Abstract

A new optimization strategy is introduced which randomly samples a Boltzmann distribution at an array of fixed temperatures. The design and expected behavior of the resulting algorithm is analyzed theoretically. The algorithm implemented and applied to a standard test problem from the global optimization literature.

1 Introduction

Simulated annealing[6] is a global optimization algorithm based on the Metropolis sampling procedure[7]. Simulated annealing converges to the global optimum of combinatorial problems under certain conditions[4]. Simulated annealing was later generalized[1] with the Hastings sampling procedure[5] to obtain an algorithm may be faster if an appropriately biased generator can be obtained[2]. In all cases, guaranteed convergence requires a cooling schedule of the form $T_k \geq C/\ln(k + 1)$, which requires exponential time to fall to a specified temperature. Cooling must be slow enough to permit the algorithm to eventually escape a local minimum at any stage of execution. In this paper, I will present an alternative to annealing that can easily escape local minima.

*Supported in part by U. S. Army Contract DAAL03-89-D-003-0004 and by the Center for Communications and Signal Processing at North Carolina State University.
2 Biased generators

The Markov chain on a finite set $X$ resulting from the transition matrix from $x \in X$ to $y \in X$

$$P(y|x) = \begin{cases} \gamma(y|x)a(y|x) & \text{if } y = x, \\ 1 - \sum_{z \neq x} P(z|x) & \end{cases}$$

(1)

with arbitrary positive generator $\gamma$ and acceptance probability

$$a(y|x) = \min \left( 1, \frac{\gamma(x|y)\pi(y)}{\gamma(y|x)\pi(x)} \right)$$

(2)

will eventually be distributed proportional to $\pi$. Consider the dynamics of a probability distribution stochastically evolving at fixed temperature $T$ according to the transition matrix Equation 1. Because $T$ is fixed, so is the Gibbs distribution $\pi$. The generator $\gamma$ is also assumed fixed. The rate at which the probability $p(x, t)$ of a particular state $x$ changes is the difference of the rate of transitions into $x$ per unit time minus the rate out of $x$

$$\frac{dp(x,t)}{dt} = \sum_{y} P(x|y)p(y, t) - \sum_{y} P(y|x)p(x, t).$$

(3)

If the generator is identical to the target distribution $\gamma(y|x) = \pi(y), \forall x, y$, then $P(y|x) = \gamma(y|x)\min(1, 1) = \pi(y)$, so that Equation 3 reduces to the special case

$$\frac{dp(x,t)}{dt} = \sum_{y} \pi(x)p(y, t) - \sum_{y} \pi(y)p(x, t) = \pi(x) - p(x, t),$$

(4)

a differential equation with a well known exponential solution, $p(x, t) = \pi(x) + (p(x, 0) - \pi(x))e^{-t}$. In this case, $p(x, t)$ samples from an arbitrary initial distribution to the Gibbs distribution $\pi(x)$ exponentially fast with a characteristic time of unity. This motivates the following result.

**Theorem:** If the generator $\gamma$ of the finite transition matrix of Equations 1 and 2 approximates the Gibbs distribution in the sense that the deviation

$$\delta(x|y) = \frac{\gamma(x|y)}{\pi(x)} - 1$$

(5)
obeys $|\delta(x|y)| \leq \delta, \forall x, y$ for some positive constant $\delta < \frac{1}{2}$, then the Markov chain induced by Equations 1 and 2 converges from arbitrary initial $p$ to the Gibbs distribution faster than an exponential with characteristic time $\tau = 1/(1 - 2\delta)$.

