TECHNICAL RESEARCH REPORT

Scale Invariance Properties in the Simulated Annealing Algorithm

by Mark Fleischer, Sheldon Jacobson

CSHCN TR 2002-8
(ISR TR 2002-15)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Scale Invariance Properties in the Simulated Annealing Algorithm

M. A. Fleischer
Institute for Systems Research
University of Maryland, College Park
*(*mfleisch@isr.umd.edu*)*

S. H. Jacobson
Department of Mechanical and Industrial Engineering
University of Illinois at Urbana-Champaign
*(*shj@uiuc.edu*)*

Abstract. The Boltzmann distribution used in the steady-state analysis of the simulated annealing algorithm gives rise to several scale invariant properties. Scale invariance is first presented in the context of parallel independent processors and then extended to an abstract form based on lumping states together to form new aggregate states. These lumped or aggregate states possess all of the mathematical characteristics, forms and relationships of states (solutions) in the original problem in both first and second moments. These scale invariance properties therefore permit new ways of relating objective function values, conditional expectation values, stationary probabilities, rates of change of stationary probabilities and conditional variances. Such properties therefore provide potential applications in analysis, statistical inference and optimization. Directions for future research that take advantage of scale invariance are also discussed.

Keywords: scale invariance, self-similarity, simulated annealing, Markov chains, branch and probability bound, nested partition algorithms

Acknowledgements

The authors would like to thank Drs. Alan Weiss, Armand Makowski and Marvin Nakayama, as well as the Associate Editor and an anonymous referee for their insightful and helpful comments that lead to a much improved manuscript.

The first author was supported in part by the Institute for Systems Research at the University of Maryland, College Park, Northrop-Grumman, Inc. and through collaborative participation in the Collaborative Technology Alliance for Communications & Networks sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAD19-01-2-0011. The second author was supported in part by the National Science Foundation (DMI-9907980) and the Air Force Office of Scientific Research (F49620-01-1-0007).

1. Introduction

In recent years, the concept of scale invariance, often described in terms of *self-similarity*, has received much attention. It describes such disparate phenomena as the structure of a snowflake to the behavior of the stock market (Mandelbrot, 1983). These two cases represent the extremes with which scale invariance is apparent. The repeating geometrical patterns at different scales often found in nature are quite compelling even while the mathematics that fully describes this scale invariance is quite arcane. At the other extreme is the scale invariance associated with stochastic and diffusion processes. In this realm, scale invariance departs from the visual and wends its way into the abstract where the probabilistic nature of sample paths becomes the cornerstone of pattern finding.

The notion of pattern finding is key in discovering and utilizing the concept of scale invariance. Thus, the perspective of an analyst is paramount: finding any pattern depends on how you look at things. Indeed, it is often possible to find scale invariance in almost any phenomenon if one stretches definitions sufficiently; whether such scale invariance is useful or even real depends on the vantage point from which such scale invariance is discovered and whether interesting patterns persist in related areas.

This paper identifies a form of scale invariance in the simulated annealing (SA) algorithm. This scale invariance is manifest in several ways. (Fleischer, 1999) shows scale invariance associated with parallel, independent processors. This article describes a scale invariance based on lumping solution states together to form aggregate states.

To fully describe this scale invariance requires a definition of the type of scales used. Section 2 provides this necessary background and illustrates a scale invariance with respect to a system of independent processors thereby providing the basis for comparisons. It also describes the indexing method used in conjunction with aggregate states. These methods are then used in Section 3 to show scale invariance in SA between individual solution states and aggregate solution states. Section 4 describes potential applications in the areas of analysis, statistical inference and in optimization. Finally, Section 5 provides a summary of this article, describes areas of future research, and some concluding remarks.

2. Background and Motivation

The concept of scale invariance in the literature on dynamical systems pertains to phenomena that retains some property at different scales. Demonstrating this property therefore requires a comparison of some phenomenon at different scales and, hence, an appropriate description of the phenomenon

that is being compared and also a description of the type of scaling used. In other words, it must be shown that *something* is similar to something else even though its scale or definition is different. The following equations provide the foundations for these comparisons and are well known (see *e.g.*, (Mitra et al., 1986; Aarts and Korst, 1989)) in the context of SA. These results concern the stationary probabilities associated with state i in a discrete optimization problem.

If the SA algorithm is executed at some fixed temperature t, then the frequency of visits to some particular state $i \in \Omega$ where Ω is the state space is given by the stationary probability distribution

$$\pi_i(t) = \frac{e^{-f_i/t}}{\sum_{j\in\Omega} e^{-f_j/t}}, \tag{1}$$

where f_i is the objective function value associated with state i. (Mitra et al., 1986) shows that the rate of change of the stationary probabilities associated with some state i with respect to the temperature parameter t is

$$\frac{\partial}{\partial t}\pi_i(t) = \frac{\pi_i(t)}{t^2}\left[f_i - \langle f\rangle(t)\right] \tag{2}$$

(Mitra et al., 1986; Aarts and Korst, 1989) where $\langle f\rangle(t)$ is the expected objective function value at temperature t.

Mitra also shows how this rate of change depends on the quantity in brackets (Mitra et al., 1986, p.755-6). For an optimal state i^*, $f_{i^*} < \langle f\rangle(t)$ for $t > 0$, hence the derivative is negative. Consequently, the stationary probabilities of the optima monotonically increase with *decreases* in temperature values.

Equations (1) and (2) form the basis of the scale invariance first described in (Fleischer, 1993), and further developed in this paper. These equations have been used in many of the theoretical results on the convergence of SA and its finite-time performance, yet the scale invariance associated with (1) and (2) has yet to be fully described and exploited.

The next section provides some background on the concept of scale invariance, the definitions that are needed for identifying aggregate states, and describes the scale invariance in SA using several lemmas and theorems.

3. Scale Invariance in SA

The term *scale invariance* is usually employed to describe phenomena and properties that seem to exist or persist on different scales. These scales can be the physical dimensions of an object, time, or some other property that is associated with the phenomenon of interest. Indeed, in *self-similar systems*

as they are sometimes called, it is often impossible to determine the scale of the relevant phenomenon in question merely by observing it. For example, in diffusion processes (such as Wiener Processes), the scale of the physical dimension of sample paths cannot be ascertained simply by viewing the sample paths as they appear and mathematically behave the same at any scale (Ross, 1970). The same is true in fractal geometry: the patterns that emerge from the application of recursive functionals are repeated at all levels of magnification or scale (Mandelbrot, 1983). The system on one scale appears like itself on another scale. Such properties, whether they are the behavior or the attributes of some system that are invariant in terms of scale indicate some form of scale invariance.

The foregoing description of scale invariance is unfortunately rather vague and a more concrete description of scale invariance is desirable. A more appealing way to define scale invariance is in the following abstract terms: if statement A implies B, then scale invariance exists if a transformation applied to A resulting in A', and applied to B resulting in B', implies that A' implies B'. This definition suggests that exploring valid examples of scale invariance requires reasonable definitions of various mathematical elements and a showing of how they relate to those mathematical quantities that reflect scale invariance.

In SA, this scale is not based in terms of physical dimensions or time, but is more abstract and relates to the states of discrete optimization problems. The scale is based on the *level of aggregation* of states be it in terms of the states of several processors or in terms of the states in a single processor system. The invariant properties associated with these aggregate states involves the relationships between their stationary probabilities, objective function values, the rate of change of their stationary probabilities, and the variance of objective function values when the SA algorithm is applied to a discrete optimization problem. Before exploring the scale invariance of aggregated states, however, a brief description of scale invariance in the context of parallel processing is presented.

