Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

THEORY AND APPLICATION OF ANNEALING ALGORITHMS FOR CONTINUOUS OPTIMIZATION

Saul B. Gelfand
Peter C. Doerschuk
Mohamed Nahhas-Mohandes

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907-1285, U.S.A.

ABSTRACT

Simulated annealing algorithms for optimization over continuous spaces come in two varieties: Markov chain algorithms and modified gradient algorithms. Unfortunately, there is a gap between the theory and the application of these algorithms: the convergence conditions cannot be practically implemented. In this paper we suggest a practical methodology for implementing the modified gradient annealing algorithms based on their relationship to the Markov chain algorithms.

1 INTRODUCTION

Simulated annealing is a popular approach to global optimization of functions with multiple local minima. One type of annealing algorithm for continuous optimization involves simulating a Markov chain using a generalized Metropolis (or related) method. We refer to these algorithms as Markov chain annealing algorithms (MCAA's). There is a large amount of theoretical analysis and practical methodology developed for the MCAA's (Vanderbilt and Louie, 1984; Bohachevsky et al., 1986; Corana et al., 1987; Brooks and Verdini, 1988; Press and Teukolsky, 1991; Gelfand and Mitter, 1992). However, the feasibility of MCAA's for high-dimensional problems is questionable.

Another type of annealing algorithm for continuous optimization involves modifying gradient-type search algorithms. Let U(•) be a smooth cost function on $\mathbb{R}^D$. A standard gradient algorithm for finding a local minimum of U(•) (and hence a global minimum if U(•) is convex) is given by

$$z_{k+1} = z_k - \mu \nabla U(z_k)$$

where μ is a step-size parameter. A modified gradient algorithm for finding a global (or near global) minimum of U(•) is given by

$$X_{k+1} = X_k - \mu \nabla U(X_k) + \sqrt{2T\mu}\, W_k$$

where $\{W_k\}$ is a white Gaussian noise sequence and T is a "temperature" parameter which is slowly decreased as the algorithm proceeds. The idea behind this algorithm is that by artificially adding in the noise term (via Monte Carlo simulation) it is possible to escape from strictly local minima. We refer to this modified gradient algorithm as a gradient annealing algorithm (GAA). Now there is some theoretical analysis developed for the gradient annealing algorithm (Kushner, 1987; Gelfand and Mitter, 1991b,c), but no practical methodology that we are aware of. On the other hand, there may be some hope of using GAA for high-dimensional problems with smooth well-behaved cost functions, as it attempts to exploit the smoothness by its use of derivatives. The goal of this paper is to use some theory from Gelfand and Mitter (1991a) relating the MCAA and GAA, and some practical methodology from Johnson et al. (1989) for the MCAA, to develop a practical methodology for GAA.

2 MARKOV CHAIN ANNEALING ALGORITHMS

Most of the theory and application of MCAA deals with discrete (combinatorial) optimization. The literature on MCAA's for continuous optimization is by and large a straightforward generalization of the discrete case. It is this point of view we discuss in this section. The discussion is very brief and the reader is referred to the literature for more details.

Let U(•) be a cost function on $\mathbb{R}^D$. We wish to find an element of $\mathbb{R}^D$ which minimizes U(•). A general description of the MCAA for solving this problem is as follows (we only consider the Metropolis procedure here):

Given a current solution $x \in \mathbb{R}^D$ generate a candidate solution $y \in \mathbb{R}^D$

If $U(y) \le U(x)$ then accept y as the next solution.

If $U(y) > U(x)$ then accept y as the next solution with probability $\exp(-(U(y) - U(x))/T)$; (otherwise the next solution is the current solution x).

Here the candidate solution is usually a probabilistically generated perturbation of the current solution. Also, the "temperature" parameter T is slowly decreased as the algorithm proceeds, making transitions to higher cost states less likely. The algorithm stops subject to some termination criterion.

