $p_{ij} = \alpha_{ij} q_{ij}$ for $i \neq j$, and let $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. Then $P = [p_{ij}]$ is a matrix of transition probabilities on D with the interpretation that we accept a proposed transition from state i to state j with probability α_{ij}. More specifically,

- We always accept the proposed transition if $\pi(j) \geq \pi(i)$, i.e. if state j is at least as probable as state i.
- Otherwise, we accept the proposed transition with probability equal to the odds ratio of state j to state i. For example, if state j is much less probable than state i, then we are unlikely to accept the proposed transition.

With this construction, if $j \neq i$, then

$$\begin{aligned}
\pi(i)p_{ij} &= \pi(i)\alpha_{ij}q_{ij} \\
&= \pi(i)\min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\} q_{ij} \\
&= \left\{\begin{array}{ll} \pi(j)q_{ij} & \text{if} \quad \pi(j) \leq \pi(i) \\ \pi(i)q_{ij} & \text{if} \quad \pi(j) \geq \pi(i) \end{array}\right\} \\
&= \left\{\begin{array}{ll} \pi(i)q_{ji} & \pi(i) \leq \pi(j) \\ \pi(j)q_{ji} & \pi(i) \geq \pi(j) \end{array}\right\} \\
&= \pi(j)\min\left\{\frac{\pi(i)}{\pi(j)}, 1\right\} q_{ji} \\
&= \pi(j)\alpha_{ji}q_{ji} \\
&= \pi(j)p_{ji}.
\end{aligned}$$

It follows that

$$\sum_i \pi(i)p_{ij} = \sum_{i \neq j} \pi(i)p_{ij} + \pi(j)p_{jj} = \sum_{i \neq j} \pi(j)p_{ji} + \pi(j) - \pi(j) \sum_{i \neq j} p_{ji} = \pi(j),$$

or more succinctly that $\pi' P = \pi'$. Thus, π satisfies the Chapman-Kolmogorov equations and $\pi_t \to \pi$ from any initial distribution π_0.

The Metropolis algorithm allows us to estimate I by observing the states assumed by the X_t and averaging the corresponding values of $g(X_t)$. For future reference, we note that the Metropolis algorithm depends only on the ratios $\pi(j)/\pi(i)$.

5 Global Optimization

By now even the most patient reader must be wondering what Sections 2–4 have to do with optimization. In this section we establish a connection between optimization and integration. The reader may well be skeptical of this connection, for quadrature is widely regarded as more difficult than optimization. Nevertheless, for better or for worse, this connection is the basis for simulated annealing.

The Discrete Case We first suppose that D is a finite subset of n points in $\Re^k$.

Theorem 4 *If x_* is the unique global minimizer of $f : D \to \Re$, then*

$$x_* = \lim_{\lambda \to \infty} \frac{\sum_{x \in D} x \exp\left[-\lambda f(x)\right]}{\sum_{x \in D} \exp\left[-\lambda f(x)\right]}.$$

Proof: Let

$$I_1 = \sum_{x \in D} |x - x_*| \exp\left[-\lambda f(x)\right]$$

and

$$I_2 = \sum_{x \in D} \exp\left[-\lambda f(x)\right].$$

It evidently suffices to prove that $I_1/I_2 \to 0$ as $\lambda \to \infty$.

Let $D^- = D \setminus \{x_*\}$. Because x_* is the unique global minimizer of f, there exists $\delta > 0$ such that $f(x) \geq f(x_*) + \delta$ for every $x \in D^-$. Then

$$I_1 = \sum_{x \in D^-} |x - x_*| \exp\left[-\lambda f(x)\right] \leq (n-1) \cdot \max\left\{|x - x_*| : x \in D^-\right\} \cdot \exp\left(-\lambda\left[f\left(x_*\right) + \delta\right]\right)$$

and

$$I_2 \geq \exp\left[-\lambda f\left(x_*\right)\right].$$

Hence, there exists $M < \infty$ for which

$$0 \leq \frac{I_1}{I_2} \leq M \exp(-\lambda\delta) \to 0$$

as $\lambda \to \infty$. □

The Continuous Case The case of continuous optimization is beyond the scope of our elementary exposition. Nevertheless, for historical reasons, we state the continuous version of Theorem 4.

