Simulated Annealing with Inaccurate Cost Functions

DANIEL R. GREENING*

University of California, Los Angeles
IBM T.J. Watson Research Center, Yorktown Heights, New York

Abstract. Simulated annealing is an algorithm which generates near-optimal outcomes to combinatorial optimization problems. It is commonly thought to be slow. Cost-function approximation and parallel processing increase simulated annealing speed, but they can cause inaccuracies that degrade the outcome.

Prior theoretical work has not adequately related cost-function inaccuracy to the run-time or quality of the outcome.

We prove these results about annealing with inaccurate cost-functions: 1) Expected cost at equilibrium is exponentially affected by γ/T, where γ limits cost-function range-errors and T gives the temperature. 2) Expected cost at equilibrium is exponentially affected by $(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2$, when the errors have a Gaussian distribution. 3) Constraining γ to a constant factor of T guarantees convergence under a $1/\log t$ temperature schedule. 4) A similar constraint guarantees convergence for a fractal space with a geometric temperature schedule. 5) Inaccuracies worsen the expected outcome, but additional iterations can compensate.

Categories and Subject Descriptors: D.1.3 [**Concurrent Programming**]: *parallel programming*; F.2.2 [**Nonnumerical Algorithms and Problems**]: *routing and layout, computation on discrete structures*; G.3 [**Probability and Statistics**]: *probabilistic algorithms (including Monte Carlo)*

General Terms: Algorithms, performance.

Additional Key Words and Phrases: Parallel simulated annealing, Gaussian noise, combinatorial optimization, statistical mechanics, Boltzmann machine, conductance, mixing speed.

1. *Introduction*

Combinatorial optimization problems present this task: We are given a finite set of feasible states S, where each state $s \in S$ is represented by n state-variables, so that $s = (v_1, v_2, \ldots, v_n)$. We are given a cost-function $C: S \to \mathbb{R}$. Now, find a state with minimum cost. Many such problems are NP-complete or worse. Current algorithms to solve NP-complete problems require exponential time, based on n.

A near-optimal state is often good enough in practice. Several algorithms require only polynomial-time to produce a near-optimal state. One polynomial-time heuristic for these problems is the "greedy algorithm." Although it doesn't always produce a satisfactory outcome, it is the basis for simulated annealing.

* Currently at Novell. Internet address: `dgreen@cs.ucla.edu`

1.1. *The Greedy Algorithm*

A greedy algorithm for combinatorial optimization has a generator, which outputs a randomly-chosen state from an input state. The set of output states produced from input state s is called the neighborhood of s.

The algorithm randomly chooses its first current-state, then starts a loop. The loop first calls the generator to obtain a trial-state from the current-state. If the trial-state has lower cost, the algorithm selects the trial-state as its new current-state, otherwise, it selects the old current-state. The greedy algorithm continues its loop, generating trial-states and selecting the next state until some stopping criteria is met, commonly when no further improvement is seen for several iterations. The greedy algorithm returns the final current-state as its outcome.

The greedy algorithm has a flaw: many problems have high-cost local minima. When applied to NP-complete problems it will likely return one.

1.2. *Simulated Annealing*

Simulated annealing augments the greedy algorithm with a randomized escape from local minima. The escape is controlled through a value called "temperature." Higher temperature values make the algorithm more likely to increase cost when selecting a trial-state. In this way, simulated annealing can "climb out" of a local minimum.

```
1.    T ← T0;
2.    s ← s0;
3.    E ← C(s);
4.    while not stopping-criteria()
5.        s' ← generate(s) with probability G_ss';
6.        E' ← C(s');
7.        Δ ← E' − E;
8.        if (Δ ≤ 0) ∨ (random() < e^(−Δ/T))
9.            s ← s';
10.           E ← E';
11.       T ← reduce-temperature(T);
12.       end while;
```

FIG. 1. *Simulated Annealing*

Figure 1 shows the simulated annealing algorithm. Line 1 sets the initial temperature to T_0. Lines 2 and 3 set the current-state s and its cost E. The loop at lines 4–12 generates a trial-state s', evaluates the change in cost Δ, selects the next current-state, and reduces the temperature until the stopping criteria is met.

Line 8 shows how simulated annealing accepts a trial-state. The first term, $(\Delta \leq 0)$, expresses greed: we always accept a lower-cost trial state. The `random` function returns a uniformly-distributed random value between

0 and 1. The second term of line 6, (`random()` $< e^{-\Delta/T}$), improves the likelihood of accepting a costlier trial-state. When the stopping criteria is met, simulated annealing returns current-state s as its outcome.

At high initial temperatures, the second term of line 8 lets the algorithm explore the entire state space. As the temperature drops we explore big valleys, then smaller and smaller sub-valleys to reach the outcome.

Simulated annealing has a useful property: if you fix the temperature, it "equilibrates." That is, it approaches a stationary probability distribution, or "equilibrium." Simulated annealing's equilibrium is the "Boltzmann distribution," a probability distribution depending solely on the cost-function.

These terms—annealing, equilibrium, temperature, Boltzmann distribution, etc.—come from thermodynamics. Though we describe simulated annealing as an algorithm, it behaves like a thermodynamic system. Many publications on simulated annealing appear in physics journals.

1.3. *Inaccuracies in Circuit Placement*

Simulated annealing is frequently applied to the circuit-placement problem. Circuit-placement is NP-complete [20]. In a simplified version of circuit-placement, we are given a set of m uniform-size cells. The cells must be placed on a grid, with no more than one cell per grid position. The state variables indicate the positions of the cells, so a state is $(x_1, y_1, x_2, y_2, \ldots, x_m, y_m)$, with (x_i, y_i) indicating the position of cell i. We are also given a set of nets (wires), each of which lists two or more cells. The cost-function gives the sum of the net lengths. The goal is to place the cells on the grid to obtain the minimum total net length.

Simulated annealing takes time. It can require hours or days to place a large circuit with simulated annealing. The loop at lines 4–12 of Figure 1 dominates the execution time, so programmers try to speed up the cost computation at line 6 (along with the other calculations in the loop).

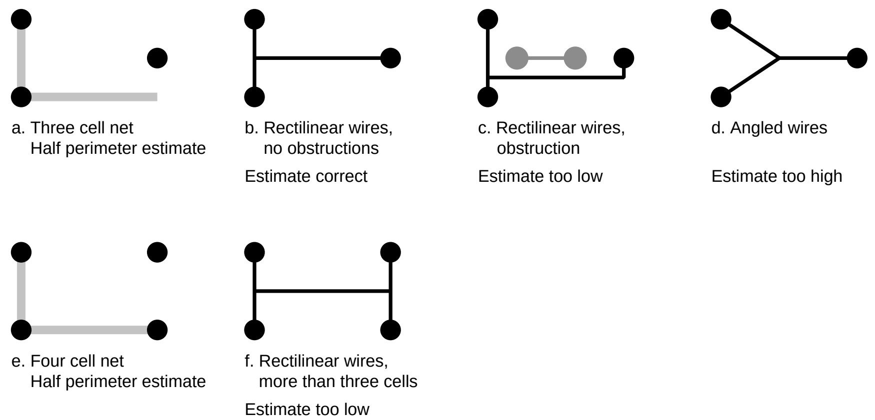


FIG. 2. *Wire Length Estimates*

Inaccuracy can speed up simulated annealing. Programmers have used faster, but less accurate cost-functions. For example, circuit placement algorithms often use the "half perimeter" method, which estimates a net's length as half the perimeter of the smallest rectangle containing all its cells. If the wires in the net parallel the x or y axis, if the net holds no more than 3 cells, and if the wires encounter no obstructions, the half perimeter method is accurate. Otherwise, it may be inaccurate.

Figure 2 shows an example. In the first row, *a* shows the half perimeter estimate for a three-cell net. *b* shows the rectilinear wiring we would use with no obstruction. *c* adds an obstruction. *d* allows the wires to be placed at angles. In the second row, *e* shows the estimate for a four-cell net. *f* shows the rectilinear wiring we would use with no obstruction. Only in case *b* is the estimate correct.

The problem statement itself is an approximation of the real problem: place transistors on silicon so that a circuit is as fast as possible. Longer wires slow down the circuit, making total net length a plausible cost-function, but long-wires are not the only contributors to circuit slowness.

1.4. *Inaccuracies in Parallel Algorithms*

Parallel processing can speed up simulated annealing. One parallel method partitions ownership of the state variables among several processors. A processor then generates trial states by manipulating only its own state variables.

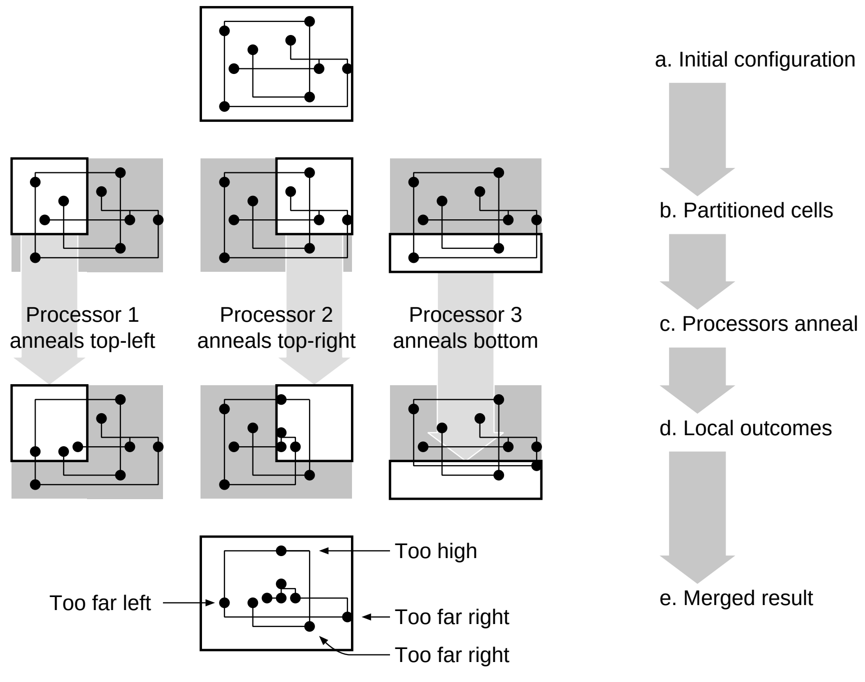


FIG. 3. *Parallel Annealing with Stale State-Variables*

Each processor maintains local copies of other processors' state variables. If the value of a local copy of a state variable differs from the value in the owning processor, we call the local data "stale." Periodically, the processors reassign cell ownership and refresh local copies.

Figure 3 shows an example. Row a shows the initial configuration. Each dot in the figure is a cell. The lines connecting the dots are nets.

Row b shows how the cells have been partitioned among the processors, in this case with some spatial locality.

In row c, each processor performs a fixed number of annealing loop iterations and then stops. A processor manipulates only state-variables for cells it owns (in the white area), but can refer to local copies of other state-variables (in the gray area). A processor may move a cell only within the white area.

Row d shows the result of each processor's annealing, with the cell locations it used to perform cost-function calculation. Notice that all local positions of unowned cells are wrong, and so many of the cost-function results were inaccurate.

Row e shows the merged outcome. If the iterations occurred at an infinitesimally low temperature (a greedy algorithm), stale state-variables caused mistakes in moving cells. The diagram points out four in the merged outcome.

The parallel algorithm allows the cost-function to use stale state-variables, and those causes the cost-function to return inaccurate results [8]. Experiments have characterized these inaccuracies [9, 12, 4].

To avoid using stale state-variables, the processors must communicate more frequently, perhaps even by locking state-variables. These operations increase execution time. Thus, as in the half perimeter wire-length estimate, we tradeoff accuracy for speed.

1.5. *Errors*

We will find it useful to measure the "error" of a cost-function computation: the difference between a state's true-cost and the value returned by the cost-function (the "observed cost"). Practitioners do not directly measure all errors in commercial applications: that would require computation of the true-cost, in addition to the inaccurate cost-function. If you can afford to compute the true-cost everytime, there is no benefit in performing the inaccurate cost computation.

Instead, we can sample the errors or execute the algorithm on a smaller representative problem to get error estimates. Quantifying these errors is valuable: the results of this paper depend on the ability to estimate errors.

This paper assumes that errors fall into two categories: range-errors and Gaussian-errors. I draw conclusions about both, but range-errors receive the most complete discussion.

We say a cost-function has "range-errors" if the difference between the true-cost $C(s)$ and the cost-function $\dot{C}(s)$ is confined to a fixed range.

We say a cost-function has "Gaussian-errors" if the difference between the true-cost $C(s)$ and the cost-function $\ddot{C}(s)$ is a Gaussian random variable with mean $C(s)$.

Throughout this text, symbols with one dot above signify range-errors, such as $\dot{C}$. Symbols with two dots above signify functions and variables with Gaussian errors. Symbols without dots refer to functions and variables without errors.

Range-errors are easier to measure and manipulate. Past experimental work has looked at range-errors. I do not know of any experimental work with Gaussian errors, however one analytic paper discusses them [6].

1.6. *The Problem*

Experiments show that larger errors increase the outcome cost, but the literature contains contradictions [19, 10] and there are several unanswered questions. We address this by exploring three general areas:

¶ *Fixed temperature behavior*: If the cost-function errors have known properties, what is the equilibrium distribution at a fixed temperature? What is the expected cost? At what rate does the system equilibrate?

¶ *Decreasing temperature schedules*: If we moderate the errors in the cost-function during execution, by changing the frequency of interprocessor communication, by altering the precision of computed values, or by introducing a cost-function with a different accuracy, can we produce as good an outcome as with the true-cost? To get the same expected outcome as with an accurate cost function, how should we alter the temperature? How will the total execution time change?

¶ *Error characteristics*: What types of inaccuracies might we expect? How do we measure them?

1.7. *Prior Work*

Appealing to physical science analogies, Grover presented an early analysis showing the effects of range-errors on the partition function [10]. Durand and White analyzed equilibrium properties for range-errors on a restricted algorithmic class [5].

Gelfand and Mitter showed that slowly decreasing state-independent Gaussian errors, under some conditions, does not affect asymptotic convergence under long schedules, for discrete [6] and continuous [7] state spaces. We do not assume state-independence; in our experience simulated annealing programs exhibit strongly state-dependent errors.