The MCAA can be precisely formulated as a continuous state Markov chain as follows. Let $q(x,y)$ be a transition probability density from x to y ($x,y \in \mathbb{R}^D$); $q(x,y)$ is a probability density for the candidate state y given the current state x. The continuous state annealing chain $\{Y_k\}$ (at a fixed temperature T) has 1-step transition probability density from x to y given by

$$p(T,x,y) = s(T,x,y)q(x,y) + m(T,x)\delta(y - x) \quad (2.1)$$

where

$$s(T,x,y) = \exp\left[-\frac{[U(y) - U(x)]^+}{T}\right]$$

and $m(T,x)$ is chosen to provide the correct normalization. Here $[\bullet]^+$ denotes positive part and $\delta(\bullet)$ is a Dirac-delta function. For a fixed temperature T this annealing chain $\{Y_k\}$ has a Gibbs equilibrium distribution with density function

$$\pi(T,x) = \frac{1}{Z(T)} \exp\left[-\frac{U(x)}{T}\right];$$

$$Z(T) = \int \exp\left[-\frac{U(x)}{T}\right] dx (< \infty),$$

and as the temperature T tends to zero we get $\pi(T,\bullet)$ converging to a density $\pi^*(\bullet)$, which is concentrated on the global minima of $U(\bullet)$. If the rate of temperature decrease is slow enough, then $\{Y_k\}$ remains near the equilibrium distributions and also concentrates on the global minima for k large (Gelfand and Mitter, 1992) (we note that the proof of convergence in the continuous case naturally requires many more technical assumptions and details than the discrete case).

Unfortunately, there is a large gap between the theory and application of the MCAA. The main problem is that the theoretically appropriate rates of decrease for the temperature are far too slow for practical implementation. In practice, one needs a temperature schedule, a candidate generator and a termination criterion which achieve desirable tradeoffs between complexity and performance.

A practical methodology for continuous state MCAA's can be adopted with relatively few changes from the methodology for discrete state MCAA's developed by Johnson et al. (1989) (refinements of this latter methodology form the basis for most implementations of MCAA's in both continuous and discrete state-space). A key quantity in this methodology is the acceptance probability $P_A(T)$ which is estimated by

$$\hat{P}_A(T) = \frac{N_A(T)}{N(T)} \quad (2.2)$$

where $N_A(T)$ is the number of moves accepted, and $N(T)$ is the number of moves attempted, at temperature T. Although there is some motivation for allowing $N(T)$ to increase with decreasing T, the experiments by Johnson et al. (1989) suggest that there is no real advantage to doing so, and hence $N(T)$ is fixed at some number N. The methodology proceeds by making this fixed number of iterations (attempted moves) at each of a sequence of geometrically decaying temperatures, and the initial and final temperatures are selected by requiring that the acceptance probability be specified values. For termination in the discrete case it is also required that the running cost for the best solution has not decreased over the 5 previous temperature values; in the continuous case considered here we modify this to only require that the running cost for the best solution has not decreased by more than a small threshold. A summary description of the algorithm is given below.

Markov chain annealing algorithm methodology

Input parameters: p_0 (initial acceptance probability), p_F (final acceptance probability), ρ (geometric ratio in temperature schedule), N (number of iterations at any temperature), ε (termination threshold)

1. Find initial temperature T_0 such that $\hat{P}_A(T_0) = p_0$ ($\hat{P}_A(T_0)$ given by Equation (2.2)).
2. Set $j = 0$.
3. Run the annealing chain N iterations at temperature $T_j = \rho^j T_0$
4. Let Y_j^* be the best solution found through temperature T_j
   If $\hat{P}_A(T_j) \le p_F$ and $U(Y_j^*) \ge U(Y_{j-5}^*) - \varepsilon$
   then terminate the search and output Y_j^* and $U(Y_j^*)$
   else set $j = j + 1$ and go to 3. □

Although there have been some successes reported with MCAA's of the general type described above, it has been observed that the method is very inefficient for high dimensional problems, essentially because it does not exploit the smoothness of the cost function. In the next section, we discuss the GAA which may overcome this inefficiency in some problems.