3.1. Parallel Processing in SA

To motivate the notion of aggregating states in SA, consider a system of parallel processors each running the SA algorithm on a given combinatorial optimization problem. Such a system of processors also exhibits a form of scale invariance (see (Fleischer, 1999)) in that the stationary probabilities and the rate change of the stationary probabilities associated with a *system* state has the same form as the analogous quantities in a single processor system. For a system of p independent processors, each of which is in some state i, the system state can be represented in a product space by $i_1, i_2, \ldots, i_p$ and its stationary probability represented as $\pi_{i_1,i_2,\ldots,i_p}(t)$. In (Fleischer, 1999)

showed that

$$\pi_{i_1,i_2,\ldots,i_p}(t) = \frac{e^{-f_{i_1,i_2,\ldots,i_p}/t}}{\sum_{i_1,i_2,\ldots,i_p} e^{-f_{i_1,i_2,\ldots,i_p}/t}} \tag{3}$$

where $f_{i_1,i_2,\ldots,i_p} = \sum_{m=1}^{p} f_{i_m}$ represents the system's objective function value (the sum of the objective function values associated with each processor). Note that the form of (3) is similar to (1).

The following equations extend this result by showing that the rate of change of the stationary probability with respect to temperature t for p parallel processors has the same form as the rate of change of the stationary probability associated with a single processor. Taking the derivative of (3) with respect to temperature t,

$$\begin{aligned}
\frac{\partial\, \pi_{i_1,i_2,\ldots,i_p}(t)}{\partial t} &= \frac{\partial}{\partial\, t} \frac{e^{-f_{i_1,i_2,\ldots,i_p}/t}}{\sum_{i_1,i_2,\ldots,i_p} e^{-f_{i_1,i_2,\ldots,i_p}/t}} \\
&= \frac{\pi_{i_1,i_2,\ldots,i_p}(t)}{t^2} \left[\sum_{m=1}^{p} f_{i_m} - \left\langle \sum_{m=1}^{p} f_{i_m} \right\rangle (t) \right] \\
&= \frac{\pi_{i_1,i_2,\ldots,i_p}(t)}{t^2} \left[f_{i_1,i_2,\ldots,i_p} - \left\langle f_{i_1,i_2,\ldots,i_p} \right\rangle (t) \right] \qquad (4)
\end{aligned}$$

where (4) is similar to (2). This similarity is apparent simply because of how $f_{i_1,i_2,\ldots,i_p}$ has been defined. Thus, by making meaningful and logical definitions of other elements associated with SA it is possible to extend the similarity apparent from the aggregation of states of multiple processors to the aggregation of states in a discrete optimization problem.

3.2. Aggregating States in SA

To show how lumping states together into an aggregate state exhibits scale invariance, it is necessary to identify these aggregate states. This requires some method for indexing these states so they can be uniquely identified. How this is done is crucial towards demonstrating scale invariance.

In many discrete optimization problems, the index associated with a state is either arbitrary and merely used to distinguish between states (such as in a proof) or used to indicate some other information about the state it represents. In such a case, some specific attribute that not only uniquely describes the particular state but also provides other useful information must be devised. In SA, and in particular, in terms of the stationary probability associated with states, this is often done by using an arbitrary i for a non-optimal state and an i^* for an optimal state. This index is of limited use however for aggregate states as more information than simply distinguishing an optimal state from other states is necessary.

In lumping states together, not only is some method of designating them required, but their objective function values must also be designated. The notion that such aggregate states have an objective function value must therefore be considered. The following definitions provide the indexing conventions used in the next several sections. These conventions denote states, collections of states, stationary probabilities, and objective function values, and are based on both arbitrary denotations and denotations reflecting some ordering of objective function values.

3.3. DEFINITIONS

The following definitions are needed to show how lumped states have self-similar properties as individual states. These basic definitions are used later to define new characteristics of lumped states such as their indices, objective function values and stationary probabilities.

Definitions:

Ω the entire set of states in a discrete optimization problem.

i an arbitrary state in a discrete optimization problem that identifies a particular state.

$\pi_i(t)$ the stationary probability of state i at temperature t.

f_i the objective function value associated with state i.

F a random objective function value produced by the SA algorithm. Thus, its probability distribution is $\Pr\{F = f\} = \sum_{i:f_i=f} e^{-f_i/t} / \sum_{j\in\Omega} e^{-f_j}$.

$\langle f \rangle(t) = \sum_{i\in\Omega} \pi_i(t) f_i$**,** expected objective function value at temperature t.

3.4. AGGREGATING STATES

To show scale invariance based on lumping states together to form aggregate states, these aggregate states must have similar attributes and similar mathematical relationships associated with individual states. It is therefore necessary to assign stationary probabilities and objective function values to these states based on some reasonable and logical criteria and then investigate their relationships to determine whether they are similar to the corresponding relationships associated with individual states.

3.4.1. *Stationary Probabilities*

Let $A = \{i_1, i_2, \ldots, i_m\}$ be some arbitrary set of m states where $A \subset \Omega$. A reasonable approach for defining the stationary probability of A is based

on the frequency of occurrence of any state in the set A. Using indicator variables, define

$$\mathbf{1}_A(t) = \begin{cases} 1 & \text{if the current state } i \in A \text{ at temperature } t \\ 0 & \text{otherwise} \end{cases}.$$

Note that for each state i, $\mathbf{1}_i(t) = 1(0)$ if the current state in SA at temperature t is (not) i. Thus, the stationary probability may be defined using indicator variables and the law of total probability:

$$\begin{aligned} \pi_A(t) &= \mathsf{E}\{\mathbf{1}_A(t)\} = \Pr\{\mathbf{1}_A(t) = 1\} \\ &= \sum_{i \in A} \pi_i(t) = \sum_{l=1}^{m} \pi_{i_l}(t). \end{aligned} \tag{5}$$

These definitions formalize the notion that when the SA algorithm visits any state *in* A, the algorithm visits A itself, *i.e.,* the frequency of visits *to* A, $\pi_A(t)$, is the sum of the frequency of visits to states *in* A.

3.4.2. *Objective Function Values*

Defining objective function values for aggregate states is not quite as simple as in the case involving stationary probabilities. In this case, a visit to a state in A gives A an associated objective function value that may differ from that of a previous visit. Thus, instead of simply counting visits to any state to establish the relative frequency of that state, the attribute of each state—its objective function value—must be taken into account. Note also that the frequency of visits to each element of set A is dependent on the temperature t.

Given that the objective function values associated with set A may vary over the course of an SA experiment, the most reasonable approach for assigning an objective function value is to take the time average of the objective function values obtained *over the course of visits to set* A. This suggests an expression based on conditional expectation. Define the objective function value of lumped node A by

$$\begin{aligned} f_A(t) &= \mathsf{E}\{F \mid \mathbf{1}_A(t) = 1\} \\ &= \sum_{i \in \Omega} f_i \Pr\{F = f_i \mid \mathbf{1}_A(t) = 1\} \\ &= \sum_{i \in \Omega} \frac{f_i \Pr\{F = f_i \wedge \mathbf{1}_A(t) = 1\}}{\Pr\{\mathbf{1}_A(t) = 1\}} \\ &= \frac{\sum_{l=1}^{m} \pi_{i_l}(t) f_{i_l}}{\sum_{l=1}^{m} \pi_{i_l}(t)} = \sum_{i \in A} \frac{\pi_i(t) f_i}{\pi_A(t)}. \end{aligned} \tag{6}$$

Thus, the objective function value of set A is the weighted average or convex combination of the objective function values of the states it contains, *i.e.,*

the expected objective function value of states *in* set A. Note that using the definition in (6), the identity

$$f_\Omega(t) \equiv \langle f \rangle(t)$$

holds for all $t > 0$.

3.4.3. *Consistency*
For scale invariance to exist, the values for $\pi_A(t)$, its rate of change, and $f_A(t)$ must have similar relationships as the corresponding values for individual states. Further, true scale invariance must be *consistent*, that is it must be evident at all scales. This means that aggregations of aggregate states should obey the same relational rules. To illustrate, let A and B be disjoint aggregate states. Then from (6), the objective function value of aggregate state $A \cup B$ is

$$\begin{aligned} f_{A\cup B}(t) &= \sum_{i\in A\cup B} \frac{\pi_i(t) f_i}{\pi_{A\cup B}(t)} \\ &= \sum_{i\in A} \frac{\pi_i(t) f_i}{\pi_{A\cup B}(t)} + \sum_{i\in B} \frac{\pi_i(t) f_i}{\pi_{A\cup B}(t)} \\ &= \frac{\pi_A(t) f_A(t) + \pi_B(t) f_B(t)}{\pi_A(t) + \pi_B(t)} \end{aligned} \tag{7}$$

where (7) is obtained by dividing and multiplying each summation by $\pi_A(t)$ and $\pi_B(t)$, respectively and using the definitions of $f_A(t)$ and $f_B(t)$. Scale invariance is manifest in (7) because this equation of the objective function of unions of disjoint aggregate states has the same formulation in terms of the aggregate states as (6) has in terms of individual states.