Romeo and Sangiovanni-Vincentelli give conditions on errors such that transition probabilities of an inaccurate annealing process converge to those of the accurate process as $T \to 0$ [18]. That result can be usefully applied to long schedules where regardless of the starting temperature, the system is guaranteed to approach the equilibrium distribution. Application to shorter schedules seems limited. We generalize this result slightly in §5.

Experiments have shown that when errors are proportionally constrained to temperature, outcomes improve. Invoking these observations, researchers have modified their algorithms to obtain better outcomes [13, 1, 3]. The analytic results we obtain confirm a significant temperature-dependence.

1.8. *Contents*

§2 introduces simulated annealing as a Markov process, and presents its fundamental properties.

§3 derives equilibrium properties for annealing. §3.1 shows the effects of range-errors on expected cost, and §3.2 shows the effects of Gaussian errors on expected cost.

§4 shows how inaccuracies affect conductance. Conductance is a value that approximates the rate a system approaches equilibrium (the mixing rate). §4.1 establishes the conductance for annealing with range-errors. §4.2 establishes the conductance for annealing under Gaussian noise.

§5 introduces a framework for simulated annealing [24]. Using this framework, we prove that limiting cost errors proportionally to the temperature guarantees the minimum cost outcome, using a $T(t) = d/\log t$ temperature schedule. The result is only mildly useful, since the execution time is exponential.

§6 describes fractal state spaces. Fractal spaces and practical problems are statistically similar [22, 16]. Annealing on fractal spaces have a convenient property: use of the geometric $T(t) = c^t T_0$ schedule, used in many commercial annealing programs, results in convergence to the global minimum. We prove that with a $T(t) = c^t T_0$ temperature schedule on a fractal state space, if errors are strictly limited to values proportional to the temperature, the system will converge. §6.1 analyzes "confined annealing," and §6.2 extends the result to "unconfined annealing." In §6.3, we show that inaccuracies require us to increase the annealing steps to achieve the same quality.

§7 discusses the practical issues involved in measuring errors. In this section, we also show either of two types of observed range-errors, "instantaneous" or "accumulated" errors can be used in our results.

§8 summarizes the results and offers practical insights for implementing inaccurate simulated annealing algorithms.

2. *Properties of Simulated Annealing*

First, we formalize simulated annealing. Please refer to Figure 1.

Let $N \in \mathbb{Z}^+$ be the number of states, and let $S = \{1, \ldots, N\}$ label all states. Define the *generation probability matrix* G so that $G_{ss'}$ is the probability that `generate` will return state s' when passed argument s.

Let T_t be the temperature after t iterations. Define the *acceptance probability matrix at time* t, $A^{(t)}$, so that

$$A_{ss'}^{(t)} = \begin{cases} 1 & \text{if } C(s') \leq C(s) \\ e^{(C(s)-C(s'))/T_t} & \text{otherwise.} \end{cases} \tag{1}$$

We can define an inhomogeneous Markov chain P conforming to the annealing loop of steps 3–9 in Figure 1 as follows.

$$P_{ss'}^{(t)} = \begin{cases} G_{ss'} A_{ss'}^{(t)} & \text{if } s \neq s' \\ 1 - \sum_{x \neq s'} P_{sx}^{(t)} & \text{otherwise.} \end{cases} \tag{2}$$

Let s_0 be the starting state. $P^{(t)} P^{(t-1)} \cdots P^{(0)} s_0$ gives the state probability vector after t iterations.

If we fix the temperature, then $A^{(t)} = A^{(t+1)} = A$ and $P^{(t)} = P^{(t+1)} = P$, and the annealing algorithm simulates a homogeneous Markov chain.

Even analysis of realistic accurate annealing programs is difficult: a time-dependent acceptance matrix introduces problems, programs often use time-dependent `generate` functions [25, 15], and the temperature schedule may not even be monotonic [23].

Fortunately, annealing programs attempt to bring state probabilities near equilibrium at each temperature [14]. Thus, we can gain information about annealing behavior by looking at the homogeneous chain.

The homogeneous annealing chain should be ergodic, otherwise the minimum cost state could be unreachable or periodic. With this property, each state s has an *equilibrium probability* $\boldsymbol{\pi}_s$ independent of the initial state. We can model many annealing applications as ergodic processes. We can guarantee ergodicity if the following properties hold.

$$
\begin{aligned}
&(3) \qquad \text{probability} \qquad \forall s \in S, \sum_{s' \in S} G_{ss'} = 1 \\
&(4) \qquad \text{coverage} \qquad \forall s, s' \in S, \exists k \geq 1, [(G^k)_{ss'} \neq 0] \\
&(5) \qquad \text{aperiodicity} \qquad \exists s \in S, [P_{ss} \neq 0] \\
&(6) \qquad \text{finiteness} \qquad |S| \in \mathbb{Z}^+
\end{aligned}
$$

(3) simply states that G_s is a probability vector. (4) guarantees that the annealing chain is irreducible. (5) and (6) ensure that the chain is aperiodic and finite. Irreducible, aperiodic, finite Markov chains are ergodic.

We add one more, the symmetry property.

$$(7) \qquad \text{symmetry} \qquad \forall \langle s, s' \rangle \in S \times S, [G_{ss'} = G_{s's}]$$

Including (7) produces three useful results [17]: first, the equilibrium probability is otherwise independent of G; second, the annealing chain is reversible, namely,

$$(8) \qquad \forall i, j \in S, \ \boldsymbol{\pi}_i P_{ij} = \boldsymbol{\pi}_j P_{ji}$$

third, the equilibrium probabilities are given by the Boltzmann distribution, namely,

$$(9) \qquad \boldsymbol{\pi}_s(T) = \frac{e^{-C(s)/T}}{\sum_{s' \in S} e^{-C(s')/T}}.$$

Properties (3)-(7) are based on those of Otten and van Ginneken [17]: however, they ensure aperiodicity by $\forall s \in S, G_{ss} \neq 0$, a requirement not typically satisfied by annealing programs. In contrast, this aperiodicity property, (5), is trivially satisfied if $\exists s, s' \in S, C(s) \neq C(s')$.

3. *Equilibrium Properties*

In this section, we compare the expected cost at equilibrium of an inaccurate algorithm and that of an accurate algorithm, based on both range-errors and Gaussian-errors.

When errors appear in the cost-function, we refer to the true-cost, $C(s)$, of state s, and the observed-cost $\dot{C}(s) = C(s) + \epsilon(s)$, of state s. ϵ is a random variable often dependent on s, thus $\dot{C}(s)$ is a random function.

3.1. *Range-Errors*

If the true-cost of s is $C(s)$, a cost-function with range-errors, $\dot{C}(s)$, satisfies (10).

$$\forall s \in S, C(s) + \underline{\epsilon} \leq \dot{C} \leq C(s) + \overline{\epsilon} \tag{10}$$

It turns out that the results I will show depend on the value

$$\gamma = \overline{\epsilon} - \underline{\epsilon} \tag{11}$$

Let $C\colon S \to \mathbb{R}$ be the accurate cost-function for S, and let $\dot{C}\colon S \to \mathbb{R}$ be a range-error cost-function. We can use $\underline{\epsilon}$ and $\overline{\epsilon}$ from (10) to compare the equilibrium distributions of annealing with and without range-errors. The calculations depend only on $\overline{\epsilon} - \underline{\epsilon}$, so let $\gamma = \overline{\epsilon} - \underline{\epsilon}$.

THEOREM 3.1. *Let $\dot{\boldsymbol{\pi}}_s(T)$ and $\boldsymbol{\pi}_s(T)$ be the equilibrium probabilities, with range-errors and without errors respectively, of state s at temperature T. Then,*

$$e^{-\gamma/T}\boldsymbol{\pi}_s(T) \leq \dot{\boldsymbol{\pi}}_s(T) \leq e^{\gamma/T}\boldsymbol{\pi}_s(T) \tag{12}$$

Proof. Pick any state $s \in S$. By (9), if we minimize the cost of s, and maximize the cost of all other states, we will maximize the equilibrium probability of state s. Thus, we let $\dot{C}(s) = C(s) + \underline{\epsilon}$, and $\forall s' \in (S \setminus \{s\}), [\dot{C}(s') = C(s') + \overline{\epsilon}]$.

In fixing these worst-case costs, the inaccurate system satisfies the Boltzmann distribution,

$$\dot{\boldsymbol{\pi}}_s(T) \leq \frac{e^{-(C(s)+\underline{\epsilon})/T}}{e^{-(C(s)+\underline{\epsilon})/T} + \sum_{s' \neq s} e^{-(C(s)+\overline{\epsilon})/T}} \tag{13}$$

$$= \frac{e^{-\underline{\epsilon}/T} e^{-C(s)/T}}{e^{-\underline{\epsilon}/T} e^{-C(s)/T} + e^{-\overline{\epsilon}/T} \sum_{s' \neq s} e^{-C(s)/T}} \tag{14}$$

Since $e^{-\underline{\epsilon}/T} \geq e^{-\overline{\epsilon}/T}$, we have

$$\dot{\boldsymbol{\pi}}_s(T) \leq \frac{e^{-\underline{\epsilon}/T} e^{-C(s)/T}}{e^{-\overline{\epsilon}/T} \sum_{s \in S} e^{-C(s)/T}} \tag{15}$$

Finally, substituting the accurate equilibrium probability $\boldsymbol{\pi}_s(T)$ for its equivalent, we have

$$\dot{\boldsymbol{\pi}}_s(T) \leq e^{(\overline{\epsilon}-\underline{\epsilon})/T}\boldsymbol{\pi}_s(T) \tag{16}$$

The converse argument supplies the lower bounds. □

THEOREM 3.2. *Let $\bar{C}[\boldsymbol{\pi}(T)]$ be the expected true-cost C, with equilibrium distribution $\boldsymbol{\pi}$, at temperature T. So,*

$$\bar{C}[\boldsymbol{\pi}(T)] = \sum_{s\in S} C(s)\boldsymbol{\pi}_s(T) \tag{17}$$

Then,

$$e^{-\gamma/T}\bar{C}[\boldsymbol{\pi}(T)] \leq \bar{C}[\dot{\boldsymbol{\pi}}(T)] \leq e^{\gamma/T}\bar{C}[\boldsymbol{\pi}(T)] \tag{18}$$

Proof. By Theorem 3.1,

$$e^{-\gamma/T}\sum_{s\in S} C(s)\boldsymbol{\pi}_s(T) \leq \sum_{s\in S} C(s)\dot{\boldsymbol{\pi}}_s(T) \leq e^{\gamma/T}\sum_{s\in S} C(s)\boldsymbol{\pi}_s(T) \tag{19}$$

By our definition, we have the result. □

3.2. *Gaussian Errors*

Simulated annealing typically operates on structures with discrete cost-functions: errors usually appear as discrete values. However, as we add state variables and as the maximum number of uncorrected independent moves increases (through increased parallelism or longer stream lengths, for example), the errors can approach a Gaussian distribution.

In many instances, particularly in parallel applications, the probability distribution of the inaccurate cost-function $\dot{C}(s)$ is reflected about the true-cost $C(s)$ [5].

Thus, it is reasonable to investigate the effect of Gaussian-error cost-functions, where each value is a Gaussian random variable with mean $C(s)$. We will show that when the variances of the state costs do not differ greatly, annealing converges to a good solution.

LEMMA 3.3. *Let the cost of each state s be $\ddot{C}(s) = C(s) + X_s$, where X_s is an independent random variable. Execute the simulated annealing algorithm in Figure 1, with lines 3 and 6 sampling random variables $\ddot{C}(s)$ and $\ddot{C}(s')$ respectively, and with T fixed. If (3)–(7) are satisfied and $T > 0$, the resulting homogeneous Markov chain $\ddot{P}$ is ergodic, and the equilibrium probabilities are given by*

$$\ddot{\boldsymbol{\pi}}_i = \frac{e^{-C(i)/T}\mathrm{E}[e^{-X_i/T}]}{\sum_{j\in S} e^{-C(j)/T}\mathrm{E}[e^{-X_j/T}]}. \tag{20}$$

Proof. For any two states $s, s' \in S$, the acceptance probability $\ddot{A}_{s,s'}$ under random cost-function $\ddot{C}$ satisfies $0 < \ddot{A}_{s,s'} < 1$.

$$(P^k)_{ss'} = \sum_{s_1} \cdots \sum_{s_{k-1} \in S} G_{ss_1} \ddot{A}_{s,s_1} \cdots G_{s_{k-1}s'} \ddot{A}_{s_{k-1},s'}. \tag{21}$$

Since (4) is satisfied and all terms of A are non-zero, there is some k for which $(\ddot{P}^k)_{ss'}$ is non-zero. Thus, $\ddot{P}$ is irreducible.

(3)–(7) ensure that $\ddot{P}$ is irreducible, aperiodic, and finite, so $\ddot{P}$ is ergodic. Since $\ddot{P}$ is ergodic, it satisfies the *detailed balance condition* (8). Thus, $\ddot{\pi}_i \ddot{P}_{ij} = \ddot{\pi}_j \ddot{P}_{ji}$. If $\ddot{P}_{ij}$ is non-zero, we have $\ddot{\pi}_i = \ddot{\pi}_j \ddot{P}_{ji}/\ddot{P}_{ij} = \ddot{\pi}_j G_{ji}\ddot{A}_{ji}/(G_{ij}\ddot{A}_{ij})$. By (7), $\ddot{\pi}_i = \ddot{\pi}_j \ddot{A}_{ji}/\ddot{A}_{ij}$.

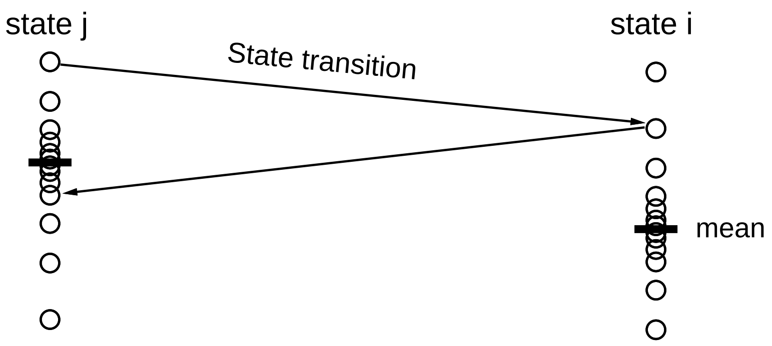


FIG. 4. *State i Sampled and Fixed.*

Upon entering state i, we sample and fix its cost until we leave, as in Figure 4. This is equivalent to saying that the error appears in the *cost difference* between the two states.