**Proof:** Assuming positive distributions, substitution of Equation 3 into Equation 3 results in

$$\frac{dp(x,t)}{dt} = \sum_y (p(y,t)\pi(x) - p(x,t)\pi(y)) \min \left(\frac{\gamma(x|y)}{\pi(x)}, \frac{\gamma(y|x)}{\pi(y)}\right), \quad (6)$$

after a little algebra. If the deviation of the evolving distribution from the asymptotic Gibbs distribution is defined as

$$\epsilon(x, t) \equiv \frac{p(x,t)}{\pi(x)} - 1 \quad (7)$$

then Equation 6 can be rewritten in terms of $\delta$ and $\epsilon$ as

$$\frac{d\epsilon(x,t)}{dt} = -\epsilon(x,t) + \sum_y \pi(y)(\epsilon(y,t) - \epsilon(x,t)) \min (\delta(y|x), \delta(x|y)). \quad (8)$$

At each time define $\bar{\epsilon}(t) = \max_x (|\epsilon(x,t)|)$. If $p \neq \pi$ the largest component of the fractional deviations $\epsilon(x,t)$ is either positive or negative. If it is positive then $\exists \bar{x}$ such that $\bar{\epsilon}(t) = \epsilon(\bar{x},t)$ which is changing in time at the rate

$$\frac{d\bar{\epsilon}(t)}{dt} = -\bar{\epsilon}(t) + \sum_y (\pi(y)\epsilon(y,t) - \pi(y)\bar{\epsilon}(t)) \min (\delta(y|x), \delta(x|y)). \quad (9)$$

But $\epsilon(y,t) \leq |\epsilon(y,t)| \leq \bar{\epsilon}(t)$ and $\min (\delta(x|y), \delta(y|x)) \leq \delta$ so that

$$\frac{d\bar{\epsilon}(t)}{dt} \leq -\bar{\epsilon}(t) + 2 \sum_y \pi(y)\bar{\epsilon}(t)\delta \quad (10)$$

and $\sum_y \pi(y) = 1$ so that

$$\frac{d\bar{\epsilon}(t)}{dt} \leq -\left(1 - 2\delta\right)\bar{\epsilon}(t) = -\frac{\bar{\epsilon}(t)}{\tau}. \quad (11)$$

On the other hand, if $\bar{\epsilon}(t)$ is attained for a negative $\epsilon(x,t)$, then for some $\bar{x}$, $\bar{\epsilon}(t) = -\epsilon(\bar{x},t)$ which leads again to Equation 11. Therefore for arbitrary probability distribution $p$ and arbitrary time, the bound $\bar{\epsilon}(t)$ on the magnitude of the fractional deviation of $p$ relative to $\pi$ decreases faster than an exponential with characteristic time $\tau$ as claimed. $\square$
3 The Boltzmann distribution

The Boltzmann distribution at for a real-valued objective \( f \) and temperature \( T \) on the set \( X \) is

\[
\pi_T(x) = \frac{1}{Z_T} \exp \left( - \frac{f(x)}{T} \right) \tag{12}
\]

where

\[
Z_T = \sum_{x \in X} \exp \left( - \frac{f(x)}{T} \right) \tag{13}
\]

depends on \( T \).

Consider the Markov chain generated by the transition matrix of Equation 1 for the target distribution \( \pi_T \) of Equation 12 with an equilibrated Markov chain at a slightly higher temperature as a candidate generator, so that

\[
\gamma(x) = \pi_{T + \Delta_T}(x).
\]

Then

\[
\frac{\gamma(x)}{\pi(x)} = \frac{Z_T}{Z_{T + \Delta_T}} \exp \left( - \frac{f(x)}{T + \Delta_T} + \frac{f(x)}{T} \right).
\]

For small enough \( \Delta_T \), this is approximately

\[
\frac{\gamma(x)}{\pi(x)} = 1 - \Delta_T \frac{d}{dT} \left( \frac{f(x)}{T} + \ln Z_T \right).
\]

By Equation 13,

\[
\frac{d}{dT} \ln Z_T = \sum_{x \in X} \frac{f(x)}{T^2} \exp \left( - \frac{f(x)}{T} \right) = - \frac{\langle f \rangle}{T^2},
\]

so that Equation 5 becomes

\[
\bar{\delta} = \max_{x \in X} \left( \frac{\gamma(x)}{\pi(x)} - 1 \right) \leq \frac{\Delta_T \Delta_f}{T^2},
\]

where

\[
\Delta_f = \max_{x \in X} (f(x) - \langle f \rangle) \leq \max_{x \in X} f(x) - \min_{x \in X} f(x).
\]

This implies that the second Markov chain equilibrates about \( \tau \) iterations after the first Markov chain equilibrates.
4 An array of samplers