3.4.4. *Scale Invariant Relationships*
These scale invariant relationships become further apparent using (5) and (6) to determine the rate change of $\pi_A(t)$ with respect to t,

$$\begin{aligned} \frac{\partial \pi_A(t)}{\partial t} &= \frac{\partial}{\partial t} \sum_{l=1}^{m} \pi_{i_l}(t) \\ &= \sum_{l=1}^{m} \frac{\partial}{\partial t} \pi_{i_l}(t) && (8) \\ &= \sum_{l=1}^{m} \left[\frac{\pi_{i_l}(t)}{t^2} \left(f_{i_l} - \langle f \rangle(t) \right) \right] && (9) \\ &= \sum_{l=1}^{m} \left[\frac{\pi_{i_l}(t) f_{i_l}}{t^2} \right] - \sum_{l=1}^{m} \left[\frac{\pi_{i_l}(t) \langle f \rangle(t)}{t^2} \right] && (10) \end{aligned}$$

where (2) is substituted into (8) for each $\dot{q}$ to yield (9). Noting the definitions in (5) and (6) and substituting these expressions into (10) yields

$$\begin{aligned}\frac{\partial \pi_A(t)}{\partial t} &= \frac{\pi_A(t) f_A(t)}{t^2} - \frac{\pi_A(t)\langle f\rangle(t)}{t^2} \\ &= \frac{\pi_A(t)}{t^2}\left[f_A(t) - \langle f\rangle(t)\right]. \end{aligned} \tag{11}$$

The similarities in (2) and (11) suggest a scale invariant structure due to the parallel relationships between individual states and aggregate states in terms of their stationary probabilities, derivatives with respect to temperature, and objective function values. The following lemma and its corollary expands on these relationships in a property referred to as *objective function complementarity*. This lemma establishes a general relationship between aggregate states, their complement aggregate states, and their objective function values. The following definitions are needed:

Definition. Aggregate states A and $\Omega \setminus A$ are said to be *complementary* states. If $A \subset B \subseteq \Omega$ then sets A and $B \setminus A$ are said to be *complementary relative to set* B or simply are *relative complements* with respect to B.

LEMMA 1. **(Objective Function Complementarity)** *Given any non-empty aggregate states A and $\Omega \setminus A$, for all $t > 0$,*

$$f_A(t) - \langle f\rangle(t) = f_A(t) - f_\Omega(t) = \pi_{\Omega\setminus A}(t)\left[f_A(t) - f_{\Omega\setminus A}(t)\right] \tag{12}$$

Proof. From the definitions of $f_A(t)$ and $\langle f\rangle(t)$,

$$\begin{aligned} f_A(t) - \langle f\rangle(t) &= f_A(t) - \left(\sum_{i\in A}\pi_i(t) f_i + \sum_{i\in\Omega\setminus A}\pi_i(t) f_i\right) \\ &= f_A(t) - \frac{\pi_A(t)\sum_{i\in A}\pi_i(t) f_i}{\pi_A(t)} - \frac{\pi_{\Omega\setminus A}(t)\sum_{i\in\Omega\setminus A}\pi_i(t) f_i}{\pi_{\Omega\setminus A}(t)} \\ &= f_A(t) - \pi_A(t) f_A(t) - \pi_{\Omega\setminus A}(t) f_{\Omega\setminus A}(t) \\ &= \left[1 - \pi_A(t)\right] f_A(t) - \pi_{\Omega\setminus A}(t) f_{\Omega\setminus A}(t) \end{aligned}$$

Since A and $\Omega \setminus A$ are complement sets, then $1 - \pi_A(t) = \pi_{\Omega\setminus A}(t)$ and the result follows. ■

Observe in the lemma that the aggregate state A is obviously contained in Ω. This provides a clue for generalizing the property of Objective Function Complementarity by considering aggregate states that have some complementary relationship with respect to some subset of Ω. This generalization is stated in the following corollary.

Corollary 1 to Lemma 1: *Let sets A and $B \setminus A$ be non-empty relative complements where $A \subset B \subseteq \Omega$. Then for all $t > 0$*

$$f_A(t) - f_B(t) = \frac{\pi_{B\setminus A}(t)}{\pi_B(t)} \left[f_A(t) - f_{B\setminus A}(t) \right].$$

Proof. Applying the lemma and noting that

$$f_\Omega(t) = \pi_B(t) f_B(t) + \pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t)$$

and substituting this into the left-hand side of (12) and expanding the right-hand side yields

$$f_A(t) - \left(\pi_B(t) f_B(t) + \pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t) \right) = \pi_{\Omega\setminus A}(t) f_A(t) - \pi_{\Omega\setminus A}(t) f_{\Omega\setminus A}(t) \quad (13)$$

Adding $\pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t)$ and subtracting $\pi_{\Omega\setminus B}(t) f_B(t)$ to both sides of (13) yields

$$\begin{aligned} f_A(t) - f_B(t) \;=\; & \pi_{\Omega\setminus A}(t) f_A(t) - \pi_{\Omega\setminus A}(t) f_{\Omega\setminus A}(t) \\ & + \pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t) - \pi_{\Omega\setminus B}(t) f_B(t) \end{aligned} \quad (14)$$

Now, since $A \subset B$, then $\Omega\setminus A = (B\setminus A) \cup (\Omega\setminus B)$, a union of two disjoint sets. Thus, from the consistency property described earlier (see Section 3.4.3),

$$f_{\Omega\setminus A}(t) = \frac{\pi_{B\setminus A}(t) f_{B\setminus A}(t) + \pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t)}{\pi_{B\setminus A}(t) + \pi_{\Omega\setminus B}(t)}. \quad (15)$$

Note that $\pi_{B\setminus A}(t) + \pi_{\Omega\setminus B}(t) = \pi_{\Omega\setminus A}(t)$, hence (15) becomes

$$\pi_{\Omega\setminus A}(t) f_{\Omega\setminus A}(t) = \pi_{B\setminus A}(t) f_{B\setminus A}(t) + \pi_{\Omega\setminus B}(t) f_{\Omega\setminus B}(t)$$

and substituting this into (14) along with the expansion of $\pi_{\Omega\setminus A}(t)$ above and simplifying yields

$$f_A(t) - f_B(t) = \pi_{B\setminus A}(t) \left[f_A(t) - f_{B\setminus A}(t) \right] + \pi_{\Omega\setminus B}(t) \left[f_A(t) - f_B(t) \right]. \quad (16)$$

Since $\pi_{\Omega\setminus B}(t) = 1 - \pi_B(t)$ then upon further re-arranging and simplifying of (16) the result follows. ■

Note that when $B = \Omega$ this corollary reduces to the statement in Lemma 1 (Although this indicates that the corollary is a more general statement, it better demonstrates consistency when stated as a corollary).