Let $a(C_i, C_j)$ be the probability of accepting a move from a state with cost C_i to a state with cost C_j. By (1) this is

$$a(C_i, C_j) = \begin{cases} 1 & \text{if } C_j \le C_i \\ e^{(C_i - C_j)/T} & \text{otherwise.} \end{cases} \tag{22}$$

Let ϕ_i be the probability density function for random variable $\ddot{C}(i)$. By (22) we have

$$\begin{aligned}
\frac{\ddot{A}_{ji}}{\ddot{A}_{ij}} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_i(x_i)\phi_j(x_j) \frac{a(x_j, x_i)}{a(x_i, x_j)} dx_i dx_j && (23) \\
&= \int_{-\infty}^{\infty} \phi_i(x_i) \left[\int_{x_j > x_i} \phi_j(x_j) \frac{1}{e^{(C(i)+x_i-C(j)-x_j)/T}} dx_j \right. && (24) \\
&\qquad \left. + \int_{x_j \le x_i} \phi_j(x_j) \frac{e^{(C(j)+x_j-C(i)-x_i)/T}}{1} dx_j \right] dx_i \\
&= \int_{-\infty}^{\infty} \phi_i(x_i) e^{(-C(i)-x_i)/T} dx_i \int_{-\infty}^{\infty} \phi_j(x_j) e^{(C(j)+x_j)/T} dx_j && (25)
\end{aligned}$$

$$(26) \qquad = \frac{e^{C(j)/T}\mathrm{E}[e^{X_j/T}]}{e^{C(i)/T}E[e^{X_i/T}]}$$

For any $G_{ij} \neq 0$, the detailed balance equation (8) is now

$$(27) \qquad \ddot{\boldsymbol{\pi}}_i e^{C(i)/T}\mathrm{E}[e^{X_i/T}] = \ddot{\boldsymbol{\pi}}_j e^{C(j)/T}\mathrm{E}[e^{X_j/T}].$$

By (20), $\ddot{\boldsymbol{\pi}}$ is a probability distribution ($\sum_{s\in S}\ddot{\boldsymbol{\pi}}_s = 1$). (20) satisfies (27) for all state pairs. Since $\ddot{P}$ is ergodic, (20) must be the unique probability distribution. □

COROLLARY 3.4. *If all X_s are identically distributed, and the resulting Markov chain is ergodic, then $\ddot{\boldsymbol{\pi}} = \boldsymbol{\pi}$.*

Proof. This follows directly from the proof of Lemma 3.3. □

LEMMA 3.5. *If X is a Gaussian random variable with mean μ and standard deviation σ, then*

$$(28) \qquad \mathrm{E}[e^{-X/T}] = e^{\sigma^2/2T^2-\mu/T}$$

Proof. By definition,

$$\mathrm{E}[e^{-X/T}] = \int_{-\infty}^{\infty} e^{-x/T}\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}dx$$

$$(29) \qquad = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-x/T-x^2/2\sigma^2+2x\mu/2\sigma^2-\mu^2/2\sigma^2}dx$$

$$(30) \qquad = \frac{e^{-\mu^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-[x^2-2x\mu+2\sigma^2x/T]/2\sigma^2}dx$$

$$(31) \qquad = \frac{e^{-\mu^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-[x^2+2x(\sigma^2/T-\mu)+(\sigma^2/T-\mu)^2-(\sigma^2/T-\mu)^2]/2\sigma^2}dx$$

$$(32) \qquad = \frac{e^{-\mu^2/2\sigma^2+(\sigma^2/T-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(x+\sigma^2/T-\mu)^2/2\sigma^2}dx$$

Let $u = (x+\sigma^2/T-\mu)/\sqrt{2}\sigma$ and $du = dx/\sqrt{2}\sigma$. We then have

$$(33) \qquad \mathrm{E}[e^{-X/T}] = \frac{e^{\sigma^2/2T^2-\mu/T}}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-u^2}du.$$

By a standard mathematical trick, we convert Cartesian to polar coordinates and obtain

$$(34) \qquad \mathrm{E}[e^{-X/T}] = e^{\sigma^2/2T^2-\mu/T}.$$

□

THEOREM 3.6. *Let $C\colon S \to \mathbb{R}$ be a cost-function with equilibrium distribution $\boldsymbol{\pi}$. Consider a random cost-function $\ddot{C}\colon S \to \mathbb{R}$, where each random variable $\ddot{C}(s)$ is an independent Gaussian distribution with mean $C(s)$ and variance $\sigma^2(s)$ and equilibrium distribution $\ddot{\boldsymbol{\pi}}$. Then $\ddot{\boldsymbol{\pi}}_s$ is bounded by*

$$(35) \qquad e^{(\underline{\sigma}^2-\overline{\sigma}^2)/2T^2}\boldsymbol{\pi}_s \le \ddot{\boldsymbol{\pi}}_s \le e^{(\overline{\sigma}^2-\underline{\sigma}^2)/2T^2}\boldsymbol{\pi}_s$$

Proof. By Lemmas 3.3 and 3.5, we have

$$\ddot{\boldsymbol{\pi}}_s = \frac{e^{\sigma^2(s)/2T^2 - C(s)/T}}{\sum_{s' \in S} e^{\sigma^2(s')/2T^2 - C(s')/T}} \tag{36}$$

Fix s. To maximize $\ddot{\boldsymbol{\pi}}_s$, we maximize the numerator at $e^{\overline{\sigma}^2/2T^2 - C(s)/T}$ and minimize the denominator at $Z = e^{\overline{\sigma}^2/2T^2} e^{-C(s)/T} + \sum_{s' \neq s} e^{\underline{\sigma}^2/2T^2} e^{C(s')/T}$. The inequality

$$Z \geq \sum_{s' \in S} e^{\underline{\sigma}^2/2T^2} e^{C(s')/T} \tag{37}$$

allows us to state that

$$\ddot{\boldsymbol{\pi}}_s \leq \frac{e^{\overline{\sigma}^2/2T^2} e^{C(s)/T}}{\sum_{s' \in S} e^{\underline{\sigma}^2/2T^2} e^{C(s)/T}} = e^{(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2} \boldsymbol{\pi}_s \tag{38}$$

A similar argument supplies the lower bound. □

THEOREM 3.7. *Let $\bar{C}[\boldsymbol{\pi}(T)]$ be the expected true-cost C, with equilibrium distribution $\boldsymbol{\pi}$, at temperature T. Then,*

$$e^{-(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2} \bar{C}[\boldsymbol{\pi}(T)] \leq \bar{C}[\ddot{\boldsymbol{\pi}}(T)] \leq e^{(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2} \bar{C}[\boldsymbol{\pi}(T)] \tag{39}$$

Proof. By Theorem 3.6,

$$e^{-(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2} \sum_{s \in S} C(s) \boldsymbol{\pi}_s(T) \leq \sum_{s \in S} C(s) \ddot{\boldsymbol{\pi}}_s(T) \leq e^{(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2} \sum_{s \in S} C(s) \boldsymbol{\pi}_s(T) \tag{40}$$

Definition (17) gives us the result. □

4. *Mixing Speed*

Mixing speed is the rate at which we approach equilibrium in an ergodic process. A higher mixing speed indicates that we can more rapidly reduce the temperature while maintaining near-equilibrium. Errors affect mixing speed, not necessarily for the worse.

To quantify mixing speed, we must first have a measure for the distance from equilibrium. Let P be the transition matrix for the annealing chain, let $\boldsymbol{\pi}$ be the stationary probability vector over state space S. Let t denote time, and $\mathbf{x}(t)$ denote state probabilities at time t, so $\mathbf{x}(t) = P^t \mathbf{x}(0)$. Define the discrepancy vector $\mathbf{d}(t) = \mathbf{x}(t) - \boldsymbol{\pi}$. The distance from equilibrium can be defined

$$\|\mathbf{d}(t)\| = \sum_{s \in S} \mathbf{d}_s^2(t). \tag{41}$$

If P is a Markov chain and $\boldsymbol{\pi}_i$ is the stationary probability of state i, define the *conductance of a subset*, Φ_V, $V \in S$ as

$$\Phi_V = \frac{\sum_{i \in V} \sum_{j \notin V} \boldsymbol{\pi}_i P_{ij}}{\sum_{i \in V} \boldsymbol{\pi}_i} \tag{42}$$

In words, the conductance of a state subset V is the conditional probability that some transition will leave the subset, given that we start there.

Let $\boldsymbol{\pi}_V = \sum_{v \in V} \boldsymbol{\pi}_v$, and let $\mathbf{S}_{1/2} = \{V \subset S | \boldsymbol{\pi}_V \leq 1/2\}$. Define the *global conductance*, Φ, as the minimum conductance over all subsets with stationary probability below 1/2, thus,

$$\Phi = \min_{V \in \mathbf{S}_{1/2}} \Phi_V \tag{43}$$

Sinclair and Jerrum showed that any initial distribution of a strongly aperiodic Markov chain, such as defined in (3)–(7), with discrepency vector $\mathbf{d}(0)$, satisfies (44) [21].

$$\|\mathbf{d}(t)\| \leq (1 - \Phi^2)^t \cdot \|\mathbf{d}(0)\| \tag{44}$$

Thus, global conductance provides a good measure for mixing speed.

Another commonly-used mixing speed measure is the "dominant eigenvalue." The convergence rate of a simulated annealing algorithm is inversely related to the eigenvalues of its underlying Markov chain [2, 17]. The eigenvalues measure the "ruggedness" of the cost landscape. The second largest eigenvalue (or "dominant eigenvalue"), in particular, provides a first-approximation to the slowness of reaching equilibrium from an arbitrary starting state.

Calculating the eigenvalues for annealing spaces is typically intractable, however a useful bound relates the more easily computed global conductance to the dominant eigenvalue [21]. Sinclair and Jerrum showed that for an ergodic reversible Markov chain, the dominant eigenvalue $\boldsymbol{\lambda}_2$ of the transition matrix satisfies (45).

$$\boldsymbol{\lambda}_2 \leq 1 - \frac{\Phi^2}{2}. \tag{45}$$

Thus, by raising the global conductance, we reduce the dominant eigenvalue and increase the convergence rate.

Others have examined the effect of adjusting the move spaces to obtain better mixing speed [22, 17]. Here, we show how the errors modify mixing speed, using global conductance as a measure.

4.1. *Mixing Speed Under Range-Errors*

Consider two functions: the *true-cost* given by C, and the *observed-cost* given by random function $\dot{C}$. Suppose $\dot{C}$ is bounded relative to C: $\forall s \in S, C(s) + \underline{\epsilon} \leq \dot{C}(s) \leq C(s) + \overline{\epsilon}$. As before, let $\gamma = \overline{\epsilon} - \underline{\epsilon}$. What is the effect of these errors on the global conductance?

LEMMA 4.1. *Consider two annealing chains P and $\dot{P}$, with G and S defined identically. They have cost-functions C and $\dot{C}$, respectively. The two cost-functions are related by $C(s) + \underline{\epsilon} \leq \dot{C} \leq C(s) + \overline{\epsilon}$. Let Φ_V and $\dot{\Phi}_V$ be the corresponding conductances for subset $V \subseteq S$. Then,*

$$e^{-3\gamma/T} \dot{\Phi}_V \leq \Phi_V \leq e^{3\gamma} \dot{\Phi}_V \tag{46}$$

Proof. The conductance of $V \subset S$, under cost-function $\dot{C}$ is

$$\dot{\Phi}_V = \frac{\sum_{i\in V}\sum_{j\notin V} \dot{\boldsymbol{\pi}}_i G_{ij}\dot{A}_{ij}}{\sum_{i\in V}\dot{\boldsymbol{\pi}}_i} \tag{47}$$

Let us construct an upper bound on Φ_V. Let $V' = S \setminus V$. For any vertex $i \in V$, we can lower its transition probability to a $j \in V'$ by a factor of at most $e^{-\gamma}$. This follows directly from the definition of A and $\dot{C}$. We minimize the acceptance term thus: $\dot{A}_{ij} = A_{ij}e^{-\gamma/T}$. We assume that the generation probability matrix G does not change. Thus, we have

$$\dot{\Phi}_V \geq e^{-\gamma/T}\frac{\sum_{i\in V}\sum_{j\in V'} \dot{\boldsymbol{\pi}}_i G_{ij}A_{ij}}{\sum_{i\in V}\dot{\boldsymbol{\pi}}_i} \tag{48}$$

Using Theorem 3.1,

$$\dot{\Phi}_V \geq e^{-\gamma/T}\frac{\sum_{i\in V}\sum_{j\in V'} e^{-\gamma/T}\boldsymbol{\pi}_i G_{ij}A_{ij}}{\sum_{i\in V}e^{+\gamma}\boldsymbol{\pi}_i}, \tag{49}$$

which implies

$$\Phi_V \leq e^{+3\gamma/T}\dot{\Phi}_V. \tag{50}$$

A similar construction supplies the lower bound. □

THEOREM 4.2. *Let Φ and $\dot{\Phi}$ be the global conductances of annealing chains P and $\dot{P}$. Then,*

$$e^{-3\gamma/T}\Phi \leq \dot{\Phi} \leq e^{+3\gamma/T}\Phi \tag{51}$$

Proof. Let $\dot{\mathbf{S}}_{1/2} = \{V \subset S | \dot{\boldsymbol{\pi}}_V \leq 1/2\}$. Suppose

$$\min_{V\subset\dot{\mathbf{S}}_{1/2}} \dot{\Phi}_V = \dot{\Phi}_A \tag{52}$$

for some set $W \in \dot{\mathbf{S}}_{1/2}$. There are two cases.

Case 1, $W \in \mathbf{S}_{1/2}$.

By Lemma 4.1, Theorem 4.2 holds.

Case 2, $W \notin \mathbf{S}_{1/2}$.