Instead of the two samplers of the preceding section, consider an array of $K$ samplers, each operating at its own fixed temperature $T_k$ for $k = 1, 2, \ldots, K$. We take $T_k$ to be a monotonically decreasing function of $k$. If the difference in temperatures between consecutive samplers is small enough to treat $T_k$ as a continuous function of $k$, we can solve for that difference

$$-\frac{dT}{T^2} = \frac{dk}{\Delta_f/\delta}$$

which can be integrated to obtain

$$\frac{1}{T_k} = \frac{1}{T_1} + \frac{k}{\Delta_f/\delta}$$

which determines $T_k$. The Markov chain generated by each sampler in such an array would equilibrate about $\tau$ time after its predecessor had equilibrated. We can choose the temperature of the first sampler high enough to equilibrate immediately, say $T_1 = 2\Delta_f$). In principle, we can choose the last temperature $T_K = T_{fin}$ low enough so that the corresponding Boltzmann distribution of Equation 12 is dominated by acceptably close to the global minimum. The number of samplers $K$ in the array can be obtained from Equation 4

$$K = \frac{\Delta_f}{T_{fin}\delta}.$$

The total time for the last sampler to equilibrate is of order $K\tau$. For fixed $\delta$, say $\delta = 1/4$, this means that the entire array equilibrates in a time of order $\frac{\Delta_f}{T_{fin}}$.

5 Experimental results

A simple program was written to implement this idea on a DECstation 3100. In each time step, each unit in the array was designed to update its current state first with the state of the preceding sampler, then with a uniformly generated candidate. When the current state of preceding sampler $k - 1$ is used as a candidate for sampler $k$, the acceptance probability must be
adjusted for the slight difference in temperatures according to Equation 2 as
the minimum of unity and

\[ \exp \left( -(f_{k-1} - f_k) \times \left( \frac{1}{T_k} - \frac{1}{T_{k-1}} \right) \right). \]

For the uniformly generated candidate, the acceptance probability of Equation 2 reduces to the usual Metropolis criterion of the minimum of unity and

\[ \exp \left( -\frac{f' - f_k}{T_k} \right), \]

where \( f' = f(x') \) and \( x' \) is a candidate generated uniformly in a region of specified size, stepsizes, around the current state \( x_k \).

The algorithm was tested on the Shekel function with five poles in 4 dimensions[8]. This is a standard test function for global optimization that is particularly difficult for conventional simulated annealing because it exhibits deep remote minima[3]. Several experiments were run to explore the sensitivity of the algorithm to the various parameters, as tabulated below.

<table>
<thead>
<tr>
<th>Samplers</th>
<th>Sweeps</th>
<th>Initial T</th>
<th>Step size</th>
<th>Runs</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>2.0</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.5</td>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1</td>
<td>0.5</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1</td>
<td>2.0</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>250</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>300</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>500</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>0.1</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>0.1</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.1</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>0.1</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

The implementation is not sensitive to the values of \( \Delta_f \) or \( \tilde{\delta} \) except in the resulting number of samplers in the array. The algorithm was run for a
specified number of sweeps from a specified initial temperature using differ­
et seeds for the random number generator. In all runs, the samplers were
constrained to generate candidates within a 4 dimensional hypercube of side
length 10 in all dimensions. The stepsize for the usual Metropolis step was
adjusted in two of the experiments, but was usually left at a value of 1.0
as shown in the table. The fifth column shows the number of independent
runs for each set of control parameters; the last column shows the number
of those runs that obtained the global minimum.

As can be seen from the table, the algorithm is sensitive to the control pa­
rameters. However, the last few experiments show that the global minimum
is reliably found when the total number of evaluations is $10^4$. The execution
time for these runs is a few seconds on a DECstation 3100. Straightfor­
ward simulated annealing algorithms with exponential cooling schedules do
not reliably find the global minimum for this function with $10^4$ evaluations.
Tree annealing obtains the global minimum about half the time with $10^4$
evaluations[3].

References


tion. In SPIE Conference on Stochastic Methods in Sig. Proc, Image


and the Bayesian restoration of images. IEEE Transactions on PAMI,