Corollary 2 to Lemma 1: *Given any non-empty complementary aggregate states A and $\Omega \setminus A$, $f_A(t) < f_{\Omega\setminus A}(t)$ if and only if*

$$f_A(t) \; < \; \langle f \rangle(t) \; < \; f_{\Omega\setminus A}(t).$$

Proof. The implication leading to the statement $f_A(t) < f_{\Omega\setminus A}(t)$ is obvious. Proving the other direction, if $f_A(t) < f_{\Omega\setminus A}(t)$ then from (12) it follows that $f_A(t) < \langle f \rangle(t)$ since both sides of (12) must be negative. Switching the sets A and $\Omega \setminus A$ and applying the Lemma again, (12) becomes

$$f_{\Omega\setminus A}(t) - \langle f \rangle(t) = f_{\Omega\setminus A}(t) - f_\Omega(t) = \pi_A(t) \left[f_{\Omega\setminus A}(t) - f_A(t) \right]$$

with both sides positive and the result follows. ■

Corollary 3 to Lemma 1: *For any non-empty aggregate states A and $\Omega \setminus A$,*

$$\frac{\partial \pi_A(t)}{\partial t} = \frac{-\partial \pi_{\Omega\setminus A}(t)}{\partial t}$$

for all $t > 0$.

Proof. The result follows from the simple application of (11) and Lemma 1. ■

This lemma and its corollaries show that by virtue of scale invariance, a richer and more general set of relationships among objective function values and stationary probabilities can be illuminated. Indeed, the significance of these relationships is amplified by how certain aggregate states mirror the globally optimal state. The following section establishes an important relationship between *optimal aggregate states* and other states.

3.5. Optimal Aggregate States

Define $S_0 \subset \Omega$ to be the set of optimal states. Therefore, any $i^* \in S_0$ has a special characteristic, namely, its objective function value is strictly less than all other objective function values for states not in S_0, *i.e.*,

$$f_{i^*} < f_i \text{ for all } i \notin S_0 \tag{17}$$

Note that in SA this property of the globally optima is supplemented by the fact that $f_{i^*} < f_\Omega(t)$ at *any* temperature $t > 0$ (see the text associated with (2)). Scale invariant relationships should therefore also be exhibited with respect to a globally optimal aggregate state or *supernode*.

Defining a supernode is complicated by the fact that the objective function value associated with an aggregate state is a function of temperature

t. Nonetheless, it is possible to draw analogies from the basic attributes of an optimal state in order to identify reasonable requirements for defining a supernode:

1. A supernode should have the same properties as (17), *i.e.*, an objective function value that is less than that of any *other* state or sets of states not in the supernode;

2. an objective function value always less than or equal to the expected objective function value.

One approach for defining a supernode satisfying these requirements involves ordering and indexing states according to their objective function values. This requires that the states be aggregated based on their objective function values (states with the same objective function value are thus aggregated together). Define sets S_0 through S_{p-1} as follows:

$$\text{for all } i, j \in S_k, f_i = f_j \text{ and}$$

$$f_{S_0} < f_{S_1} < \cdots < f_{S_i} < f_{S_{i+1}} < \cdots < f_{S_{p-1}} \tag{18}$$

for p distinct objective function values, where $f_{S_k} = f_i$ for all $i \in S_k$.

Therefore, S_i is the *set* of states with the i^{th} best (after the optimal) objective function value. A supernode, $\hat{S}_k$, is therefore defined as the aggregation of all sets with the k lowest objective function values,

$$\hat{S}_k \equiv \bigcup_{i=0}^{k} S_i. \tag{19}$$

The stationary probability of $\hat{S}_k$ can then be defined (using (5)) as the sum of the stationary probabilities of the states within the supernode

$$\pi_{\hat{S}_k}(t) = \sum_{i=0}^{k} \pi_{S_i}(t).$$

The objective function value associated with supernode $\hat{S}_k$ can be defined (using (6)) as

$$f_{\hat{S}_k}(t) = \frac{\sum_{i=0}^{k} \pi_{S_i}(t) f_{S_i}(t)}{\sum_{i=0}^{k} \pi_{S_i}(t)}$$

Using these definitions, the supernode $\hat{S}_k$ has all the attributes and properties of the globally optimal state: an index k, a stationary probability $\pi_{\hat{S}_k}(t)$, and an objective function value $f_{\hat{S}_k}(t)$. Moreover, this supernode has analogous relationships to other states in terms of these attributes and properties. Note that from the ordering in (18), $f_{\hat{S}_k}(t)$ has a lower value than the objective

function value of any $S_i \notin \hat{S}_k$. Lemma 2 further shows that like the optimal states, the $\hat{S}_k(t)$ has an objective function value that is also less than the expected objective function value for all $t > 0$.

LEMMA 2. *The objective function value of a supernode as defined in (19) is less than the expected objective function value for all $t > 0$. Thus, given a state space with p objective function values, for all $k < p - 1$ and all temperatures $t > 0,\ f_{\hat{S}_k}(t) < \langle f \rangle(t)$.*

Proof. Let $\hat{S}_k$ be a supernode and $\Omega \setminus \hat{S}_k$ be the corresponding complement aggregate state. From the ordering of sets S_i and the definition of $\hat{S}_k$, every $i \in \hat{S}_k$ is such that $f_i < f_j$ for every $j \in \Omega \setminus \hat{S}_k$. Consequently, any convex combination of objective function values of states in $\hat{S}_k$ is less than any convex combination of objective function values of states in $\Omega \setminus \hat{S}_k$. Therefore, for all $t > 0,\ \ f_{\hat{S}_k}(t) < f_{\Omega \setminus \hat{S}_k}(t)$. From the property of objective function complementarity in Lemma 1, for all $t > 0,\ \ f_{\hat{S}_k}(t) < \langle f \rangle(t)$. ■

This lemma is used to prove the following theorem.

THEOREM 1. *The stationary probability of a supernode $\hat{S}_k$ monotonically increases with decreases in the temperature parameter t (*i.e.*, for all $t, \Delta t > 0,$ with $\Delta t < t,\ \pi_{\hat{S}_k}(t - \Delta t) > \pi_{\hat{S}_k}(t)$).*

Proof. From Lemma 2, $f_{\hat{S}_k}(t) < \langle f \rangle(t)$. Applying (11) where $\hat{S}_k$ is the aggregate node, $\partial \pi_{\hat{S}_k}(t)/\partial t < 0$ for all t, which establishes the result. ■

The scale invariance exhibited by Theorem 1 indicates an interesting relationship among the states. Recall from (2) that states with objective function values greater than $\langle f \rangle(t)$ have stationary probabilities that monotonically decrease as t decreases. (Mitra et al., 1986, p.755-6) observed that non-optimal states i with $f_i < \langle f \rangle(t)$ have stationary probabilities that *increase* as the temperature t is decreased down to some critical value where $f_i = \langle f \rangle(t)$. As $\langle f \rangle(t)$ continues to decrease with decreasing temperature $t,\ f_i > \langle f \rangle(t)$ and the stationary probabilities of these non-optimal states monotonically *decrease*.

This behavior of increasing and decreasing stationary probabilities is also exhibited *within* a supernode as the temperature is decreased. Observe that a supernode $\widehat{S}_k$ contains the non-optimal states in sets $S_1 \ldots S_k$. This means that the stationary probabilities of these non-optimal aggregate states increases and then decreases as the temperature passes through some critical temperature. Thus, the states within the supernode with objective function values less than $f_\Omega(t)$ increase while those with objective function values greater than $f_\Omega(t)$ decrease. Yet from Theorem 1, the stationary probability of a supernode

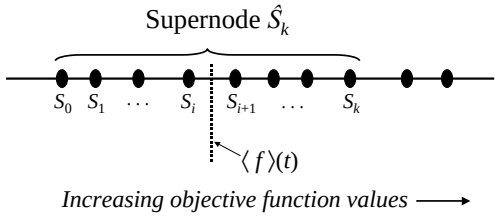

Figure 1. Aggregate States Ranked by Objective Function Value

monotonically increases. This must therefore indicate that the states within the supernode with increasing stationary probabilities more than offsets those states within the supernode with decreasing stationary probabilities.