By definition,

$$\dot{\Phi}_W = \frac{\sum_{i\in W}\sum_{j\notin W}\dot{\boldsymbol{\pi}}_i\dot{P}_{ij}}{\sum_{i\in W}\dot{\boldsymbol{\pi}}_i}. \tag{53}$$

Let $W' = S \setminus W$. $\dot{P}$ satisfies the detailed balance conditions (8), and thus

$$\sum_{i\in W}\sum_{j\notin W}\dot{\boldsymbol{\pi}}_i\dot{P}_{ij} = \sum_{i\in W'}\sum_{j\notin W'}\dot{\boldsymbol{\pi}}_j\dot{P}_{ji}, \tag{54}$$

so,

$$\dot{\Phi}_W = \frac{\sum_{i\in W'}\sum_{j\notin W'} \dot{\boldsymbol{\pi}}_i \dot{P}_{ij}}{\sum_{i\in W} \dot{\boldsymbol{\pi}}_i}. \tag{55}$$

$W \in \dot{\mathbf{S}}_{1/2}$ implies that $\Rightarrow \dot{\boldsymbol{\pi}}_W \leq 1/2$. Therefore, $\dot{\boldsymbol{\pi}}_{W'} \geq 1/2$. Thus, $\dot{\boldsymbol{\pi}}_{W'} \geq \dot{\boldsymbol{\pi}}_W$ and

$$\dot{\Phi}_W \geq \frac{\sum_{i\in W'}\sum_{j\notin W'} \dot{\boldsymbol{\pi}}_i \dot{P}_{ij}}{\sum_{i\in W'} \dot{\boldsymbol{\pi}}_i} = \dot{\Phi}_{W'}. \tag{56}$$

By Lemma 4.1, we have

$$\dot{\Phi}_W \geq \dot{\Phi}_{W'} \geq e^{-3\gamma/T}\Phi_{W'}. \tag{57}$$

Note that $W' \in \mathbf{S}_{1/2}$. By (43), we have $\dot{\Phi}_W \geq e^{-3\gamma/T}\Phi$, and $\dot{\Phi} \geq e^{-3\gamma/T}\Phi$. Thus

$$e^{-3\gamma/T}\Phi \leq \dot{\Phi}. \tag{58}$$

A similar proof, starting with $\mathbf{S}_{1/2}$ instead of $\dot{\mathbf{S}}_{1/2}$, supplies the upper bound in (51). □

This result predicts the rate that a inaccurate system equilibrates, if we know the maximum error magnitude and the rate that the accurate system equilibrates.

4.2. *Mixing Speed Under Gaussian Errors*

Let us now consider the mixing speed of Gaussian cost-functions. Substitute $\ddot{C}$ for $\dot{C}$ in (47). The conductance of a subset $V \subset S$ is defined by

$$\ddot{\Phi}_V = \frac{\sum_{i\in V}\sum_{j\notin V} \ddot{\boldsymbol{\pi}}_i G_{ij} \ddot{A}_{ij}}{\sum_{i\in V} \ddot{\boldsymbol{\pi}}_i}. \tag{59}$$

LEMMA 4.3 (Conductance Reversibility). *Consider state space S, and subsets $V \subset S$ and $V' = S \setminus V$. Let Φ_V be the conductance of V, and $\boldsymbol{\pi}_V$ be the equilibrium probability of set V. If the underlying Markov space has the reversibility property, then*

$$\Phi_V \boldsymbol{\pi}_{V'} = \Phi_{V'} \boldsymbol{\pi}_V \tag{60}$$

Proof. The conductance of V is

$$\Phi_V = \frac{\sum_{i\in V}\sum_{j\in V'} \boldsymbol{\pi}_i P_{ij}}{\sum_{i\in V} \boldsymbol{\pi}_i}. \tag{61}$$

The conductance of V' is

$$\Phi_V = \frac{\sum_{i\in V}\sum_{j\in V'} \boldsymbol{\pi}_j P_{ji}}{\sum_{i\in V'} \boldsymbol{\pi}_i}. \tag{62}$$

By the reversibility property, the numerators of (61) and (62) are equal. Algebra produces (60). □

LEMMA 4.4. *Apply Gaussian-error cost-function to states $i, j \in S$, as described previously, then the acceptance probability $\ddot{A}$ behaves according to the inequality*

$$\ddot{A}_{ij} \leq e^{\overline{\sigma}^2/T^2} A_{ij}. \tag{63}$$

Proof. Let random variables X and Y be defined so $X = \ddot{C}(i)$ and $Y = \ddot{C}(j)$. By the definition of A we have

$$\ddot{A}_{ij} = \int_{-\infty}^{\infty} \left[\int_{x}^{\infty} e^{(x-y)/T} \phi_j(y)\, dy \;+\; \int_{-\infty}^{x} \phi_j(y)\, dy \right] \phi_i(x) dx \tag{64}$$

Since $y \geq x \Rightarrow e^{(x-y)/T} \leq 1$,

$$\ddot{A}_{ij} \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_j(y) \phi_i(x)\, dy\, dx. \tag{65}$$

Separating independent integrations,

$$\ddot{A}_{ij} \leq \int_{-\infty}^{\infty} \phi_i(x)\, dx \int_{-\infty}^{\infty} \phi_j(y)\, dy. \tag{66}$$

By Lemma 3.5,

$$\ddot{A}_{ij} \leq e^{\sigma_i^2/2T^2 + C(i)/T} e^{\sigma_j^2/2T^2 - C(j)/T}, \tag{67}$$

and therefore

$$\ddot{A}_{ij} \leq e^{\sigma_i^2/2T^2 + \sigma_j^2/2T^2} e^{(C(i)-C(j))/T}. \tag{68}$$

Suppose $C(i) \leq C(j)$, then $e^{(C(i)-C(j))/T} = A_{ij}$ and Equation 63 holds. Suppose $C(i) \geq C(j)$ then $A_{ij} = 1$. Since $e^{2\overline{\sigma}^2/T^2} \geq 1$ and since $\ddot{A}_{ij} \leq 1$, Equation 63 holds here also. □

LEMMA 4.5. *Let $\ddot{C}$ be a random function where each random variable $\ddot{C}(s)$ is a Gaussian. $\ddot{C}(s)$ has mean value $C(s)$. Let A_{ij} and $\ddot{A}_{ij}$ be the acceptance probabilities for an annealing transition from state i to state j under cost-functions C and $\ddot{C}$ respectively. Then A and $\ddot{A}$ are related by related by*

$$\ddot{A}_{ij} \geq \frac{1}{2} A_{ij}. \tag{69}$$

Proof. Let us examine the transition from state a to state b. There are two cases:

Case 1, $C(i) \geq C(j)$.

In this case, $A_{ij} = 1$. Let ϕ_i and ϕ_j be the probability densities of random variables $\ddot{C}(i)$ and $\ddot{C}(j)$. Restating (64), the transition probability under $\ddot{C}$ is

$$
\begin{aligned}
A_{ij} &= \int_{-\infty}^{\infty} \phi_i(x) \int_{-\infty}^{x} \phi_j(y) \cdot 1 \, dy \, dx \\
&\quad + \int_{-\infty}^{\infty} \phi_i(x) \int_{x}^{\infty} \phi_j(y) e^{-(y-x)/T} dy \, dx
\end{aligned} \tag{70}
$$

If $C(i) \geq C(j)$ then the first term always exceeds $1/2$. Therefore, $\ddot{A}_{ij} \geq \frac{1}{2} A_{ij}$.

Case 2, $C(i) < C(j)$.

Now, $A_{ij} = e^{-(C(j)-C(i))/T}$. We restate (70) slightly differently for this case:

$$
\begin{aligned}
A_{ij} &= \int_{-\infty}^{\infty} \phi_i(x) \int_{-\infty}^{x} \phi_j(y) \cdot 1 \, dy \, dx \\
&+ \int_{-\infty}^{\infty} \phi_i(x) \int_{x}^{x+C(j)-C(i)} \phi_j(y) e^{-(y-x)/T} dy \, dx \\
&+ \int_{-\infty}^{\infty} \phi_i(x) \int_{x+C(j)-C(i)}^{\infty} \phi_j(y) e^{-(C(j)-C(i))/T} dy \, dx
\end{aligned} \tag{71}
$$

Examine the first term of (71), and note that

$$
\begin{aligned}
& e^{-(C(j)-C(i))/T} \int_{-\infty}^{\infty} \phi_i(x) \int_{-\infty}^{x} \phi_j(y) \, dy \, dx \\
& \leq \int_{-\infty}^{\infty} \phi_i(x) \int_{-\infty}^{x} \phi_j(y) \, dy \, dx.
\end{aligned} \tag{72}
$$

Examine the second term of (71). We assert that

$$
\begin{aligned}
& \int_{-\infty}^{\infty} \phi_i(x) \int_{x}^{x+C(j)-C(i)} \phi_j(y) e^{-(C(j)-C(i))/T} dy \, dx \\
& \leq \int_{-\infty}^{\infty} \phi_i(x) \int_{x}^{x+C(j)-C(i)} \phi_j(y) e^{-(y-x)/T} dy \, dx
\end{aligned} \tag{73}
$$

Combining (71) with (72) and (73), we obtain

$$
\begin{aligned}
\ddot{A}_{ij} &\geq e^{-(C(j)-C(i))/T} \int_{-\infty}^{\infty} \phi_i(x) \int_{-\infty}^{x+C(j)-C(i)} \phi_j(y) \cdot 1 \, dy \, dx \\
&\quad + \int_{-\infty}^{\infty} \phi_i(x) \int_{x+C(j)-C(i)}^{\infty} \phi_j(y) e^{-(y-x)/T} dy \, dx
\end{aligned} \tag{74}
$$

Since the means of both ϕ_i and ϕ_j are less than $x + C(j) - C(i)$, the first term is no less than $\frac{1}{2} e^{-(C(j)-C(i))/T}$. Thus we conclude

$$
\ddot{A}_{ij} \geq \frac{1}{2} A_{ij}. \tag{75}
$$

□

LEMMA 4.6. *Consider two annealing chains P and $\ddot{P}$, with the same generation probability matrix G and state space S. Chain P has cost-function C, and chain $\ddot{P}$ has Gaussian random cost-function $\ddot{C}$. The two cost-functions are related by* $\mathrm{E}[\ddot{C}(s)]C(s)$. *The variance of each random variable $\ddot{C}(s)$, σ_s^2, is constrained by $\underline{\sigma}^2 \leq \sigma_s^2 \geq \overline{\sigma}^2$. Then the conductances of a subset $V \in S$ in the two chains are related by*

$$\frac{1}{2}e^{(\underline{\sigma}^2-\overline{\sigma}^2)/T^2}\Phi_V \leq \ddot{\Phi}_V \leq e^{(2\overline{\sigma}^2-\underline{\sigma}^2)/T^2}\Phi_V. \tag{76}$$

Proof. By definition,

$$\ddot{\Phi}_V = \frac{\sum_{i\in V}\sum_{j\notin V}\ddot{\boldsymbol{\pi}}_i G_{ij}\ddot{A}_{ij}}{\sum_{i\in V}\ddot{\boldsymbol{\pi}}_i}. \tag{77}$$

By Theorem 3.6 and Lemma 4.4,

$$\ddot{\Phi}_V \leq \frac{\sum_{i\in V}\sum_{j\notin V}\boldsymbol{\pi}_i e^{(\overline{\sigma}^2-\underline{\sigma}^2)/2T^2}G_{ij}A_{ij}e^{\overline{\sigma}^2/T^2}}{\sum_{i\in S}e^{(\underline{\sigma}^2-\overline{\sigma}^2)/2T^2}\boldsymbol{\pi}_i}. \tag{78}$$

By simple algebra we obtain

$$\ddot{\Phi}_V \leq e^{(2\overline{\sigma}^2-\underline{\sigma}^2)/T^2}\Phi_V. \tag{79}$$

By Theorem 3.6 and Lemma 4.5,

$$\frac{\sum_{i\in V}\sum_{j\notin V}\boldsymbol{\pi}_i e^{(\underline{\sigma}^2-\overline{\sigma}^2)/2T^2}G_{ij}\frac{1}{2}A_{ij}}{\sum_{i\in S}e^{(\overline{\sigma}^2-\underline{\sigma}^2)/2T^2}\boldsymbol{\pi}_i} \leq \ddot{\Phi}_V \tag{80}$$

By simple algebra we obtain

$$\frac{1}{2}e^{(\underline{\sigma}^2-\overline{\sigma}^2)/T}\Phi_V \leq \ddot{\Phi}_V. \tag{81}$$

Combining (79) and (81), we obtain (76). □

THEOREM 4.7. *Consider two annealing chains P and $\ddot{P}$, with the same generation probability matrix G and state space S. Chain P has cost-function C, and chain $\ddot{P}$ has Gaussian random cost-function $\ddot{C}$. The two cost-functions are related by* $\mathrm{E}[\ddot{C}(s)] = C(s)$. *The variance of each random variable $\ddot{C}(s)$, σ_s^2, is constrained by $\underline{\sigma}^2 \leq \sigma_s^2 \geq \overline{\sigma}^2$. Then the global conductances of the two chains are related by*

$$\frac{1}{2}e^{(\underline{\sigma}^2-\overline{\sigma}^2)/T^2}\Phi \leq \ddot{\Phi} \leq e^{(3\overline{\sigma}^2-2\underline{\sigma}^2)/T^2}\Phi. \tag{82}$$

Proof. By definition,

$$\ddot{\Phi} = \min_{V\in\ddot{\mathbf{S}}_{1/2}}\ddot{\Phi}_V. \tag{83}$$

Choose any $V \in \ddot{\mathbf{S}}_{1/2}$. There are two cases.

Case 1, $V \in \mathbf{S}_{1/2}$.

Let $W_V = V$. By Lemma 4.6,

$$\frac{1}{2} e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2} \Phi_{W_V} \leq \ddot{\Phi}_V. \tag{84}$$

This ends case 1.

Case 2, $V \notin \mathbf{S}_{1/2}$.

Let $W_V = S \setminus V$. If $V \notin \mathbf{S}_{1/2}$, then $\boldsymbol{\pi}_V > 1/2$. Therefore, $1 - \boldsymbol{\pi}_V = \boldsymbol{\pi}_{W_V} < 1/2$ and $W_V \in \mathbf{S}_{1/2}$.

By Lemma 4.3

$$\Phi_V \boldsymbol{\pi}_{W_V} = \Phi_{W_V} \boldsymbol{\pi}_V, \tag{85}$$

therefore

$$\Phi_{W_V} \frac{\boldsymbol{\pi}_V}{\boldsymbol{\pi}_{W_V}} = \Phi_V. \tag{86}$$

$\boldsymbol{\pi}_{W_V} < 1/2$ and $\boldsymbol{\pi}_V > 1/2$, so

$$\Phi_{W_V} \frac{(1/2)}{(1/2)} < \Phi_V. \tag{87}$$

By Lemma 4.6,

$$\frac{1}{2} e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2} \Phi_{W_V} < \ddot{\Phi}_V. \tag{88}$$

This ends case 2.