Theorem 5 *(Pincus, 1968) Suppose that $D \subset \Re^k$ is compact and that $f : D \to \Re$ is continuous. If x_* is the unique global minimizer of f in D, then*

$$x_* = \lim_{\lambda \to \infty} \frac{\int_D x \exp\left[-\lambda f(x)\right] dx}{\int_D \exp\left[-\lambda f(x)\right] dx}.$$

Summary Let π_λ denote the probability density function on D defined by

$$\pi_\lambda(x) \propto \exp\left[-\lambda f(x)\right].$$

Then

$$\pi_\lambda(x) = \frac{\exp\left[-\lambda f(x)\right]}{\sum_{y \in D} \exp\left[-\lambda f(y)\right]}$$

in the discrete case and

$$\pi_\lambda(x) = \frac{\exp\left[-\lambda f(x)\right]}{\int_D \exp\left[-\lambda f(y)\right] dy}$$

in the continuous case. We can summarize the global optimization criteria in Theorems 4 and 5 by writing

$$x_* = \lim_{\lambda \to \infty} E_\lambda X,$$

where X is a random vector with values in D and E_λ denotes expectation with respect to π_λ.

6 Simulated Annealing

We now describe the heuristics of simulated annealing in the case of discrete optimization. (The case of continuous optimization requires replacing finite-state Markov chains with continuous-state Markov processes, with attendant difficulties.) For $\lambda > 0$ fixed, we can approximate

$$I(\lambda) = E_\lambda X = \sum_{x \in D} x \pi_\lambda(x) = \sum_{x \in D} x \frac{\exp\left[-\lambda f(x)\right]}{\sum_{y \in D} \exp\left[-\lambda f(y)\right]}$$

by using the Metropolis algorithm to construct a Markov chain on D that has π_λ as its limiting distribution, then estimating $I(\lambda)$ by observing $\hat{I}_{N-m}(\lambda)$ as $N \to \infty$. Notice that this construction does *not* require knowledge of the "integration constant"

$$\sum_{y \in D} \exp\left[-\lambda f(y)\right],$$

because the Metropolis algorithm only requires the ratios

$$\frac{\pi_\lambda(j)}{\pi_\lambda(i)} = \frac{\exp\left[-\lambda f(j)\right]}{\exp\left[-\lambda f(i)\right]} = \exp\left(-\lambda\left[f(j) - f(i)\right]\right).$$

The Metropolis algorithm does require the user to specify transition probabilities $q_{ij} > 0$. Given a current iterate $x_c = i \in D$, the Metropolis algorithm randomly samples a trial iterate $x_d = j \in D$ from the probability distribution $(q_{i1}, \ldots, q_{in})'$ on D. If $\pi_\lambda(x_d) \geq \pi_\lambda(x_c)$, i.e. if $\exp[-\lambda f(x_d)] \geq \exp[-\lambda f(x_d)]$ or equivalently if $f(x_d) \leq f(x_c)$, then the next iterate is the trial iterate: $x_+ = x_d$. Otherwise, if $f(x_d) > f(x_c)$, then we set $x_+ = x_d$ with probability $\exp(-\lambda[f(j) - f(i)])$ or $x_+ = x_c$ with probability $1 - \exp(-\lambda[f(j) - f(i)])$.

For λ fixed, $\hat{I}_{N-m}(\lambda) \to I(\lambda)$ as $N \to \infty$. Since $I(\lambda) \to x_*$ as $\lambda \to \infty$,

The basic idea of simulated annealing is to observe the asymptotic behavior of X_t, using λ as a continuation parameter.

We are thus led to study the behavior of nonstationary Markov chains in which the transition probabilities vary with λ. Further theory, in which the critical issue is the nature of $\lambda(t)$, is then required to establish convergence to x_*. Examples include Geman and Geman (1984), Gidas (1985), and Hajek (1986, 1988). This theory is beyond the scope of our exposition, but we hope that we have persuaded the reader that such theories can be based on mathematics rather than on metaphor.