3 GRADIENT ANNEALING ALGORITHM

Let $U(\bullet)$ be a smooth cost function (at least C^2) on $\mathbb{R}^D$. We wish to find an element of $\mathbb{R}^D$ which minimizes $U(\bullet)$. Here we consider GAA as an alternative to the MCAA described in Section 2.

The GAA (with a fixed step size μ and temperature T) is given by the following stochastic recursion

$$X_{k+1} = X_k - \mu\nabla U(X_k) + \sqrt{2T\mu}\,W_k \qquad (3.1)$$

where $\{W_k\}$ is a standard D-dimensional white Gaussian noise sequence, artificially added in (via Monte Carlo simulation) to try to avoid getting trapped in local minima, and the temperature T (and possibly the step-size μ) is slowly decreased as k gets large. The asymptotic (large-time) behavior of GAA and MCAA are similar. For fixed temperature T and small step-size μ the process $\{X_k\}$ *nearly* has a Gibbs equilibrium distribution with density function $\pi(T,\bullet)$, and as the temperature T tends to zero we get $\pi(T,\bullet)$ converging to a $\pi^*(\bullet)$ which is concentrated on the global minima of $U(\bullet)$. If the rate of temperature and step-size decrease are chosen appropriately, then $\{X_k\}$ remains near the equilibrium distributions and also concentrates on the global minima for large k (Kushner, 1987; Gelfand and Mitter, 1991b,c).

GAA is plagued by the same gap between theory and application as the MCAA's. Theoretically appropriate rates of decrease for the temperature schedule are too slow for practical implementation, and no results are available concerning the important case of fixed step-size which by analogy with standard gradient algorithms is necessary for rapid convergence. In practice one needs a temperature schedule, a step-size (assumed known and fixed here) and a termination criterion which achieve desirable tradeoffs between complexity and performance. Practical implementation of GAA appears not to have received any attention in the literature.

We shall suggest a practical methodology for GAA based on the methodology for MCAA's discussed in Section 2, and the relationship between GAA and MCAA which we shall elaborate on below. We shall show that GAA and a certain class of MCAA interpolated into continuous time (with step-size/interpolation interval μ) both have a diffusion limit (as $\mu \to 0$), and these diffusion limits are linearly time-scaled versions of one another. Hence by taking into account the appropriate time-scaling, we can use the MCAA methodology as a basis for a GAA methodology. An important feature of this approach is that it allows us to implicitly associate the idea of acceptance probability with GAA - a critical quantity in developing temperature schedules and initialization and termination criterion for most practical annealing schemes.

3.1 Diffusion Limits for MCAA and GAA

We first formulate a MCAA which has the appropriate structure and scaling to admit a diffusion limit. Referring to the general version of the MCAA in Section 2 we consider here a transition density $q(\bullet,\bullet)$ which corresponds to selecting a coordinate direction at random, and then making a Gaussian perturbation along that coordinate. Let x_i denote the i-th coordinate of $x \in \mathbb{R}^D$. We choose

$$q(x,y) = \frac{1}{D}\sum_{i=1}^{D} N(x_i,\alpha\mu)(y_i)\prod_{j\neq i}\delta(y_j - x_j) \qquad (3.2)$$

where $N(m,\sigma^2)(\bullet)$ denotes a (scalar) Gaussian density with mean m and variance σ^2. Note that the variance of the Gaussian perturbation along the selected coordinate is $\alpha\mu$ where α does not depend on T (so that $q(\bullet,\bullet)$ does not depend on T) and is to be specified. Let $\{Y_k\}$ be a Markov chain with 1-step transition density $p(T,\bullet,\bullet)$ given by Equations (2.1) and (3.2). Interpolate $\{Y_k = Y_k^\mu:\ k = 0,1,...\}$ into a continuous-time process $\{Y^\mu(t):\ t \geq 0\}$ by