Let $\mathbf{S}_W = \{W_V : V \in \ddot{\mathbf{S}}_{1/2}\}$, and let

$$\Phi_{\mathbf{S}_W} = \min_{W \in \mathbf{S}_W} \Phi_{W_V}. \tag{89}$$

Then by (84), (88), and the definition of $\ddot{\Phi}$,

$$\frac{1}{2} e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2} \Phi_{\mathbf{S}_W} < \ddot{\Phi} \tag{90}$$

But $\mathbf{S}_W \subseteq \mathbf{S}_{1/2}$, so by definition of $\mathbf{S}_{1/2}$,

$$\frac{1}{2} e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2} \Phi \leq \frac{1}{2} e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2} \Phi_{\mathbf{S}_W} < \ddot{\Phi} \tag{91}$$

This proves the leftmost inequality in (82).

Now we prove the rightmost inequality. This second half of the proof follows the reasoning of the first. We provide it because the factors and results are slightly different.

By definition,

$$\Phi = \min_{V \in \mathbf{S}_{1/2}} \Phi_V. \tag{92}$$

Choose any $V \in \mathbf{S}_{1/2}$. There are two cases.

Case 1, $V \in \ddot{\mathbf{S}}_{1/2}$.

Let $X_V = V$. By Lemma 4.6,

$$\ddot{\Phi}_V \leq e^{(2\overline{\sigma}^2 - \underline{\sigma}^2)/T^2} \Phi_V. \tag{93}$$

This ends case 1.

Case 2, $V \notin \ddot{\mathbf{S}}_{1/2}$.

Let $X_V = S \setminus V$. If $V \notin \ddot{\mathbf{S}}_{1/2}$, then $\ddot{\boldsymbol{\pi}}_V > 1/2$. Therefore, $1 - \ddot{\boldsymbol{\pi}}_V = \ddot{\boldsymbol{\pi}}_{X_V} < 1/2$ and $X_V \in \ddot{\mathbf{S}}_{1/2}$.

By Lemma 4.3

$$\ddot{\Phi}_V \ddot{\boldsymbol{\pi}}_{X_V} = \ddot{\Phi}_{X_V} \ddot{\boldsymbol{\pi}}_V, \tag{94}$$

therefore

$$\ddot{\Phi}_{X_V} \frac{\ddot{\boldsymbol{\pi}}_V}{\ddot{\boldsymbol{\pi}}_{X_V}} = \ddot{\Phi}_V. \tag{95}$$

$\ddot{\boldsymbol{\pi}}_{X_V} < 1/2$ and $\ddot{\boldsymbol{\pi}}_V > 1/2$, so

$$\ddot{\Phi}_{X_V} \frac{(1/2)}{(1/2)} < \ddot{\Phi}_V. \tag{96}$$

By Lemma 4.6,

$$\ddot{\Phi}_{X_V} < e^{(2\overline{\sigma}^2 - \underline{\sigma}^2)/T^2} \Phi_V. \tag{97}$$

This ends case 2.

Let $\ddot{\mathbf{S}}_X = \{X_V : V \in \mathbf{S}_{1/2}\}$, and let

$$\Phi_{\ddot{\mathbf{S}}_X} = \min_{X \in \ddot{\mathbf{S}}_X} \ddot{\Phi}_{X_V}. \tag{98}$$

Then by (93), (97), and the definition of Φ,

$$\Phi_{\ddot{\mathbf{S}}_X} < e^{(2\overline{\sigma}^2 - \underline{\sigma}^2)/T^2} \Phi \tag{99}$$

But $\ddot{\mathbf{S}}_X \subseteq \ddot{\mathbf{S}}_{1/2}$, so by definition of $\ddot{\mathbf{S}}_{1/2}$,

$$\ddot{\Phi} \leq \Phi_{\ddot{\mathbf{S}}_X} < e^{(2\overline{\sigma}^2 - \underline{\sigma}^2)/T^2} \Phi. \tag{100}$$

With (91) and (100), we have proved the theorem. □

How do we interpret this result? Φ is equivalent to the speed at which the accurate annealing system equilibrates. $\ddot{\Phi}$ is the speed at which the inaccurate system equilibrates. Suppose we have an annealing schedule which keeps the accurate system at temperature T for time t_T. To guarantee the same level of equilibration in the inaccurate case, we must keep the system at temperature T for time

$$\ddot{t}_T = 2e^{(\overline{\sigma}^2 - \underline{\sigma}^2)/T^2} t_T \tag{101}$$

5. *Convergence on General Spaces*

On ergodic spaces, simulated annealing will provably converge monotonically to an optimum, with an appropriate temperature schedule. The most general results don't constrain the state space, other than by (3)–(7). We pay a time penalty for generality. The $T(t) = d/\log t$ temperature schedule, annealing a reversible graph, converges to a minimum cost state [11]. Tsitsiklis generalized this result to non-reversible graphs [24].

We first present a result by Tsitsiklis without proof: interested readers can refer to the original paper. We then prove that constraining errors to a constant factor of temperature guarantees convergence.

Define global optima set S_0 such that $s \in S_0 \Rightarrow C(s_0) = \min\{C(s')|s' \in S\}$. For state space S, generation matrix G, and cost-function C, construct a delta matrix D^0 as follows.

$$D^0_{ij} = \begin{cases} \infty & \text{if } G_{ij} = 0 \\ C(j) - C(i) & \text{if } G_{ij} \neq 0 \text{ and } C(j) \geq C(i) \\ 0 & \text{otherwise} \end{cases} \tag{102}$$

If (3)–(7) are satisfied, D^0 captures enough information to compute equilibrium probabilities at any fixed temperature.

Now, construct the set of *transient states* $\not R$ so that

$$\not R = \{i \in S | \exists j \in S \text{ such that } D^0_{ij} = 0 \text{ and } D^0_{ji} \neq 0\} \tag{103}$$

All other states, $R = S \setminus \not R$, are termed *recurrent*. These recurrent states are the local minima. We can construct a partitioning over R: for any $i \in R$, its equivalency class is $R_i = \{j \in R | D^0_{ij} = 0\}$.

Figure 5 shows a pruning algorithm. At pruning depth $d+1$, we eliminate local minima of depth d. Informally, we measure the depth by starting at some $i \in R$, and following the shortest path to a lower local minimum j. The difference between the greatest cost reached on the path and $C(i)$, is the depth of i.

Hajek describes this process as "how cups runneth over": the state variable has to "run over" some boundary to escape a local minima. The higher the boundary, the longer it takes to crest.

The recurrent state sets constructed by this algorithm have important properties. First, $R^{d+1} \subset R^d$. Second, if the system satisfies (3)–(7), then some $d' \in \mathbb{Z}^{0+}$ produces $R^{d'} = S_0$, the set of all minimum cost states.

Figure 6 shows an example space, and its delta matrix. We prune the annealing space by iterating with *Prune Minima*. The first iteration yields $\not R^0 = \not R, R^0 = R, C^0$, and D^1.

$$\not R^0 = \{b, d\}, \quad R^0 = \{a, c, e\} \tag{107}$$

Input: $N \times N$ matrix D^d.
Output: $N \times N$ matrices D^{d+1}, sets $\not R^d, R^d \subset S$, equivalency sets R_i^d.

1. Find the transient states $\not R^d$, recurrent states R^d, and equivalence classes R_i^d of D^d.
2. Let

$$(104) \qquad C_{ij} = \begin{cases} D_{ij}^d - 1 & \text{if } i,j \in R^d \text{ and } j \notin R_i^d \\ D_{ij}^d & \text{otherwise} \end{cases}$$

3. For all $i \in R^d, j \in R^d$ solve the shortest path problem from i to j over D^d. Let D_{ij}^{d+1} be its length. (Note that $i \in R^d \Rightarrow D_{ii}^{d+1} = 0$.)
4. For all $i \in R^d, j \in \not R^d$, let

$$(105)\ D_{ij}^{k+1} = \min_{k \in R^d} \{D_{ik}^{d+1} + C_{kj}\} = \min_{k \in R^d} \{D_{ik}^{d+1} + D_{kj}^d\}$$

5. If $i \in \not R^k$, let

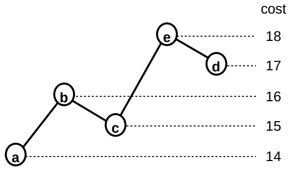
$$(106)\ D_{ij}^{k+1} = \min_{k \in R^d} \{C_{ik} + D_{kj}^{d+1}\} = \min_{k \in R^d} \{D_{ik}^d + D_{kj}^{d+1}\}$$

FIG. 5. Prune Minima *Algorithm*

$$(108)\ C^0 = \begin{bmatrix} \infty & 2 & \infty & \infty & \infty \\ 0 & \infty & 0 & \infty & \infty \\ \infty & 1 & \infty & 3 & \infty \\ \infty & \infty & 0 & \infty & 0 \\ \infty & \infty & \infty & 1 & \infty \end{bmatrix}, \quad D^1 = \begin{bmatrix} 0 & 2 & 2 & 5 & 5 \\ 0 & 1 & 0 & 3 & 3 \\ 1 & 1 & 0 & 3 & 3 \\ 1 & 1 & 0 & 1 & 0 \\ 2 & 2 & 1 & 1 & 0 \end{bmatrix}$$

The second iteration yields $\not R^1, R^1, C^1$, and D^2.

$$(109) \qquad \not R^1 = \{b, d\}, \quad R^1 = \{a, c, e\}$$



$$D^0 = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{bmatrix} \infty & 2 & \infty & \infty & \infty \\ 0 & \infty & 0 & \infty & \infty \\ \infty & 1 & \infty & 3 & \infty \\ \infty & \infty & 0 & \infty & 0 \\ \infty & \infty & \infty & 1 & \infty \end{bmatrix} \end{array}$$

FIG. 6. *Pruning Example*

$$
(110)\; C^1 = \begin{bmatrix} 0 & 2 & 1 & 5 & 4 \\ 0 & 1 & 0 & 3 & 3 \\ 0 & 1 & 0 & 3 & 2 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 \end{bmatrix}, \quad D^2 = \begin{bmatrix} 0 & 2 & 1 & 4 & 3 \\ 0 & 1 & 0 & 3 & 2 \\ 0 & 1 & 0 & 3 & 2 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}
$$

The third iteration yields

$$
(111) \qquad \not R^2 = \{b, c, d, e\}, \;\; R^2 = \{a\}, \;\; C^2 = D^2, \;\; D^3 = D^2.
$$

Further iterations produce no changes. Thus, $R^2 = \{a\}$ is the set of global minima for our example, and the maximum depth of local minima is 1.

Suppose that we are given a stochastic matrix P^ε (whose ijth entry is denoted by p^ε_{ij}), parameterized by a positive parameter ε and assume that there exist positive constances C_1, C_2, and a collection $\mathcal{A} = \{\alpha_{ij} : 1 \leq i, j \leq N\}$ such that $\alpha_{ij} \in \mathcal{N}_0 \cup \{\infty\}, \forall i, j$ and such that $p^\varepsilon_{ij} = 0$ whenever $\alpha_{ij} = \infty$, and $C_1 \varepsilon^{\alpha_{ij}} \leq p^\varepsilon_{ij} \leq C_2 \varepsilon^{\alpha_{ij}}$ whenever $\alpha_{ij} < \infty$. Finally, we are given a monotonically nonincreasing function (cooling schedule) $\varepsilon : \mathcal{N}_0 \rightarrow (0,1)$. We are interested in the Markov chain $X(t)$ with transition probabilities given by $P(X(t+1) = j | X(t) = i) = p^\varepsilon_{ij}$.

Simulated annealing satisfies the preceding paragraph, if we let $\varepsilon = e^{-1/T(t)}$, and if

$$
(112) \qquad \alpha_{ij} = \begin{cases} \infty & \text{if } G_{ij} = 0 \\ \max\{0, C(j) - C(i)\} & \text{otherwise} \end{cases}
$$

Classify the states as recurrent or transient thus: State i is transient iff $\exists j \in S, \alpha_{ij} = 0 \wedge \alpha_{ji} \neq 0$. Otherwise, i is recurrent.

Assumption: If at least one of the states i, j, and k are recurrent, we have $\alpha_{ik} \leq \alpha_{ij} + \alpha_{jk}$.

THEOREM 5.1 (Tsitsiklis). *Assume that for some integer $d \geq 0$,*

$$
(113) \qquad \sum_{t=0}^{\infty} \varepsilon^d(t) = \infty
$$

$$
(114) \qquad \sum_{t=0}^{\infty} \varepsilon^{d+1}(t) < \infty
$$

Then

1. $\forall i \in S, \lim_{t\rightarrow\infty} P(X(t) \in R^d | X(0) = i) = 1.$
2. *For any* $i \in R^d$, $\limsup_{t\rightarrow\infty} P(X(t) = i | X(0) = i) > 0.$

COROLLARY 5.2 (Tsitsiklis). *Let the transition probabilities for simulated annealing be given by (1) and (2). Consider temperature schedules of the form $T(t) = c/\log t$. For any initial state, the algorithm converges in probability to the set of global minima, S_0, if and only if there exists some d such that $R^d \subset S_0$, and c is larger than or equal to the smallest such d, to be denoted by d^*.*

THEOREM 5.3. *Suppose a inaccurate simulated annealing chain satisfies (3)–(7) and has a cost-function $\dot{C}(s,t)$, with time-dependent errors*

$$C(s) + \underline{\epsilon}(t) \leq \dot{C}(s,t) \leq C(s) + \overline{\epsilon}(t). \tag{115}$$

Let S_0 be the set of minimum cost states in S.

Let $b_1 T(t) \leq \underline{\epsilon}(t) \leq \overline{\epsilon}(t) \leq b_2 T(t)$, where b_1 and b_2 are constants. Let the temperature schedule be of the form $T(t) = d/\log t$, where $d \in \mathbb{Z}^{0+}$. Then the algorithm converges in probability to the set of global minima if $R^d \subset S_0$.