To see this, consider Figure 1 where aggregate states $S_0 \ldots S_k$ form supernode $\widehat{S}_k$ and where $f_{S_i} < \langle f \rangle(t) < f_{S_{i+1}}$ for states S_i and S_{i+1} both contained within the supernode (thus, $i+1 \leq k$). In this case, for states S_j with $j \leq i$, $\partial \pi_{S_j}(t)/\partial t < 0$, hence these aggregate states have *increasing* stationary probabilities in accordance with (2) and Theorem 1. However, for aggregate states $S_j \subseteq \widehat{S}_k$ with $j > i$, $\pi_{S_j}(t)/\partial t > 0$, hence have monotonically *decreasing* stationary probabilities. But from Theorem 1, the entire supernode has $\partial \pi_{\widehat{S}_k}(t)/\partial t < 0$. From this, and based on (5),

$$\frac{\partial \pi_{\widehat{S}_k}(t)}{\partial t} = \frac{\partial \pi_{\widehat{S}_i}(t)}{\partial t} + \frac{\partial \pi_{\widehat{S}_k \setminus \widehat{S}_i}(t)}{\partial t} < 0,$$

hence

$$\frac{\partial \pi_{\widehat{S}_i}(t)}{\partial t} < \frac{-\partial \pi_{\widehat{S}_k \setminus \widehat{S}_i}(t)}{\partial t}$$

and the magnitude of the rate of increasing probability of $\widehat{S}_i$ is greater than the magnitude of the rate of decreasing probability of state $\widehat{S}_k \setminus \widehat{S}_i$. The relationships between these rates suggests that any aggregate state that contains a supernode will have a monotonically increasing stationary probability. This point motivates the following discussion and theorem.

Theorem 1 is based on the objective function value ordering in (18) and the definition of a supernode *i.e.*, all objective function values of states contained in the supernode are strictly less than objective function values for all states not in the supernode. Although this ordering preserves the properties of the optimal states, it is also somewhat restrictive. The following theorem

generalizes Theorem 1 by only requiring that the aggregate state contain all the optimal states, *i.e.*, also allows it to contain other states.

Define a *partial* supernode $\mathcal{A} = \widehat{S}_k \cup A$ where A is a set of non-optimal states and $\widehat{S}_k$ and A are *separated*, *i.e.*, there exists some intervening states with objective function values greater than the maximum objective function value in $\widehat{S}_k$ and less than the minimum objective function value in A (see Figure 2 for an illustration).

THEOREM 2. *Given a supernode $\hat{S}_m \subset \Omega$ and a* partial *supernode $\mathcal{A} \subset \hat{S}_m$, where the objective function values over the entire state space are non-negative ($f_i \geq 0$ for all $i \in \Omega$), then there exists a temperature t' where $f_\Omega(t') \leq \min\{f_i : i \in \hat{S}_m \setminus \mathcal{A}\}$ such that the stationary probability of the partial supernode monotonically increases,* i.e., *for decreasing $0 < t \leq t'$,* i.e., $\partial \pi_{\mathcal{A}}(t)/\partial t < 0$.

Proof. To clarify the proof, let $B = \hat{S}_m \setminus \mathcal{A}$, the relative complement of $\mathcal{A}$ with respect to supernode $\hat{S}_m$. Since temperature t' is such that $f_\Omega(t') \leq \min\{f_i : i \in B\}$ then it follows that for all $0 < t \leq t',\ f_\Omega(t) \leq f_B(t)$. But

$$f_B(t) - f_\Omega(t) = f_B(t) - f_{\hat{S}_m}(t) + f_{\hat{S}_m}(t) - f_\Omega(t) > 0 \qquad (20)$$

Re-writing (20) using Lemma 1

$$\frac{\pi_{\mathcal{A}}(t)}{\pi_{\hat{S}_m}(t)} \left[f_B(t) - f_{\mathcal{A}}(t)\right] + \pi_{\Omega \setminus \hat{S}_m}(t) \left[f_{\hat{S}_m}(t) - f_{\Omega \setminus \hat{S}_m}(t)\right] > 0 \qquad (21)$$

where $\mathcal{A} = \hat{S}_m \setminus B$.

Now note that the second term in (21) is always negative (from Lemmas 1 and 2), hence the first term must be positive. Consequently, in reversing $f_B(t)$ and $f_{\mathcal{A}}(t)$ in (21) and adding the second term yields

$$\frac{\pi_B(t)}{\pi_{\hat{S}_m}(t)} \left[f_{\mathcal{A}}(t) - f_B(t)\right] + \pi_{\Omega \setminus \hat{S}_m} \left[f_{\hat{S}_m}(t) - f_{\Omega \setminus \hat{S}_m}(t)\right] < 0 \qquad (22)$$

Applying Lemma 1 and this time noting that $B = \hat{S}_m \setminus \mathcal{A}$, the terms in (22) become

$$\left[f_{\mathcal{A}}(t) - f_{\hat{S}_m}(t)\right] + \left[f_{\hat{S}_m}(t) - f_\Omega(t)\right] < 0$$

hence, for all $0 < t \leq t',\ f_{\mathcal{A}}(t) - f_\Omega(t) < 0$ and therefore from (11), $\partial \pi_{\mathcal{A}}(t)/\partial t < 0$ and the stationary probability of the partial supernode monotonically increases. ■

The monotonic behavior of the objective function values for the supernode and the *partial* supernode, as well as the form of the rate of change of the stationary probability demonstrate scale invariance in the SA algorithm.

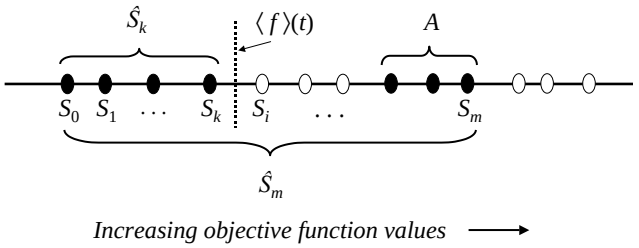

Figure 2. Partial Supernode with arbitrary and separated aggregate state A

3.6. Scale Invariance in Second Moments

The scale invariance described so far involved only the first moments of objective function values—expressions and formulations with f_i. This section presents results on scale invariance involving second moments with terms containing f_i^2. Once again, a basis for comparison is needed. One useful relationship is the derivative of the expected objective function value with respect to temperature:

$$\frac{\partial \langle f \rangle(t)}{\partial t} = \frac{\partial f_\Omega(t)}{\partial t} = \frac{\langle f^2 \rangle(t) - [\langle f \rangle(t)]^2}{t^2} = \frac{\sigma_\Omega^2(t)}{t^2} \tag{23}$$

(Aarts and Korst, 1989, p.20) where the second moment of f is defined by

$$\langle f^2 \rangle(t) \equiv \sum_{i \in \Omega} \pi_i(t) f_i^2.$$

and $\sigma_\Omega^2(t)$ is the variance of objective function values over the entire state space at temperature t.

As noted earlier, scale invariance requires showing that reasonable definitions of certain quantities for aggregated states have similar relationships to other quantities as do the analogous quantities for individual states. To that end and using the same approach and justifications as in (6), define the variance of the objective function of a lumped node A by the conditional variance

$$\sigma_A^2(t) \;=\; \mathsf{E}\{(F - f_A(t))^2 \mid \mathbf{1}_A(t) = 1\}$$

$$
\begin{aligned}
&= \sum_{i\in\Omega}[f_i - f_A(t)]^2 \Pr\{F = f_i \mid \mathbf{1}_A(t) = 1\} \\
&= \sum_{i\in\Omega} \frac{[f_i - f_A(t)]^2 \Pr\{F = f_i \wedge \mathbf{1}_A(t) = 1\}}{\Pr\{\mathbf{1}_A(t) = 1\}} \\
&= \sum_{i\in A} \frac{[f_i^2 - 2f_i f_A(t) + [f_A(t)]^2]\,\pi_i(t)}{\pi_A(t)} \\
&= \frac{\sum_{i\in A}\pi_i(t) f_i^2}{\pi_A(t)} - \frac{\sum_{i\in A}\pi_i(t)[f_A(t)]^2}{\pi_A(t)} \\
&= \langle f^2\rangle_A(t) - [f_A(t)]^2 \qquad (24)
\end{aligned}
$$

where the second moment of the objective function of lumped node A at temperature t is indicated by $\langle f^2\rangle_A(t)$.