$$Y^\mu(t) = Y_k^\mu, \qquad t \in [k\mu,(k+1)\mu),\ k = 0,1,...$$

It can be shown (Gelfand and Mitter, 1991a) that $Y^\mu(\bullet)$ has a diffusion limit, i.e., $Y^\mu(\bullet) \to Y(\bullet)$ as $\mu \downarrow 0$ (in law), where $Y(\bullet)$ satisfies the Ito equation

$$dY(t) = -\frac{\alpha}{2TD}\nabla U(Y(t))dt + \sqrt{\frac{\alpha}{D}}\,dV(t)$$

and $V(\bullet)$ is a standard D-dimensional Wiener process.

Next, consider the GAA $\{X_k\}$ given by Equation (3.1). Interpolate $\{X_k = X_k^\mu:\ k = 0,1,...\}$ into a continuous-time process $\{X^\mu(t):\ t \geq 0\}$ by

$$X^\mu(t) = X_k^\mu, \qquad t \in [k\mu,(k+1)\mu),\ k = 0,1,...$$

It is easy to show that $X^\mu(\bullet)$ also has a diffusion limit, i.e., $X^\mu(\bullet) \to X(\bullet)$ as $\mu \downarrow 0$ (in law), where $X(\bullet)$

satisfies the Ito equation

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2T}\,dW(t)$$

and $W(\bullet)$ is a standard D-dimensional Wiener process.

Now consider the process defined by linearly scaling time by a factor $\beta > 0$ in the process $X(\bullet)$:

$$\tilde{X}(t) = X(\beta t)$$

By standard calculations $\tilde{X}(\bullet)$ satisfies the Ito equation

$$d\tilde{X}(t) = -\beta\nabla U(\tilde{X}(t))dt + \sqrt{2T\beta}\,d\tilde{W}(t)$$

where $\tilde{W}(\bullet)$ is a standard D-dimensional Wiener process. From the (assumed) uniqueness of the Ito equation solution, it is seen that if we take

$$\beta = \beta(T) = \frac{\alpha}{2TD}$$

then $Y(\bullet) = \tilde{X}(\bullet)$ (in law), i.e., $Y(\bullet)$ is a linearly time-scaled version of $X(\bullet)$, with scale-factor $\beta(T)$ depending inversely on the temperature T.

3.2 Toward a Methodology for GAA

In view of the limit diffusion behavior exhibited by both GAA and MCAA, we have that under suitable conditions, GAA is close to a linearly time-scaled version of MCAA. This suggests that we can use the MCAA methodology to guide the GAA methodology by correcting for the scaling (this is not to say that GAA and MCAA perform the same; see the discussion in Section 4).

The idea is the following. Suppose we run the MCAA $\{Y_k\}$ for N iterations at temperature T, and consider the GAA $\{X_k\}$ also at temperature T. Then since the limit diffusion $Y(\bullet)$ for $\{Y_k\}$ is a linearly time-scaled version of the limit diffusion $X(\bullet)$ for $\{X_k\}$ with scale factor $\beta(T)$, the suggestion is to run the GAA $\{X_k\}$ for $N(T) = \beta(T)N$ iterations at temperature T to compensate for the time scaling. Now we still need to choose the parameter α in the variance of the MCAA. We do this by choosing $\beta(T_0) = 1$, i.e., we choose the GAA and MCAA to run at the same time scale at the initial (high) temperature value T_0 (this choice is somewhat arbitrary but avoids introducing additional parameters). Hence $\alpha = 2T_0D$ and so $\beta(T) = T_0/T$ and thus $N(T) = (T_0/T)N$, and we run the GAA $\{X_k\}$ for

$$N_j = \rho^{-j}N_0$$

iterations at temperature $T_j = \rho^j T_0$. From a practical point of view, we may impose a ceiling on the number of iterations the GAA can make at any temperature.