Proof. Suppose the transition matrix P^t for some Markovian system at time t is constrained by (116).

$$c_1 e^{-D^0_{ij}/T(t)} \leq P^t_{ij} \leq c_2 e^{-D^0_{ij}/T(t)}, \tag{116}$$

where c_1 and c_2 are positive constants. Assume for some integer $d \geq 0$ that (117) and (118) hold.

$$\sum_{t=0}^{\infty} e^{-d/T(t)} = \infty \tag{117}$$

$$\sum_{t=0}^{\infty} e^{-d-1/T(t)} < \infty \tag{118}$$

We then conclude by Theorem 5.2 that (119) and (120) are true.

$$\forall i \in S, \ \lim_{t\to\infty} P[X(t) \in R^d | X(0) = i] = 1 \tag{119}$$

$$\forall i \in R^d, \ \limsup_{t\to\infty} P[X(t) = i | X(0) = i] > 0 \tag{120}$$

Let $c_1 = e^{-b_1}$ and $c_2 = e^{-b_2}$. These values satisfy (115) and (116).

Let d satisfy $R^d \subset S_0$. Such a d must exist, since (3)–(7) are satisfied. Let $T(t) = d/\log t$. This satisfies (117) and (118). By (119) the inaccurate simulated annealing algorithm converges in probability to the set of global minima. □

This result shows that if errors are constrained above and below, by constant factors of the temperature, annealing under a $c/\log t$ temperature schedule will converge to an optimum. This result applies to any state space satisfying (3)–(7).

The convergence time makes this schedule uninteresting for practical matters. Its primary advantage, and the reason it appears in so many simulated annealing papers, is that it applies generally and yields theoretical results.

Recent theory papers have restricted the annealing space and obtained less tedious schedules. One such framework appears in the next section. It too shows a dependence between temperature and the restrictions which must be imposed on the errors.

6. *Deterministic Fractal Spaces*

Sorkin has postulated that many common annealing spaces have a fractal structure. This characterization appeals because a geometric temperature schedule, ubiquitous in annealing programs, provably drives annealing to the optimum in a fractal space. Statistical comparisons suggest that the more closely an optimization space approximates a fractal structure, the more likely simulated annealing will find a good solution [22].

In this section, we analyze the performance of inaccurate simulated annealing on a deterministic fractal space. Following Sorkin's explication, suppose we have integer base $b > 0$, scaling factor $r \in (0, 1)$, and arbitrary, fixed function $F: \{0, \ldots, b-1\} \to [0, 1]$. The *a-complement function* of x, $\mathrm{comp}_a(x)$, is defined below.

$$\mathrm{comp}_a(x) = \begin{cases} x & \text{if } a \text{ is even,} \\ 1 - x & \text{otherwise.} \end{cases} \tag{121}$$

We construct a cost-function $C: [0, 1] \to [0, 1/(1-r)]$ as follows. Represent state s in base b as $0.s_1 s_2 \cdots$. If state s has two representations, such as $0.0011111\cdots$ and 0.01 for $b = 2$, use the terminating one. Then

$$C(s) = F(s_1) + r \cdot C(\mathrm{comp}_{s_1}(0.s_2 s_3 \cdots)). \tag{122}$$

Thus, we have $C: [0, 1] \to [0, 1/(1-r)]$.

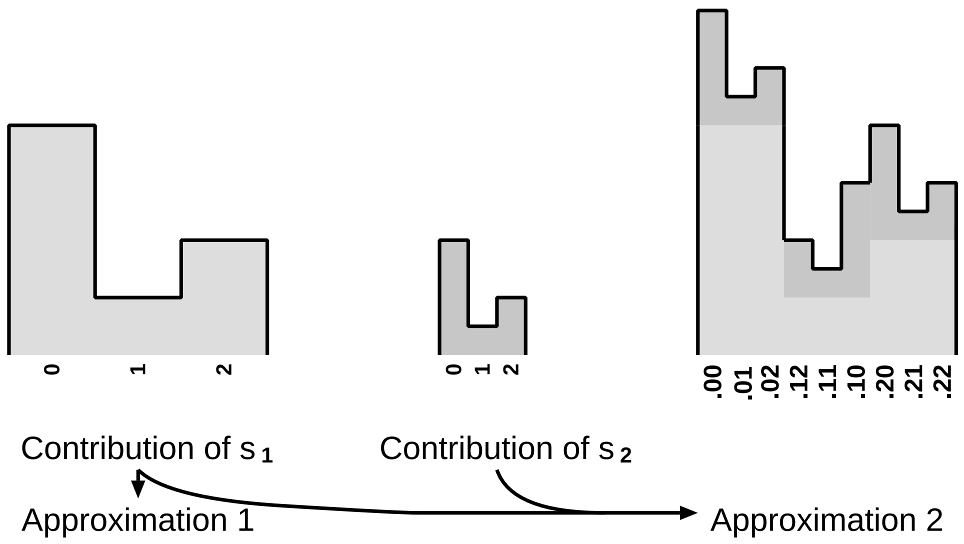


FIG. 7. *Base 3 Fractal Example*

Spaces of this form exhibit a deterministic fractal structure. Figure 7 shows a simple example with base $b = 3$ and scaling factor $r = 0.5$. The contribution of first digit s_1 is added to the contribution of s_2 to form an approximation for the cost of s. If we compute the approximation of $C(s)$

using digits s_1 through s_k (i.e., presuming $C(\mathrm{comp}_{s_k}(0.s_{k+1}\cdots)) = 0$), the value we obtain is off by no more than

$$\sum_{i=k+1}^{\infty} r^i = \frac{r^k+1}{1-r}. \tag{123}$$

(122) constructs the state space to anneal, but we have not presented a generating function. This analysis considers two generating functions, for confined and unconfined annealing. Confined annealing is unrealistically simple, but results for confined annealing will apply to unconfined annealing.

Suppose that we can perturb a single digit s_k of state s, by adding or subtracting b^{-k-1}, to generate a new state s'. If the perturbation would cause a carry or borrow into digit s_{k-1} we move back to the original digit. Formally, the move set for state s at stage k is

$$S_k^{\mathbf{c}}(s) = \begin{cases} \{s - b^{-k-1}, s + b^{-k-1}\} & \text{if } 0 < s_k < b-1, \\ \{s, s + b^{-k-1}\} & \text{if } s_k = 0, \\ \{s - b^{-k-1}, s\} & \text{if } s_k = b-1. \end{cases} \tag{124}$$

The corresponding move generating function is

$$\mathtt{generate}_k^{\mathbf{c}}(s) \equiv \text{Choose randomly from } S_k^{\mathbf{c}} \tag{125}$$

Using this generating function produces what Sorkin calls confined annealing, so named because moves at stage k are confined by the first $k-1$ digits in the state.

By contrast, unconfined annealing has a less restricted move set,

$$S_k^{\mathbf{u}}(s) = \begin{cases} \{s - b^{-k-1}, s + b^{-k-1}\} & \text{if } b^{-k-1} \leq s \leq 1 - b^{-k-1}, \\ \{s, s + b^{-k-1}\} & \text{if } s < b^{-k-1}, \\ \{s - b^{-k-1}, s\} & \text{if } s > 1 - b^{-k-1}. \end{cases} \tag{126}$$

It has the obvious generating function $\mathtt{generate}_k^{\mathbf{u}}$. An example showing the differences between confined and unconfined annealing appears in Figure 8.

Our algorithm will change its generating function, starting with $\mathtt{generate}_0$ and incrementing the subscript with each change in temperature. This feature is not exotic: tying the move scale to temperature has been common since the earliest years of simulated annealing [25].

6.1. *Confined Annealing*

First, let us presume each state exhibits a fixed error. That is, that there is a single-valued function $\varepsilon\colon S \to \mathbb{R}$. Further presume that the accurate state space has a fractal structure. We can transform it into a new problem of the same class.

Definition 6.1. *The total variational distance between two probability vectors P and P' is*

$$\|P' - P\|_{\mathrm{tvd}} = \frac{1}{2}\sum_{v\in V} |P'_v - P_v| \tag{127}$$

or equivalently,

$$\|P' - P\|_{\text{tvd}} = \sum_{v \in V : P'_v > P_v} |P'_v - P_v| \tag{128}$$

LEMMA 6.2. *If P_1, P_2, and P_3 are probability vectors on the same space, then*

$$\|P_1 - P_3\|_{\text{tvd}} \leq \|P_1 - P_2\|_{\text{tvd}} + \|P_2 - P_3\|_{\text{tvd}} \tag{129}$$

Proof. Apply the definition of $\|_{\text{tvd}}$ and the triangle inequality for absolute value. □

DEFINITION 6.3. *The minimum non-zero value of function $f: V \to \mathbb{R}$ is defined by*

$$\Delta f = \min_{v \in V} f(v) \tag{130}$$

LEMMA 6.4. *Let $\gamma = (\overline{\epsilon} - \underline{\epsilon})$. If $\gamma \leq -T \ln(1 - \varepsilon)$, then $\|\pi - \pi_\gamma\|_{\text{tvd}} \leq \varepsilon$.*

Proof. Let $a_v = \boldsymbol{\pi}_{\gamma v} / \boldsymbol{\pi}_v$. By our previous definition of $\boldsymbol{\pi}$, we have

$$\begin{aligned}
\|\pi - \pi_\gamma\|_{\text{tvd}} &= \frac{1}{2} \sum_{v \in V} |\boldsymbol{\pi}_v - \boldsymbol{\pi}_{\gamma v}| && (131)\\
&= \frac{1}{2} \sum_{v \in V} |\boldsymbol{\pi}_v - a_v \boldsymbol{\pi}_v| && (132)\\
&= \frac{1}{2} \sum_{v \in V} |(1 - a_v) \boldsymbol{\pi}_v| && (133)
\end{aligned}$$

By Theorem 3.1, we have $e^{-\gamma/T} \leq a_v \leq e^{\gamma/T}$. To maximize the effect of a_v, we reduce some $\boldsymbol{\pi}_v = 1$ by a factor of $e^{-\gamma/T}$ and raise some other $\boldsymbol{\pi}_v = 0$ correspondingly by a factor of $e^{\gamma/T}$.

So we have

$$\|\pi - \pi_\gamma\|_{\text{tvd}} \leq \frac{1}{2} \cdot 2 \cdot (1 - e^{-\gamma/T}) \tag{134}$$

By our premise, we have

$$1 - e^{-\gamma/T} \leq \varepsilon \tag{135}$$

With (135), (131), and transitivity of $\leq$, we prove it. □

Here we state several results from [22] without proof.:

LEMMA 6.5 (Sorkin 3.8.1). *For a function $f: V \to [f_{\min}, f_{\max}]$ with $f_{\text{range}} = f_{\max} - f_{\min}$, an arbitrary distribution P_t on V, and π_T and π_0 the stationary distributions at temperatures T and 0:*

$$\mathbf{E}_{P_t}[f] \leq \mathbf{E}_{\pi_T}[f] + \|P_t - \pi_T\|_{\text{tvd}} \cdot f_{\text{range}} \tag{136}$$

DEFINITION 6.6. *For an annealing graph G with b vertices let $\hat{T}(\varepsilon, \Delta, b) = \Delta / \ln(b^2/\varepsilon)$. If G is regular, substitute b for b^2.*

LEMMA 6.7 (Sorkin 3.8.4). *Given Δ, let $T \leq \hat{T}(\varepsilon, \Delta, b)$. If $\Delta \leq \varepsilon$ or $\Delta \leq \min\{C(v) | C(v) > \varepsilon\}$, then*

$$\mathbf{E}_{\pi_T}[C] \leq 2\varepsilon \tag{137}$$

and if $\Delta \leq \Delta C$ as defined in Definition 6.3, then

$$\mathbf{E}_{\pi_T}[C] \;\leq\; \|\pi_T - \pi_0\|_{\text{tvd}} \;\leq\; \varepsilon. \tag{138}$$

DEFINITION 6.8. *For an annealing graph G with b vertices and for a given temperature T, let*

$$\hat{t}(T, \varepsilon, b) = 2\left(\ln\left(\frac{b^2}{\varepsilon}\right) + \frac{1}{T}\right)\frac{1}{\Phi^2(\infty)}e^{2/T} \tag{139}$$

If G is regular, substitute b for b^2. Also, if G is regular we may substitute $1/b$ for $\Phi(\infty)$.

LEMMA 6.9 (Sorkin 3.8.6). *If $t \geq \hat{t}(T, \varepsilon, n)$ then beginning from any distribution P_0 and annealing at temperature T for time t, the outcome distribution satisfies*

$$\|P_t - \pi_T\|_{\text{tvd}} \leq \varepsilon \tag{140}$$

We will need to fix the temperature until the total variational distance from the present distribution to the stationary distribution is less than some distance $\varepsilon \in [0, 1]$.

THEOREM 6.10 (Sorkin 3.8.7). *Consider a graph G with b vertices, a cost-function C from G to $\mathbb{R}$ having minimum 0 and maximum 1, and a small value ε. Let Δ be any of ΔC, ε, or $\min\{C(v) \colon C(v) > \varepsilon\}$. With the functions $\hat{T}$ and $\hat{t}$ as defined above, begin from an arbitrary initial distribution P_0 and anneal at temperature $T \leq \hat{T}(\varepsilon, \Delta, b)$ for $t \geq \hat{t}(T, \varepsilon, b)$ steps. If the outcome distribution is denoted P_t, we have:*

$$\begin{align} \mathbf{E}_{\pi_T}[C] &\leq 2\varepsilon \tag{141}\\ \|P_t - \pi_T\|_{\text{tvd}} &\leq \varepsilon \tag{142}\\ \mathbf{E}_{P_t}[C] &\leq 3\varepsilon \tag{143}\end{align}$$

If $\Delta = \Delta C$, the bounds in (141) and (143) can be improved to ε and 2ε respectively.