Now that a suitable definition for the variance of a lumped node has been defined, scale invariance can be seen in the following similar relationship as in (23):

$$
\begin{aligned}
\frac{\partial f_A(t)}{\partial t} &= \frac{\partial}{\partial t}\left(\frac{\sum_{i\in A}\pi_i(t) f_i}{\pi_A(t)}\right) \\
&= \sum_{i\in A}\frac{\partial}{\partial t}\left(\frac{\pi_i(t) f_i}{\pi_A(t)}\right) \qquad (25)
\end{aligned}
$$

Taking derivatives in (25) leads to

$$
\sum_{i\in A}\left(\frac{\frac{\partial}{\partial t}[\pi_i(t) f_i]\pi_A(t) - \pi_i(t) f_i \frac{\partial}{\partial t}\pi_A(t)}{\pi_A^2(t)}\right). \qquad (26)
$$

Recall that $\frac{\partial \pi_i(t)}{\partial t} = \frac{\pi_i(t)}{t^2}[f_i - f_\Omega(t)]$ and $\frac{\partial \pi_A(t)}{\partial t} = \frac{\pi_A(t)}{t^2}[f_A(t) - f_\Omega(t)]$. Substituting these expressions into (26) and simplifying leads to

$$
\begin{aligned}
\frac{\partial f_A(t)}{\partial t} &= \frac{1}{t^2}\sum_{i\in A}\frac{\pi_i(t) f_i}{\pi_A(t)}[f_i - f_A(t)] \\
&= \frac{1}{t^2}\sum_{i\in A}\frac{\pi_i(t) f_i^2}{\pi_A(t)} - \frac{1}{t^2}\sum_{i\in A}\frac{\pi_i(t) f_i f_A(t)}{\pi_A(t)} \\
&= \frac{1}{t^2}\left[\langle f^2\rangle_A(t) - f_A^2(t)\right] \\
&= \frac{\sigma_A^2}{t^2} \qquad (27)
\end{aligned}
$$

with scale invariance shown in the correspondence between equations (23) and (27)—scale invariance thus holds for second moments. The next section describes how scale invariance in SA can be used in a variety of ways that, in some instances, are unavailable in other optimization schemes.

4. Applications

One of the hallmarks of SA that makes it especially attractive is its generality. Indeed, it is SA's very foundation in thermodynamics that permits it to be used as a *metaheuristic*—an optimization scheme that can be applied to numerous optimization problems. It is also this foundation that gives rise to its scale invariance by virtue of the exponential form of the Boltzmann Distribution. It is this universality that provides clues as to the benefits of SA's scale invariance—benefits that provide numerous avenues for the development of new methodologies that take advantage of SA's scale invariance.

The fundamental utility of scale invariance as a property of some underlying phenomenon is that it permits inferences to easily be made on different scales based on some observed phenomenon or mathematical characteristics associated with some given scale. It can therefore enable or facilitate the *analysis* of a phenomenon at different scales or when information is available only for contrived situations. In addition, if statistically based inferences are possible at one scale, which is certainly the case with SA, scale invariance can enable or facilitate *statistical inference* on other scales. Since SA is used as an *optimization* scheme for many different types of problems, it is not surprising that its scale invariance properties offer some advantages over other optimization schemes. Scale invariance properties in SA can therefore provide tools that facilitate the solution of both theoretical and practical problems in analysis, statistical inference and optimization. This section provides examples that explore and highlight these three potential application areas of SA's scale invariance properties and offers directions for future research in the development of new experimental and computational methodologies.

Section 4.1 describes an example where SAs scale invariance was used in analysis to extend results of a contrived situation to a more general situation. Section 4.2 describes how scale invariance in SA offers new avenues for performing statistical inference with SA by showing how it is possible to define confidence intervals for the value of specified decision variables in the optimal solution without necessarily converging to the optimum. Finally, Section 4.3 extends the ideas in Section 4.2 and describes a type of branch and probability bound algorithm based on scale invariance that may improve the finite-time efficiency of SA.

4.1. Examples of SA's Scale Invariance in Analysis

Scale invariance often provides an attractive angle of attack in the analysis of problems. Indeed, this was *the* motivating factor in creating and exploring SA's scale invariance property in (Fleischer, 1993). (Fleischer, 1993) obtained an analytical result for the contrived situation where the expected objective function at temperature t, $f_\Omega(t)$, was such that $f_{S_0} < f_\Omega(t) < f_{S_1}$— between

the least cost and next-to-least cost objective function values. Such a situation arises in practical circumstances only after SA has almost converged.

The analysis to which scale invariance was applied involved the transitions between the two sets of states, f_{S_0} and f_{S_1} and the distribution of *typical* (this term is a reference to the Asymptotic Equipartition Property from information theory) annealing sequences (Goldie and Pinch, 1991). This contrived situation made the analysis more tractable since it was easier to define "typicality" (see (Fleischer and Jacobson, 1999; Fleischer, 1993)).

Extending this result to the more general circumstance where for some $k,\ f_{S_k} < f_\Omega(t) < f_{S_{k+1}}$, required a new definition of an optimal solution so that the analysis of this more general situation could proceed in an analogous manner as in the contrived situation. By lumping the states with the lowest k objective function values together thereby creating a supernode $\widehat{S}_k$ with all of the same properties as the optimal solution, the subsequent analysis was possible and made much easier due to SA's scale invariance property. Other, as yet unknown, problems in analysis may very well be facilitated by using the scale invariance property of SA.

4.2. Statistical Inference by Simulated Annealing

The scale invariance of second moments and variances described above suggest applications in the realm of statistical inference. This application, referred to as *statistical inference by simulated annealing* (SISA), constitutes a new way to perform statistical inference using SA. Since lumped states can be defined by appropriate constraints on decision variables, new methodologies and new analytical and experimental tools become available to assess quantities associated with the lumped states. One potential value of this is evident:

> *If the partitioning of the solution space is effected by putting constraints on a specified variable, the value of that variable in the optimal solution can be determined with a specified level of confidence.*

This permits values of specified decision variables to be statistically ascertained *without necessarily obtaining a good estimate of the entire ensemble of decision variables*. This possibility constitutes a feature of SA that appears to be unique among metaheuristics.

This section briefly describes the potential avenues for statistical inference based on SA's scale invariance using basic ideas and concepts. The goal here is to introduce some of the basic aspects of SISA and how various test statistics can be engineered to take advantage of SAs scale invariance. Issues regarding sampling, the use of ratio estimates, the convergence of these estimates, and the exact distribution of the relevant random variables and test statistics may therefore become active areas of future research.

It is worth noting that SISA is intimately connected to the mathematics of Markov chains. Recall that SA itself is embodied by a Markov chain. As such, the statistical methodologies associated with Markov chain Monte Carlo (MCMC) techniques are applicable (see (Norris, 1997)). These techniques require that the SA algorithm is executed at a fixed temperature and the various objective function values produced at each iteration recorded. These objective function values provide the raw data needed to calculate various test statistics.

The basic idea behind SISA is to partition the domain space of a problem into mutually exclusive subsets, say A and $\Omega \setminus A$, and to make inferences as to which subset contains the global optimum. This is done by computing estimates of $f_\Omega(t), f_A(t), f_{\Omega\setminus A}(t), \sigma^2_\Omega(t),$ and $\sigma^2_A(t)$, and testing whether $f_A(t) - f_{\Omega\setminus A}(t) = 0$. From the property of Objective Function Complementarity (see Lemma 1 and its corollaries), if $f_A(t) - f_{\Omega\setminus A}(t) < 0 (> 0)$, then it is possible to infer that A $(\Omega \setminus A)$ contains the global optimum.

Consider the following illustrative example: Let $\mathbf{P}$ be a decision problem with an n vector of decision variables $\mathbf{x} = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$, where $x_k \in \{0, 1\}$ for all k. Partition the domain space into two mutually exclusive and exhaustive sets $A = \{\mathbf{x} \in \Omega : x_k = 0\}$ and $\Omega \setminus A = \{\mathbf{x} \in \Omega : x_k = 1\}$ for some specified k. Therefore, sets A and $\Omega \setminus A$ constitute two, complementary lumped states.