The basic structure of the MCAA methodology now carries over to a GAA, except that the MCAA uses a fixed number of iterations at each of a geometrically decreasing sequence of temperatures, while the GAA uses a geometrically increasing sequence of iterations at a geometrically decreasing sequence of temperatures.

To apply the MCAA methodology to GAA it is desirable to find a good estimate of the acceptance probability, which is used to determine the initial and final temperatures. Clearly, it is not desirable to estimate the acceptance probability via a Monte Carlo simulation of a MCAA (in addition to the GAA). Now in view of the limit diffusion analysis, the appearance of the gradient term in the GAA can be viewed as a local approximation in a certain MCAA. This approximation is possible because of the (assumed) smoothness in the cost function and the smallness of the step size, and should result in significant computational and performance advantages for GAA. We shall next discuss how to make some other local approximations in the MCAA to facilitate estimation of the acceptance probability, which should make for more efficient determination of the initial and final temperatures.

The acceptance probability at temperature T is given by

$$P_A(T) = \int \pi(T,x)P_A(T\,|\,x)dx$$

where

$$P_A(T\,|\,x) = \int s(T,x,y)q(x,y)dy$$

is the conditional probability of accepting a candidate move given the current state is x. We develop an approximation to $P_A(T)$ as follows. Substituting for $q(\bullet,\bullet)$ from Equation (3.2) and setting $\alpha = 2T_0D$ we can write

$$P_A(T\,|\,x) = \frac{1}{D}\sum_{i=1}^{D} P_A(T\,|\,x,i)$$

where

$$P_A(T\,|\,x,i) = \int s(T,x,y)N(x_i,2T_0D\mu)(y_i)\prod_{j\neq i}\delta(y_j - x_j)dy$$

is the conditional probability of accepting a candidate move given the current state is x and coordinate i is selected for perturbation. Fix T, x and i for the moment and let

$$\tilde{x} = (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_D)$$

Then

$$P_A(T \mid x,i) = \int \exp\left[-\frac{[U(\tilde{x})-U(x)]^+}{T}\right] N(x_i, 2T_0 D\mu)(y_i)dy_i$$

We estimate $P_A(T \mid x,i)$ by

$$\hat{P}_A(T \mid x,i) = \int \exp\left[-\frac{[U_{x_i}(x)(y_i-x_i)+U_{x_i x_i}(x)(y_i-x_i)^2/2]^+}{T}\right]$$

$$\bullet\ N(x_i, 2T_0 D\mu)(y_i)dy_i$$

It is possible to work out expressions for $\hat{P}_A(T \mid x,i)$ in terms of the exponential and error functions. Finally we estimate $P_A(T)$ by

$$\hat{P}_A(T) = \frac{1}{N(T)} \sum_{k=1}^{N(T)} \hat{P}_A(T \mid X_k) \qquad (3.3)$$

$$\hat{P}_A(T \mid x) = \frac{1}{D} \sum_{i=1}^{D} \hat{P}_A(T \mid x,i)$$

where the average in the first equation is computed over the $N(T)$ iterations of GAA at temperature T.

A summary description of the proposed methodology for the GAA is given below.

Gradient Annealing Algorithm Methodology

Input parameters: p_0 (initial acceptance probability), p_F (final acceptance probability), ρ (geometric ratio in temperature schedule), N_0 (number of iterations at initial temperature), ε (termination threshold)

1. Find initial temperature T_0 such that $\hat{P}_A(T_0) = p_0$ ($\hat{P}_A(T_0)$ given by Equation (3.3)).
2. Set $j = 0$.
3. Run the modified gradient algorithm $N_j = \rho^{-j} N_0$ iterations at temperature $T_j = \rho^j T_0$
4. Let X_j^* be the best solution found through temperature T_j
   If $\hat{P}_A(T_j) \le p_f$ and $U(X_j^*) \ge U(X_{j-5}^*) - \varepsilon$
   then terminate the search and output X_j^* and $U(X_j^*)$
   else set $j = j+1$ and go to 3. □