LEMMA 6.11. *Define*

$$T = \hat{T}_\gamma(\varepsilon, \varepsilon, b) = \hat{T}(\varepsilon, \varepsilon, b), \tag{144}$$

per Definition 6.6, and

$$t = \hat{t}_\gamma(T, \varepsilon, b) = \hat{t}(T, \varepsilon, b) \tag{145}$$

per Definition 6.8. Then

$$\|P_{\gamma tT} - \pi_{\gamma T}\|_{\text{tvd}} \leq \varepsilon, \tag{146}$$
$$\mathbf{E}_{\pi_{\gamma T}}[C] \leq 3\varepsilon, \quad \text{and} \tag{147}$$
$$\mathbf{E}_{P_{\gamma tT}}[C] \leq 4\varepsilon. \tag{148}$$

Proof. Let $\Delta = \varepsilon$ in Lemmas 6.7 and 6.9. We have (note the absence of γ here)

$$\mathbf{E}_{\pi_T}[C] \leq 2\varepsilon \tag{149}$$

and, the first result,

$$\|P_{\gamma tT} - \pi_{\gamma T}\|_{\text{tvd}} \leq \varepsilon. \tag{150}$$

By Lemma 6.4 we have

$$\|\pi_T - \pi_{\gamma T}\|_{\text{tvd}} \leq \varepsilon \tag{151}$$

Combining these using Lemma 6.2 we obtain

$$\|\pi_T - P_{\gamma T}\|_{\text{tvd}} \leq \|\pi_T - \pi_{\gamma T}\|_{\text{tvd}} - \|\pi_{\gamma T} - \pi_T\|_{\text{tvd}} \leq 2\varepsilon \tag{152}$$

Using Lemma 6.5, we obtain our second and third results,

$$\mathbf{E}_{\pi_{\gamma T}}[C] \leq \mathbf{E}_{\pi_T}[C] + \|\pi_T - \pi_{\gamma T}\|_{\text{tvd}}, \quad \text{and} \tag{153}$$
$$\mathbf{E}_{P_{\gamma tT}}[C] \leq \mathbf{E}_{\pi_T}[C] + \|\pi_T - P_{\gamma tT}\|_{\text{tvd}} \leq 4\varepsilon \tag{154}$$

□

DEFINITION 6.12. *Construct a fractal-error cost-function with parameter* $\gamma = (\overline{\epsilon} - \underline{\epsilon})$ *as follows: Let* F *be given for the accurate space. Define* F_γ *as any function such that*

$$F + \underline{\epsilon} \leq F_\gamma \leq F + \overline{\epsilon}. \tag{155}$$

We then construct the fractal-error cost-function C_γ *substituting* F_γ *for* F *in (122).*

THEOREM 6.13. *Let deterministic fractal* C *with base* b *and scale factor* r *be given. Let error-limitation* ε *satisfy* $0 < \varepsilon < 1$. *Compute* $\hat{T} = \hat{T}(\varepsilon, \Delta C, b)$ *and* $\hat{t} = \hat{t}(\varepsilon, \Delta C, b)$. *Apply confined annealing using fractal-error cost-function* C_γ *with cooling schedule* $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{t})$ *for* $k = 1, \cdots, K$ *and* $K = \lceil \ln \varepsilon / \ln r \rceil$. *Limit the errors in each stage by* $\gamma_k \leq -T_k \ln(1 - \varepsilon)$.

Then the outcome has expected cost, evaluated under the accurate cost-function,

$$\hat{C}_{\gamma\text{con}} = \frac{\mathbf{E}_\gamma[C]}{C_{\text{range}}} \leq 5\varepsilon \tag{156}$$

and the algorithm consumes run time

$$t_{\gamma\text{con}} = \lceil \ln\varepsilon / \ln r \rceil \cdot \hat{t}. \tag{157}$$

Proof. By Lemma 6.11, annealing with $(T_1, t_1) = (\hat{T}, \hat{t})$ in generation 1 gives $\mathbf{E}_{P_{\gamma\hat{t}_1}} \leq 4\varepsilon$. Using the similarity of S_k to S_1, annealing in generation k with

$$(T_k, t_1) = (r_{k-1}\hat{T}, \hat{t}) \tag{158}$$

results in a distribution for s_k satisfying

$$\mathbf{E}_{P_{\gamma t}}[F(s_k^{k,t_k})] \leq 4\varepsilon \tag{159}$$

Therefore, the expected cost of the outcome satisfies

$$\begin{aligned} \mathbf{E}_\gamma[C] &\leq \sum_{k=1}^{K} r^{k-1}\mathbf{E}_{P_{\gamma\hat{t}_1}}[F(s_k^{k,t_k})] + \sum_{k=K+1}^{\infty} r^{k-1} \cdot 1 && (160) \\ &= \frac{1}{1-r}(4\varepsilon + r^K) && (161) \end{aligned}$$

Since C ranges from 0 to $\sum_{k=0}^{\infty} r^k = 1/(1-r)$, the relative expected cost is

$$\mathbf{E}_{\gamma\text{con}} = \frac{\mathbf{E}_\gamma[C]}{C_{\text{range}}} \leq 4\varepsilon + r^K. \tag{162}$$

If we let $K = \lceil \ln\varepsilon / \ln r \rceil$, then we have $r^K \leq \varepsilon$. This yields the result. □

Our previous result was on a state space where both the accurate cost-function and the errors have a self-similar nature. This may have some basis in realistic annealing problems, where reducing the size of the moves reduces the errors, as well.

If the errors are not naturally self-similar, then we must impose artificial adjustments to make them so.

COROLLARY 6.14. *Redefine C_γ so that $C + \underline{\epsilon} \leq C_\gamma \leq C + \overline{\epsilon}$. If we limit the errors by $\gamma_k \leq -(T_k^2/\hat{T})\ln(1-\varepsilon)$, then the outcome has expected cost, under the accurate cost-function,*

$$\hat{C}_{\gamma\text{con}} = \frac{\mathbf{E}_\gamma[C]}{C_{\text{range}}} \leq 5\varepsilon \tag{163}$$

and the algorithm consumes run time

$$t_{\gamma\text{con}} = \lceil \ln\varepsilon / \ln r \rceil \cdot \hat{t}(\varepsilon). \tag{164}$$

6.2. *Unconfined Annealing*

This section will show that annealing on an unconfined fractal space has the same outcome-distribution as the confined fractal space, appealing to the Sorkin replica model.

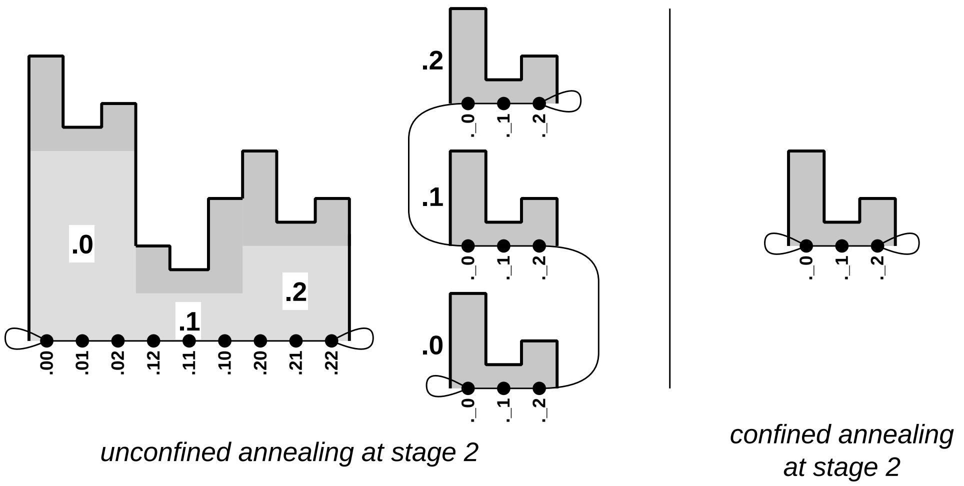


FIG. 8. *Unconfined annealing vs. Confined annealing*

The results for confined annealing encourage, but do not satisfy. In real annealing applications, a program cannot constrain moves to a given state segment. If it could, other algorithms would do a better job.

As it turns out, allowing the state to stray into an adjacent segment will not affect our result. The algorithm resolves the kth segment to with 5ε of the desired outcome in generation k, as in confined annealing. In subsequent generations, annealing continues on the kth segment with gradually decreasing temperature. The additional processing can only improve the kth segment, and so the worst-case bounds for confined annealing still hold.

The proof is virtually identical to Sorkin's. We refer the reader to §8.4 of [22], and merely state the result:

THEOREM 6.15. *Let value $0 < \varepsilon < 1$ satisfying $\hat{T} \le T_{\text{crit}}$, and deterministic fractal C with scale factor r be given. Let $\gamma \le -\ln(1-\varepsilon)$. Compute $\hat{T}$ and $\hat{t}$ as before. Apply unconfined annealing with cooling schedule $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{t})$ for $k = 1, \cdots, K$ and $K = \lceil \ln \varepsilon / \ln r \rceil$. Then the outcome has relative expected cost, evaluated with under the accurate cost-function, of*

$$\hat{C}_{\gamma\text{uncon}} = \frac{\mathbf{E}_\gamma[C]}{C_{\text{range}}} \le 5\varepsilon \tag{165}$$

and the algorithm consumes run time

$$t_{\gamma\text{uncon}} = \lceil \ln \varepsilon / \ln r \rceil \cdot \hat{t}. \tag{166}$$

Observe the proof outline from Figure 8. When performing unconfined annealing, transitions may occur from one large-scale feature to another (as from .21 to .11), even during small-scale moves, but these transitions do not affect the annealing process on the smaller scale—they behave identically to the end-loops on confined annealing. Therefore, unconfined annealing can be shown equivalent to confined annealing.

6.3. *Inaccurate Fractal: Equivalent Quality*

We have shown that constraining errors by $\gamma_k \leq -T_k \ln(1-\varepsilon)$ allows an inaccurate fractal system to converge. However, we have produced an outcome inferior to the accurate system: the inaccurate system reaches $\hat{C}_{\gamma\text{uncon}} \leq 5\varepsilon$, while the accurate system reaches $\hat{C}_{\text{uncon}} \leq 3\varepsilon$.

A more useful result determines the number of additional annealing steps needed to bring the inaccurate system to the same quality as the accurate system. We provide specifics below.

THEOREM 6.16. *If t is the total time required to reach $\hat{C}_{\text{uncon}} \leq 3\varepsilon$ with accurate annealing, and t' is the total time to reach $\hat{C}_{\gamma\text{uncon}} \leq 3\varepsilon$ under the inaccurate cost-function, then t and t' are related by the approximation*

$$\frac{t'}{t} \approx \left(\frac{5}{3}\right)^{2/\Delta} \tag{167}$$

Proof. Let $\varepsilon' = 3\varepsilon/5$. To obtain $\hat{C}_{\gamma\text{uncon}} \leq 3\varepsilon$ under inaccurate annealing, we must anneal for

$$t' = \lceil \ln \varepsilon' / \ln r \rceil \cdot \hat{t} \tag{168}$$

steps.

$$t' = \lceil \ln \frac{3\varepsilon}{5} / \ln r \rceil \cdot 2\left(\ln \frac{5b^2}{3\varepsilon} + \frac{1}{\Delta C / \ln \frac{5b^2}{3\varepsilon}}\right) e^{2\ln(5b^2/3\varepsilon)/\Delta} \tag{169}$$

$$= \lceil \ln \frac{3\varepsilon}{5} / \ln r \rceil \cdot 2\ln \frac{5b^2}{3\varepsilon}\left(1 + \frac{1}{\Delta C}\right)\left(5b^2/3\varepsilon\right)^{2/\Delta}. \tag{170}$$

Now, we have

$$\frac{t'}{t} = \frac{\lceil \ln \frac{3\varepsilon}{5} / \ln r \rceil \cdot 2\ln \frac{5b^2}{3\varepsilon}\left(1 + \frac{1}{\Delta C}\right)\left(5b^2/3\varepsilon\right)^{2/\Delta}}{\lceil \ln \varepsilon / \ln r \rceil \cdot 2\ln \frac{b^2}{\varepsilon}\left(1 + \frac{1}{\Delta C}\right)\left(b^2/\varepsilon\right)^{2/\Delta}} \tag{171}$$

$$\approx \left\lceil \frac{\ln \varepsilon + \ln \frac{3}{5}}{\ln \varepsilon} \right\rceil \frac{\ln \frac{b^2}{\varepsilon} + \ln \frac{5}{3}}{\ln \frac{b^2}{\varepsilon}} \left(\frac{5}{3}\right)^{2/\Delta}. \tag{172}$$

Suppose $5/3 \ll b^2/\varepsilon$ and $0 < \ln \epsilon \ll 3/5$, which is likely for a high-quality outcome (otherwise, why use simulated annealing). Then we can further approximate (172) as

$$\frac{t'}{t} \approx \left(\frac{5}{3}\right)^{2/\Delta}. \tag{173}$$

□

Note that (173) has $(\bar{\epsilon} - \underline{\epsilon}) \leq -T \ln(1 - \varepsilon)$ as a premise. Tightening error bounds would bring t'/t closer to one. Likewise, loosening error bounds would increase the ratio.

7. *Measuring Errors*

We have shown how errors and the temperature schedule interact to affect the outcome and speed of simulated annealing. It is rarely practical to analytically determine the error-range, instead we measure errors and estimate the error-range.

Annealing programs measure errors in one of two ways: First, by subtracting the true-cost from the computed-cost at each iteration. Second, by performing a series of iterations, accumulating a total observed cost-difference, and subtracting the total true cost-difference.

The first, called "instantaneous error," provides more detailed information about each error. The second, called "accumulated error," is more efficient, since the program computes the true-cost once per sequence, rather than once per iteration. Several researchers have performed accumulated error measurements [9, 1, 4, 12, 13]. I know of no published instantaneous error measurements.

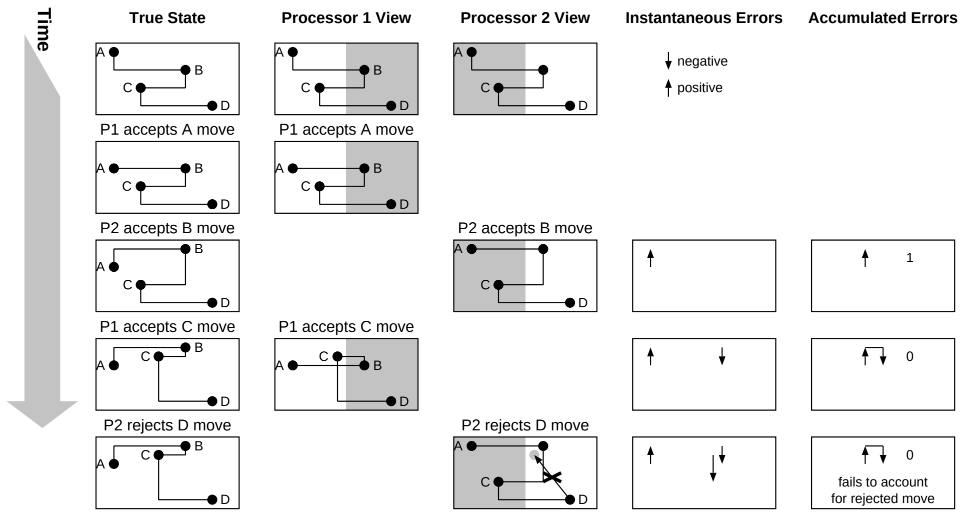


FIG. 9. *Observed Errors*

Figure 9 shows a four iteration annealing sequence performed by two processors. The first column shows the true-state of the system. The second column shows processor 1's moves, with local copies of processor 2 variables shaded. The third column shows processor 2's moves, with local copies of processor 1 variables shaded. The fourth column shows each instantaneous error, in graphical form. The fifth column shows accumulated errors.