7 Discussion

The physical interpretation of simulated annealing derives from statistical mechanics. Let $f(x) = U(x)$ denote the potential energy of a configuration x. Let

$$\lambda = \frac{1}{kT},$$

where T denotes temperature and k is the Boltzmann constant. Then π_λ is the Boltzmann probability density function for the state of the configuration at temperature T. The global minimum energy configuration is obtained by letting $T \to 0$, i.e. by cooling. Thus, the $\lambda(t)$ that arises in simulated annealing is the "cooling schedule".

Since the widely cited article by Kirkpatrick, Gelatt, and Vecchi (1983), the physical interpretation of simulated annealing has been a remarkably effective advertisement for its use. We submit that a healthier perspective is that the convergence theory for simulated annealing illuminates the natural phenomenon of physical annealing. Consider, for example, the tremendously important issue of how rapidly simulated annealing converges to the global minimizer. The Metropolis transition probabilities are

$$p_{ij} = \min\left\{\exp\left(-\lambda\left[f(j) - f(i)\right]\right), 1\right\} \cdot q_{ij}.$$

Upon letting $i = x_*$ and $j \neq x_*$, it is evident that

$$\delta = \min\left\{p_{ij} : i, j \in D\right\} \to 0$$

as $\lambda \to \infty$. Recalling that

$$|\pi_t(x) - \pi_\infty(x)| \leq (1 - n\delta)^{t-1},$$

we might suspect that the number of transitions required for $\hat{I}_{N-m}(\lambda)$ to accurately estimate $I(\lambda)$ will increase rapidly as $\lambda \to \infty$. In fact, this phenomenon is widely appreciated. The theoretical guarantees for simulated annealing can be realized only with cooling schedules so slow that they are rarely used in practice. Perhaps not surprisingly, it is also a well-known characteristic of physical annealing that the rate of convergence to equilibrium slows dramatically as temperature decreases.

The slow convergence of physical annealing upends the metaphor that commends the use of simulated annealing for global optimization. Furthermore, from the perspective of the academic optimizer, many important questions remain unanswered. In numerical optimization, algorithms for finding local minimizers perform very differently depending on how one chooses trial iterates. In the context of simulated annealing, the choice of Q, the transition probabilities for sampling a trial iterate x_d given the current iterate x_c, is surely of compelling interest. And simulated annealing, which relies exclusively on observed values of the objective function, seems especially suspect for continuous optimization.

Our view is that *no* general purpose algorithm should be expected to efficiently find global minimizers of nonconvex continuous optimization problems. *We submit that efficient global optimization is generally impossible*—see, for example, Stephens and Baritompa (1998) and Anonymous (1972). But this does not mean that one should not endeavor to devise algorithms for specific global optimization problems. In doing so, one may do well to exploit the ideas that underlie existing methods. The ideas that underlie simulated annealing are mathematically rigorous and may ultimately serve even the most vocal critics.

Acknowledgments

This exposition of simulated annealing is based on lectures delivered at Rice University in December 1996 and January 1997. I thank the members of Richard Tapia's research group for their interest and Amr El-Bakry for inviting me to reprise my presentation. I also thank Rex Kincaid for reading a draft manuscript and making helpful suggestions.

References

Anonymous (1972). A new algorithm for optimization. *Mathematical Programming*, 3:124–128.

Doob, J. L. (1953). *Stochastic Processes*. John Wiley & Sons, New York. Wiley Classics Library edition published in 1990.

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gidas, B. (1985). Nonstationary Markov chains and the convergence of the annealing algorithm. *Journal of Statistical Physics*, 39:73–131.

Hajek, B. (1986). Optimization by simulated annealing: A necessary and sufficient condition for convergence. In Van Ryzin, J., editor, *Adaptive Statistical Procedures and Related Topics*, pages 417–427. Institute of Mathematical Statistics, Hayward, CA.

Hajek, B. (1988). Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13:311–329.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Haupt, R. L. and Haupt, S. E. (1998). *Practical Genetic Algorithms*. John Wiley & Sons, New York.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Pincus, M. (1968). A closed form solution of certain programming problems. *Operations Research*, 16:690–694.

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley & Sons, New York.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Stephens, C. P. and Baritompa, W. (1998). Global optimization requires global information. *Journal of Optimization Theory and Applications*, 96:575–588.