4 CONCLUSIONS

In this paper we have developed a methodology for GAA based on the relationship between GAA and MCAA. The idea here is that GAA and a certain MCAA have diffusion limits which are linearly time-scaled versions of each other, which suggests that a GAA methodology can be obtained from a MCAA methodology by correcting for the time-scaling. This approach allows us to associate the idea of acceptance probability with GAA, a quantity which plays a critical role in temperature schedules for most practical annealing schemes. We also show how to make some local approximations which facilitate better estimation of the acceptance probability.

The experimental evaluation of the proposed GAA methodology is currently being undertaken. One interesting comparison would be between GAA and MCAA. Our intuition is that GAA will do a better job of finding a global minimum than MCAA for sufficiently smooth, well-behaved cost functions. For this to make sense, the implicit assumption that we are making is that GAA and MCAA are *close enough* to their diffusion limit to have a similar methodology (i.e., structure of their temperature schedule and termination criterion), but *far enough* from their diffusion limit to have distinctly different performance on certain problems.

REFERENCES

Bohachevsky, I.O., M.E. Johnson, and M.L. Stein. 1986. Generalized simulated annealing for function optimization. *Technometrics* 28:209-217.

Brooks, D.G. and W.A. Verdini. 1988. Computational experience with generalized simulated annealing over continuous variables. *American Journal Mathematics and Management Sciences* 8:425-449.

Corana, A., M. Marchesi, C. Martini, and S. Ridella. 1987. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software* 13:262-280.

Gelfand, S.B. and S.K. Mitter. 1991a. Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions. *Journal of Optimization Theory and Applications* 68:483-498.

Gelfand, S.B. and S.K. Mitter. 1991b. Simulated annealing type algorithms for multivariate optimization. *Algorithmica* 6:419-436.

Gelfand, S.B. and S.K. Mitter. 1991c. Recursive stochastic algorithms for global optimization in $\mathbf{R}^d$. *SIAM Journal on Control and Optimization* 29:999-1018.

Gelfand, S.B. and S.K. Mitter. 1992. Metropolis-type annealing algorithms for global optimization in $\mathbf{R}^d$. *SIAM Journal on Control and Optimization* (to appear).

Johnson, D.S., C.R. Aragon, L.Y. McGeoch and C. Schevon. 1989. Optimization by simulated annealing: an experimental evaluation: Part I, graph partitioning. *Operations Research* 37:865-892.

Kushner, H.J.. 1987. Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo. *SIAM Journal on Applied Mathematics* 47:169-185.

Press, W.H. and S.A. Teukolsky. 1991. Simulated annealing optimization over continuous spaces. *Computers in Physics* July/Aug:426-429.

Vanderbilt, D. and S.G. Louie. 1984. A Monte Carlo simulated annealing approach to optimization over continuous variables. *Journal of Computational Physics* 56:259-271.

AUTHOR BIOGRAPHIES

SAUL B. GELFAND received the Ph.D. degree from MIT in electrical engineering in 1987. Since 1987 he has been an Assistant Professor on the faculty of the School of Electrical Engineering at Purdue University. His research interests are in nonlinear and adaptive algorithms for communications and signal processing, optimization theory, and pattern recognition.

PETER C. DOERSCHUK received a Ph.D. degree from MIT in electrical engineering in 1985 and an M.D. degree from Harvard Medical School in 1987. Since 1990 he has been an Assistant Professor on the faculty of the School of Electrical Engineering at Purdue University. His research interests are in statistical signal processing, especially biological and medical problems (such as x-ray crystallography).

MOHAMED NAHHAS-MOHANDES is currently pursuing his Ph.D. in electrical engineering at Purdue University.