The accumulated error column in Figure 9 does not account for the "P2 rejects D move." In fact, accumulated errors never include contributions from rejected moves. And, although there are two instantaneous errors in accepted moves, the final accumulated error is zero—the instantaneous errors cancelled each other out.

To measure instantaneous errors we must compute both true and observed-costs of each move, annihilating the speed advantage of the inaccurate cost-function. We might compute instantaneous errors during a test run, to predict the behavior of a class of annealing problems, but not during every production run.

Measuring accumulated errors is more efficient, but it does not help estimate the instantaneous error-range. We cannot divide a series' accumulated error by the iterations to get the average instantaneous error—the accumulated error does not include contributions from rejected moves. The average instantaneous error isn't useful anyway, we need a maximum and minimum for the range.

Though accumulated error measurements do not help estimate instantaneous errors, accumulated error-bounds work directly in the results we have established. Why? The composed moves in a sequence can be seen as a "big move," and the composed accept/reject decisions as a "big decision." The result is a Boltzmann distribution with the same properties as the original. The following remarks formalize this idea.

LEMMA 7.1. *Let $X = (S, G, C, T)$ be an annealing chain, with state-set S, generation probability matrix G, cost-function C, and fixed temperature T, which satisfies (3)–(7). Choose a sequence-length n, and assume $G' = \prod_{i=1}^{n} G$ has the coverage property (4). One can construct another annealing chain $X' = (S, G', C, T)$, Then, X and X' have the same equilibrium probability density.*

Proof. We first prove that X' satisfies the annealing properties (3)–(7).

Probability (3): G is a transition probability matrix. The composition of transition probability matrices is a transition probability matrix, so G' is a transition probability matrix.

Coverage (4): By definition of G'.

Aperiodicity (5): If $\exists s, s' \in S, C(s) \neq C(s')$, then (2) ensures aperiodicity. Otherwise, all costs are equal, and the acceptance matrix defined by (1 must contain all 1s. By (2), the Markov chains $P' = G'$ and $P = G$, and G must itself be aperiodic. The composition of aperiodic matrices is aperiodic, so G' is aperiodic. Therefore, P' is aperiodic.

Finiteness (6): By definition of X'.

Symmetry (7): The composition of symmetric matrices is symmetric, therefore G' is symmetric.

The annealing constraints are satisfied. Therefore, the equilibrium probability density of chain X' is the Boltzmann distribution (9).

Since X and X' have the same cost-function, their distributions are equal. □

Let the move-sequence $\hat{s} = \langle s_0, \ldots, s_n \rangle$ designate n successively accepted moves. Let the observed cost-difference from state s to s' be denoted $\dot{\delta}(s, s')$

The total observed cost-difference is

$$\dot{\Delta} = \sum_{i=0}^{n-1} \dot{\delta}(s_i, s_{i+1}) \tag{174}$$

Knowing the true-cost at the beginning and end of the move-sequence, $C(s_0)$ and $C(s_n)$, we can compute the total true cost-difference as

$$\Delta = C(s_n) - C(s_0). \tag{175}$$

The accumulated error for the sequence is

$$\epsilon_{\hat{s}} = \dot{\Delta} - \Delta = \sum_{i=0}^{n-1} \epsilon_i. \tag{176}$$

THEOREM 7.2. *The results of* §3.1 *apply to accumulated errors, specifically: If the accumulated error is bounded above by* $\overline{\boldsymbol{\epsilon}}$ *and below by* $\underline{\boldsymbol{\epsilon}}$, *and* $\boldsymbol{\gamma} = \overline{\boldsymbol{\epsilon}} - \underline{\boldsymbol{\epsilon}}$, *then*

$$e^{-\boldsymbol{\gamma}/T} \boldsymbol{\pi}_s(T) \le \dot{\boldsymbol{\pi}}_s(T) \le e^{\boldsymbol{\gamma}/T} \boldsymbol{\pi}_s(T) \tag{177}$$

and

$$e^{-\boldsymbol{\gamma}/T} \bar{C}[\boldsymbol{\pi}(T)] \le \bar{C}[\dot{\boldsymbol{\pi}}(T)] \le e^{\boldsymbol{\gamma}/T} \bar{C}[\boldsymbol{\pi}(T)] \tag{178}$$

Proof. All moves which can generate instantaneous errors have a non-zero probability of being accepted. Therefore, any sequence of moves $\epsilon_{\hat{s}}$ has a non-zero probability $\prod_{i=0}^{n-1} P_{s_i, s_{i+1}}$ of being accepted in its entirety. Its accumulated error, $\epsilon_{\hat{s}}$, contains instantaneous error contributions from each move.

Lemma 7.1 shows us that the equilibrium properties of annealing with "big moves," created by composing `generate` n times, are equal to those obtained by annealing with "little moves", using `generate` once.

Suppose the instantaneous error of the "big move" cost-function was bounded above by $\overline{\epsilon}$ and below by $\underline{\epsilon}$. Of course, we could then use the results of §3.1 to predict the equilibrium properties of the "big move" system.

Note that the instantaneous error of the "big move" cost-function is the accumulated error of the "little moves". Then (12) implies (177), and (18) implies (178). □

The upper error bounds, $\overline{\epsilon}$ and $\overline{\epsilon}_{\hat{s}}$, are infrequently observed at low temperatures: when errors are above the mean, they decrease the probability of acceptance. Some annealing algorithms have errors symmetric about the mean, and there the mean error is zero. Parallel implementations often have these properties [5]. The upper bound can then be set to the lower bound estimate.

8. *Summary*

Calculation errors in simulated annealing, arising from parallelism or sloppy computations, affect the expected cost, the algorithm's speed, and its outcome. We have addressed several open questions about annealing under these inaccurate cost-functions.

Under equilibrium conditions, when the cost-function has range-errors, the expected true-cost is bounded. The controlling value is $\gamma = \overline{\epsilon} - \underline{\epsilon}$. In this case, the expected true-cost, C_γ, ranges from $Ce^{-\gamma/T}$ to $Ce^{+\gamma/T}$, where C is the expected true-cost when running annealing with an accurate cost-function. Therefore, the expected true-cost diverges exponentially from that obtained with an accurate cost-function, as temperature decreases. By controlling the errors so that γ is proportional to T, the expected costs will differ no more than a constant factor.

When the errors exhibit a Gaussian distribution, the controlling value is $\overline{\sigma}^2 - \underline{\sigma}^2$. In this case, the expected true-cost C_γ, ranges from $Ce^{-(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2}$ to $Ce^{+(\overline{\sigma}^2 - \underline{\sigma}^2)/2T^2}$. Again, the expected true-cost diverges exponentially as temperature decreases. Likewise, if $\overline{\sigma}^2 - \underline{\sigma}^2$ is kept proportional to T^2 the equilibrium costs will differ by no more than a constant factor.

To characterize a system with varying temperature, we must examine the rate at which it approaches equilibrium at each temperature. One convenient measure for the rate of convergence is "conductance."

Like the expected true-cost, the conductance of the inaccurate system diverges exponentially from that of the accurate system, as temperature decreases. The conductance of an inaccurate system with fixed bounds varies from $e^{-3\gamma/T}\Phi$ to $e^{+3\gamma/T}\Phi$, where Φ is the conductance of an equivalent accurate system.

The conductance of a system with Gaussian errors, $\ddot{\Phi}$, is related to the accurate system conductance by $e^{(\underline{\sigma}^2 - \overline{\sigma}^2)/T^2}\Phi/2 \leq \ddot{\Phi} \leq e^{(2\overline{\sigma}^2 - \underline{\sigma}^2)/T^2}\Phi$. Again, the conductance diverges exponentially as temperature decreases.

By relating the expected cost and conductance of the inaccurate systems to those of similar accurate systems, we obtained useful results: error constraints and annealing-time adjustments which guarantee convergence to the same quality outcome as normal annealing.

The state spaces we have considered are mathematical approximations to meaningful problems. They provide direction to be modified with experience. Based on this analytic work, simulated annealers should:

¶ Establish whether errors are introduced by approximations in the cost-function. For example, a common cost-function approximates the length of wires in a rectilinear network as half the bounding-box perimeter. This inaccurate cost-function appears in almost every VLSI circuit placement program.

¶ Instrument inaccurate parallel programs to compute and display errors during a run. Note that it is reasonable to measure only accumulated errors, since §7 shows that they are equivalent to instantaneous errors.

¶ Introduce algorithms to keep errors within a constant factor of the temperature. Use §5 and §6 to help estimate the factor.

¶ Increase the number of annealing steps performed at each temperature to compensate for errors. Use §6.3 to help estimate the number of steps.

¶ Weigh the performance benefits of parallelism or sloppy computations against the increased annealing steps to obtain the same quality. Inaccuracies can even degrade the total execution time, when considered in this light.

ACKNOWLEDGEMENTS.

Frederica Darema and Miloš Ercegovac provided helpful criticisms. Members of the simulated annealing mailing list (`anneal-request@cs.ucla.edu`) have reviewed portions of this paper.

REFERENCES

[1] Banerjee, P., Jones, M. H., and Sargent, J. S. Parallel simulated annealing algorithms for cell placement on hypercube multiprocessors. *IEEE Transactions on Parallel and Distributed Systems 1*, 1 (Jan. 1990), 91–106.

[2] Broder, A., and Shamir, E. On the second eigenvalue of random regular graphs. In *Proceedings of the 28th IEEE Symposium on the Foundations of Computer Science* (1987), pp. 286–294.

[3] Casotto, A., Romeo, F., and Sangiovanni-Vincentelli, A. A parallel simulated annealing algorithm for the placement of macro-cells. *IEEE Transactions on Computer-Aided Design CAD-6*, 5 (Sept. 1987), 838–847.

[4] Darema, F., Kirkpatrick, S., and Norton, A. V. Parallel algorithms for chip placement by simulated annealing. *IBM Journal of Research and Development 31*, 3 (May 1987), 391–402.

[5] Durand, M., and White, S. R. Permissible error in parallel simulated annealing. Tech. Rep. RC 15487, IBM T.J. Watson Research Center, Yorktown Heights, New York, 1990.

[6] Gelfand, S. B., and Mitter, S. K. Simulated annealing with noisy or imprecise energy measurements. *Journal of Optimization Theory and Applications 62*, 1 (July 1989), 49–62.

[7] Gelfand, S. B., and Mitter, S. K. Simulated annealing type algorithms for multivariate optimization. *Algorithmica 6*, 3 (1991), 419–436.

[8] Greening, D. R. Parallel simulated annealing techniques. *Physica D: Nonlinear Phenomena 42* (1990), 293–306. (also as IBM Research Report RC 15268).

[9] Greening, D. R., and Darema, F. Rectangular spatial decomposition methods for parallel simulated annealing. In *Proceedings of the International Conference on Supercomputing* (Crete, Greece, June 1989), pp. 295–302.

[10] Grover, L. K. Simulated annealing using approximate calculation. In *Progress in Computer Aided VLSI Design, volume 6*. Ablex Publishing Corp., 1989. (also as Bell Labs Technical Memorandum 52231-860410-01, 1986).

[11] Hajek, B. Cooling schedules for optimal annealing. *Mathematics of Operations Research 13*, 2 (May 1988), 311–329.

[12] Jayaraman, R., and Darema, F. Error tolerance in parallel simulated annealing techniques. In *Proceedings of the International Conference on Computer Design* (1988), IEEE Computer Society Press, pp. 545–548.

[13] Kim, S., Chandy, J. A., Parkes, S., Ramkumar, B., and Banerjee, P. Properplace: A portable parallel algorithm for standard cell placement. In *International Parallel Processing Symposium (IPPS 94)* (Cancun, Mexico, 1994).

[14] Kirkpatrick, S., C. D. Gelatt, J., and Vecchi, M. P. Optimization by simulated annealing. *Science 220*, 4598 (1983), 671–680.

[15] Lam, J. *An Efficient Simulated Annealing Schedule*. PhD thesis, Yale University, New Haven, CT, Dec. 1988.

[16] Mantica, G., and Sloan, A. Chaotic optimization and the construction of fractals: Solution of an inverse problem. *Complex Systems 3* (1989), 37–62.

[17] Otten, R., and van Ginneken, L. *The Annealing Algorithm*. Kluwer Academic Publishers, Boston, 1989.

[18] Romeo, F., and Sangiovanni-Vincentelli, A. A theoretical framework for simulated annealing. *Algorithmica 6*, 3 (1991), 302–345.

[19] Rose, J. S., Snelgrove, W. M., and Vranesic, Z. G. Parallel standard cell placement algorithms with quality equivalent to simulated annealing. *IEEE Transactions on Computer-Aided Design 7*, 3 (Mar. 1988), 387–396.

[20] Sahni, S., and Bhatt, A. The complexity of design automation problems. In *Proceedings of the 17th Design Automation Conference* (Minneapolis, MN, 1980), pp. 402–411.

[21] Sinclair, A., and Jerrum, M. Approximate counting, uniform generation, and rapidly mixing Markov chains. *Information and Computation 82* (1989), 93–133.

[22] Sorkin, G. B. *Theory and Practice of Simulated Annealing on Special Energy Landscapes*. PhD thesis, University of California, Berkeley, Department of Electrical Engineering and Computer Science, Feb. 1992.

[23] Strenski, P. N., and Kirkpatrick, S. Analysis of finite length annealing schedules. *Algorithmica 6*, 3 (1991), 346–366.

[24] Tsitsiklis, J. N. Markov chains with rare transitions and simulated annealing. *Mathematics of Operations Research 14*, 1 (Feb. 1989), 70–90.

[25] White, S. R. Concepts of scale in simulated annealing. In *Proceedings of the International Conference on Computer Design* (1984), IEEE Computer Society Press, pp. 646–651.