SISA is performed to ascertain the value of the specified decision variable x_k. Thus, SA (MCMC) experiments are executed at some fixed temperature t for m iterations and output analysis is done in the standard way (see *e.g.*, (Hobert, 2001) for recent work on MCMC). Each such experiment produces two stochastically generated sequences of objective function values and solution states of length m. For the i^{th} experimental replication of SA at temperature t

$$\{f_{i,a+j}\}_{j=1}^{m} = \{f_{i,a+1}, f_{i,a+2}, \ldots, f_{i,a+m}\}$$

constitutes the sequence of objective function values and

$$\{\mathbf{x}_{i,a+j}\}_{j=1}^{m} = \{\mathbf{x}_{i,a+1}, \mathbf{x}_{i,a+2}, \ldots, \mathbf{x}_{i,a+m}\} \tag{28}$$

constitutes the sequence of corresponding solution states, where a is some index count sufficiently high so as to ensure that the simulation achieves steady-state.

A somewhat less efficient though simpler experimental design to analyze is to run $i = 1, \ldots, r$ *independent* replications of SA with the same initial conditions and different sequences of pseudo-random variables. This produces the following realizations of objective function values (the subscript

a is ignored to simplify the expressions):

$$\begin{array}{ccccc} f_{11}, & \ldots, & f_{1j}, & \ldots, & f_{1m} \\ f_{21}, & \ldots, & f_{2j}, & \ldots, & f_{2m} \\ \vdots & & \vdots & & \vdots \\ f_{r1}, & \ldots, & f_{rj}, & \ldots, & f_{rm} \end{array} \tag{29}$$

Each column of (29) represents a set of i.i.d. realizations of objective function values (see (Law and Kelton, 1991, Ch. 9)) approximately distributed according to the Boltzmann Distribution at temperature t (to avoid cumbersome notation, the temperature parameter and the reference to the j^{th} column are dropped, where it is understood that the realizations of random variables are i.i.d. with the simulation executed at temperature t). Define the following random variables based on the random objective function value F_{ij} generated in the j^{th} iteration of experiment i as

$$F_A = \frac{\sum_{i=1}^{r} F_{ij}\mathbf{1}(i \in A)}{\sum_{i=1}^{r} \mathbf{1}(i \in A)}, \quad F_{\Omega\setminus A} = \frac{\sum_{i=1}^{r} F_{ij}\mathbf{1}(i \in \Omega \setminus A)}{\sum_{i=1}^{r} \mathbf{1}(i \in \Omega \setminus A)}$$

and

$$D = F_A - F_{\Omega\setminus A}.$$

This naturally leads to the following definitions, for a given column j, for the test statistics:

$$\widehat{f}_A = \frac{\sum_{i=1}^{r} f_{ij}\mathbf{1}_A}{\sum_{i=1}^{r} \mathbf{1}_A}, \quad \widehat{f}_{\Omega\setminus A} = \frac{\sum_{i=1}^{r} f_{ij}\mathbf{1}_{\Omega\setminus A}}{\sum_{i=1}^{r} \mathbf{1}_{\Omega\setminus A}}, \quad \widehat{f}_{\Omega} = \frac{\sum_{i=1}^{r} f_{ij}}{r} \tag{30}$$

as estimators of $f_A, f_{\Omega\setminus A},$ and f_Ω, respectively. Note that in the denominators in (30) and for any column j, $\sum_{i=1}^{r} \mathbf{1}_A = r - \sum_{i=1}^{r} \mathbf{1}_{\Omega\setminus A}$. Using the estimates in (30), define

$$\widehat{D} = \widehat{f}_A - \widehat{f}_{\Omega\setminus A}$$

as an estimate of $d = f_A - f_{\Omega\setminus A}$ with three degrees of freedom (it can be shown via the Central Limit Theorem that $D \xrightarrow{d} \mathrm{N}((f_A - f_{\Omega\setminus A}), \sigma_D^2)$. Further analysis will shed light on the exact distribution function for finite-time executions of SA).

Using these estimates, various forms of statistical inference are possible. One approach is to test the following null hypothesis against the alternative hypothesis:

$$\begin{array}{ll} H_0\text{:} & d = 0 \\ H_A\text{:} & d < 0 \text{ or } d > 0 \end{array}$$

Should the test statistics result in the decision to reject H_0 then, depending on whether $D < 0$ or $D > 0$, one could infer that $d < 0$ or $d > 0$ and from the corollary to Lemma 1, that $f_A < (>) f_{\Omega\setminus A}$. One can then conclude that

A or $\Omega \setminus A$ contains the global optimum and hence that $x_i = 0$ (1) in the optimal solution. This also suggests that computing a confidence interval on the value of d can also be useful in deciding whether $x_i = 0$ or 1 *provided it does not contain* 0. The foregoing also suggests how a confidence measure can be assigned to the entire ensemble of decision variables.

One of the implications of SISA is that confidence intervals for *each one* of the n decision variables can be obtained by using the data in (29) and appropriate definitions of lumped states. This provides an approach to optimization where one can assign a confidence level to each decision variable in a putative solution—something not readily available in other optimization schemes. Furthermore, it may be possible to improve on this approach and the efficiency of running SA by modifying the search algorithm to search only the partition deemed to have a high probability (high confidence level) of containing the optimal solution. This idea is the subject of the next section.

4.3. PARTITIONING ALGORITHMS

A practical application of the scale invariance in SA is in the design of a *partition*, or *branch and probability bound* algorithm. See (Shi and Ólafsson, 2000) for a description of a type of partition algorithm based on so-called *nested partitions*. Such a partition algorithm would use SISA on each partition to identify those subsets containing states with certain characteristics (such as the optimal objective function value; see *e.g.,* (Pinter, 1996)). For these types of algorithm, the search is continued using the remaining subsets as a new domain space. This process continues and sequentially shrinks the search space in the hopes that it provides a more efficient search.

Because the decision rule for excluding a subset is probabilistic, such a partition algorithm is also a type of *branch and probability bound* algorithm in which a subset containing some desirable feature is identified with a high probability based on some *prospectiveness criterion* (Zhigljavsky, 1991, p.147). The scale invariance in SA readily lends itself to both partition and branch and probability bound type algorithms and the development of novel prospectiveness criteria.

Whereas the example of SISA above employed partitioning the state space into two mutually exclusive subsets to determine the value of a single, specified decision variable, it is also possible and, perhaps desirable, to partition the domain space into a larger number of mutually exclusive subsets. Once this is done, the subset deemed least likely to contain the optimum is then excluded from further search. This provides a more conservative decision rule for shrinking the domain space and hence lowers the probability of Type I errors—*i.e.,* the probability of excluding portions of the domain space containing the global optimum. A natural question to then ask is how to rank the various partitions.

One way of ranking the partitions is based on the scale invariance exhibited by (11) which is a measure of the rate of change of the stationary probability with respect to temperature. An estimate of (11) for each partition provides more information than the value of $\hat{D} = \widehat{f}_A - \widehat{f}_{\Omega \setminus A}$ since (11) weights the quantity $\widehat{D}$ by the quantity $\widehat{\pi}_A / t^2$.

Using these ideas, a *recursive partitioning algorithm* can be designed in which partitioning is repeated at successive iterations to shrink the state space. Consider a partition of a discrete optimization problem defined by the sets $A_1, A_2, \ldots, A_m$ based on some scheme that may take advantage of some underlying structure (although the partitioning can in fact be completely arbitrary). Recall from Theorem 2 that an aggregate state containing an optimal state has a stationary probability that monotonically increases for sufficiently low temperatures (how low this temperature must be for monotonicity is indicated in the theorem statement). Using a fixed temperature t, an estimate of the derivative of the stationary probability of each partition can be obtained from estimates of each component of (11). Thus, given an SA experiment producing a sequence of states (see (28)), define the estimator

$$\widehat{\frac{\partial \pi_{A_k}}{\partial t}} = \frac{\widehat{\pi}_{A_k}}{t^2} \left[\widehat{f}_{A_k} - \widehat{f}_{\Omega} \right] \tag{31}$$

where the $\widehat{f}_{A_i}$ and $\widehat{f}_{\Omega}$ are defined as in (30) and

$$\widehat{\pi}_{A_k} \equiv \frac{\sum_{j=1}^{m} \mathbf{1}_{A_k}}{m} \tag{32}$$

is an estimator for $\pi_{A_k}(t)$. The estimate in (31) is used as the prospectiveness criterion in the branch and probability bound nomenclature (Zhigljavsky, 1991).

Each partition A_k is thus assigned a value given by (31) which is used in a statistical hypothesis test (similar to the one described in Section 4.2) that tests whether the optimal state $i^* \in A_k$. The partition with the highest value of (31), and hence the lowest *p*-value, is eliminated from the state space. The process is then repeated on the remaining set of states, *i.e.*, a new set of k partitions are established, SA is executed on this smaller domain, and the necessary statistics computed. The procedure is repeated until termination occurs.

Note that the estimate of the rate of change can be obtained *without actually changing the temperature*. In effect, this allows one to use *perturbation analysis* to determine those partitions in which the stationary probability is either increasing or decreasing, hence whether it is likely to contain an optimal state (Fu and Hu, 1992). Instead of running the algorithm at temperature t, obtaining statistics, and rerunning the algorithm at the lower temperature $t - \epsilon$, a single execution can be used to estimate the rate of change of the stationary probability of each partition.

This procedure recursively shrinks the state space. If the probability of a Type I error is sufficiently small, this procedure may provide an efficient way of searching for optimal states. It is worth noting that in some sense, there is an equivalence between annealing by lowering the temperature and annealing by successively reducing the search space.

5. Future Research and Conclusion

This article has explored various scale invariance properties in the SA algorithm. The type of scale invariance examined here was based on scaling the size of the state space by lumping states into aggregate states. Analysis of the SA algorithm with respect to these aggregated states indicates a form of scale invariance because the aggregated states exhibit similar behaviors as individual states. When these aggregated states are assigned an objective function value based on a conditional expectation value, various relationships are preserved between their steady-state probability and the expected objective function value. This produces a number of relational properties such as *Objective Function Complementarity* (Lemma 1 and its corollaries) and monotonicity (Theorems 1 and 2). Scale invariance properties in second moments were also described as well as relationships between the rate change of the expected objective function value with respect to temperature. These results collectively suggest that groups of nodes or sets of states can be treated or viewed as single states. Properties such as convergence in probability to *a state* can therefore be extended to convergence in probability to a *set of states*.

Scale invariance provides a new way of viewing the SA algorithm and provides a solid basis for new research into methodologies, applications, and implementations of SA that take advantage of this property. Potential applications include using these scale invariance properties in analysis. Because scale invariance properties also relate the variance of objective function values to other quantities, statistical inference is possible. SISA provides for the possibility of making inferences about the value of any specified decision variable without necessarily obtaining the optimal solution. Finally, applications of scale invariance in optimization were described.

Applications in optimization are based on the notion of recursive functionals where subsets of nodes constituting a sub-domain are partitioned into mutually exclusive subsets. The subset least likely to contain the optimum, as indicated by a prospectiveness criterion, is then excluded and SA re-executed on the remaining states. This is similar to *nested partition algorithms* and *branch and probability bound* algorithms. The approach of scaling the state space can be especially advantageous for continuous problems where this process can be repeated ad infinitum to induce SA to converge to a small neighborhood of an optimal state. SA can therefore be used recursively to

yield a more efficient use of computer resources, and hence, improve the finite-time performance of SA.

For scale invariance to achieve its true potential, however, these areas of application in analysis, statistical inference and optimization all require further investigation. In SISA, research into the distribution of test statistics for finite-time executions of SA would assist in determining the efficacy of SISA on particular problems. Experimentation may also shed light on how best to implement SISA concepts. Such research would also support efforts to use scale invariance properties in optimization.

The potential use of scale invariance in analysis hints at new discoveries to come. As is often the case, new patterns such as those exhibited by scale invariance, require some time to germinate before their true potential is achieved. Connections between SA and *random Markov fields* (see *e.g.*, (Boykov and Zabih, 1998; Li, 1995)) may provide entirely new ways of analyzing and solving complex problems.

The SA algorithm has been used to solve numerous hard discrete optimization problems. SA has been framed as a "meta-heuristic" owing to its generality. Viewing it strictly as an algorithm, however, imposes a limited perspective on SA and diminishes its significance. Thus, rather than viewing it strictly as an algorithm, SA should be used as a tool for modelling the dynamics and complexity associated with discrete optimization problems. This perspective unleashes the hidden value of SA: the analogies it draws between discrete optimization problems, information theory, and thermodynamics (Fleischer and Jacobson, 1999). The scale invariance properties examined in this paper illustrate only a few of the many potential connections between these areas of inquiry. Other ways to take advantage of the scale invariance described here and further development of the connections to information theory and thermodynamics are possible. Our hope is that this paper will encourage similar discoveries in this remarkable algorithm.

References

Aarts, E. E. and J. Korst: 1989, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons.

Boykov, Y., O. V. and R. Zabih: 1998, 'Markov Random Fields with Efficient Approximations'. Department of Computer Science, Cornell University, Ithaca, New York.

Fleischer, M. A.: 1993, 'Assessing the Performance of the Simulated Annealing Algorithm Using Information Theory'. Doctoral dissertation, Case Western Reserve University, Cleveland, Ohio.

Fleischer, M. A.: 1999, *Metaheuristics: Advances and Trends in Local Search Paradigms for Optimization*, Chapt. 28: Generalized Cybernetic Optimization: Solving Continuous Variable Problems, pp. 403–418. Kluwer Academic Publishers.

Fleischer, M. A. and S. H. Jacobson: 1999, 'Information Theory and the Finite-Time Performance of the Simulated Annealing Algorithm: Experimental Results'. *INFORMS Journal on Computing* **11**(1), 35–43.

Fu, M. and J. Hu: 1992, 'Extensions and Generalizations of Smoothed Perturbation Analysis in a Generalized Semi-Markov Process Framework'. *IEEE Transactions on Automatic Control* **37**(10), 1483–1500.

Goldie, C. and R. Pinch: 1991, *Communication Theory*. New York, N.Y.: Cambridge University Press.

Hobert, J.P., G. J. B. P. J. R.: 2001, 'On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo'. Departments of Statistics, Universities of Florida, Minnesota and Toronto. Unpublished manuscript.

Law, A. M. and W. D. Kelton: 1991, *Simulation Modeling and Analysis, 2nd Ed.* McGraw-Hill Publishing Co., Inc.

Li, S.: 1995, *Markov Random Field Modeling in Computer Vision*. Springer-Verlag.

Mandelbrot, B.: 1983, *The Fractal Geometry of Nature*. New York, NY: W.H. Freeman & Company.

Mitra, D., F. Romeo, and A. Sangiovanni-Vincentelli: 1986, 'Convergence and Finite-Time Behavior of Simulated Annealing'. *Advances in Applied Probability* **18**, 747–771.

Norris, J. R.: 1997, *Markov Chains*. Cambridge, UK: Cambridge University Press.

Pinter, J.: 1996, *Global Optimization in Action*. Norwell, MA.: Kluwer Academic Publishers, Inc.

Ross, S. M.: 1970, *Applied Probability Models with Optimization Applications*. New York, NY: Dover Publications, Inc.

Shi, L. and S. Ólafsson: 2000, 'Nested Partitions Method for Global Optimization'. *Operations Research* **48**(3), 390–407.

Zhigljavsky, A.: 1991, *Theory of Global Random Search*, Mathematics and Its Applications. Norwell, MA.: Kluwer Academic Publishers, Inc.

Disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of Northrop-Grumman, Inc., the Army Research Laboratory, the Air Force Office of Scientific Research, the National Science Foundation or the U.S. Government.