Random Heuristic Search

Michael D. Vose

C. S. Dept., 107 Ayres Hall, The University of Tennessee, Knoxville, TN 37996-1301 USA

Abstract

There is a developing theory of growing power which, at its current stage of development (indeed, for a number of years now), speaks to qualitative and quantitative aspects of search strategies. Although it has been specialized and applied to genetic algorithms, it's implications and applicability are far more general. This paper deals with the broad outlines of the theory, introducing basic principles and results rather than analyzing or specializing to particular algorithms. A few specific examples are included for illustrative purposes, but the theory's basic structure, as opposed to applications, remains the focus.

Key words: Random Heuristic Search, Modeling Evolutionary Algorithms, Degenerate Royal Road Functions.

1 Introduction

Vose [20] introduced a rigorous dynamical system model for the binary representation genetic algorithm with proportional selection, mutation determined by a rate, and one-point crossover, using the simplifying assumption of an infinite population.[1] While some of the extensions, most notably [8], are more recent, the theory's structure and basic results have been in place for a number of years. In its abstract form, the model is sufficiently general to encompass and unify a variety of search methods, from simulated annealing to genetic programming.

The abstract model, referred to as Random Heuristic Search (RHS), is really more of a general paradigm for heuristic search than a formalization of any particular search method. From an analytical perspective, the power of random heuristic search lies partially in its ability to describe a wide range of search methods at various levels of detail, from fine-grained models which capture

[1] This model has been further extended in [7,8,10,21,22,25,27–29,32].

Preprint submitted to Elsevier Preprint *26 July 1999*

complete information, to coarse approximations, which only attempt to track particular statistics. The resulting description is amenable to analysis because description within the framework of random heuristic search corresponds to mathematical formalization.

Beyond description and formalization, the framework of random heuristic search makes available a significant amount of theoretical scaffolding in the form of key concepts and theorems which provide a unified theory. Therefore, once identified as an instance of random heuristic search, a particular search strategy inherits an environment of concepts and results which speaks to the mechanisms that control its dynamics and determine its quantitative and qualitative nature. Moreover, the framework of random heuristic search is economical in that a single operator, referred to as the *heuristic*, encapsulates behavior; its properties completely determine the system (at the level of granularity it was defined), and the dynamical features of RHS are related to its differential and to its fixed points.

Originally designed to describe stochastic search methods (of which deterministic methods are a special case) over finite, discrete domains, RHS has been generalized to the infinite and continuous case. This paper does not concern such generalizations however, dealing principally with finite, time-homogeneous, Markovian search strategies.

The organization of this paper is as follows. Section two introduces random heuristic search as a general search paradigm. Section three briefly describes how a variety of search strategies are naturally instances of random heuristic search. Section four presents basic concepts and theorems which identify quantitative and qualitative properties shared by instances of RHS. Section five introduces hierarchical modeling and explains consistency concepts which can be used to tie different levels in the modeling hierarchy together. Section six illustrates some of the previous material by way of an example. [2]

Before proceeding, a few remarks will be made to define the scope and intent of this article. Whereas it is ludicrous to imply that no one else has worked on stochastic search, this article is not a survey. The main objective is, within the limited space available, to give the broad outlines of the theory of random heuristic search and to introduce the basic principles and results of its abstract framework. While some of this material has appeared elsewhere, this paper brings those scattered results together into a unified theory.

[2] The particular example considered has been previously analyzed by van Nimwegen et. al. [17,18].

2 Random Heuristic Search

This section introduces random heuristic search as an abstract search method. Whereas the emphasis here is on generality, RHS has been instantiated to particular search methods with remarkable success. The interested reader is referred to [25] for a concrete example of this abstract framework as specialized to the Simple Genetic Algorithm.

Before proceeding with the development of RHS, some preliminary remarks regarding notation will be made. Following that, random heuristic search will be introduced gradually through a series of subsections, each supplying additional refinement and detail.

2.1 Notation

Some standard mathematical notation as well as some nonstandard but useful conventions are introduced here.

The set of integers is denoted by $\mathcal{Z}$, and the set of integers modulo c is denoted by $\mathcal{Z}_c$. The symbol $\Re$ denotes the set of real numbers, and for any collection C of real numbers, vectors, or functions, the sub collection of positive members is denoted by C^+. A collection C multiplied by a number α, as in αC, denotes the collection whose members are those of C multiplied by α.

Angle brackets $\langle \cdots \rangle$ denote a tuple which is to be regarded as a column vector. The column vector of all 1 s is denoted by $\mathbf{1}$. The $n \times n$ identity matrix is I_n, and the j th column of the identity matrix is the vector e_j. For vector x, $\text{diag}(x)$ denotes the square diagonal matrix with ii th entry x_i. Indexing of vectors and matrices begins with 0.

Transpose is indicated with superscript T. The standard vector norm is $\|x\| = \sqrt{x^T x}$. Modulus (or absolute value) is denoted by $| \cdot |$. When S is a set, $| S |$ denotes the cardinality of S. More generally, $| \cdot |$ will be used as a function which returns the "cost" of a path or tributary (paths, tributaries, and their associated costs are defined in section 4.3).

Composition of functions f and g is $f \circ g(x) = f(g(x))$. The i th iterate f^i of f is defined by

$$\begin{aligned} f^0(x) &= x \\ f^{i+1}(x) &= f \circ f^i(x) \end{aligned}$$

The notation $O(f)$ denotes a function (with similar domain and codomain as f), call it g, such that pointwise $| g | \leq c \, | f |$ for some constant c. The notation

$o(f)$ represents a function (with similar domain and codomain as f), call it h, such that pointwise $|\,h\,|\;/\;|\,f\,| \to 0$. In the case where f is a vector or matrix, $|\cdot|$ is to be interpreted as a norm.

Curly brackets $\{\cdots\}$ are used as grouping symbols and to specify both sets and multisets. Square brackets $[\cdots]$ are, besides their standard use as specifying a closed interval of real numbers, used to denote an indicator function: if *expr* is an expression which may be true or false, then

$$[\mathit{expr}] = \begin{cases} 1 & \text{if } \mathit{expr} \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

The supremum is the least upper bound, and is denoted by sup. The infimum is the greatest lower bound, and is denoted by inf.

The equivalence of objects x and y is indicated by $x \equiv y$.

2.2 Framework

This material is mostly summarized from the 1994 article by Vose and Wright [28]. The interested reader is referred to [25] for more complete details.

Random heuristic search can be thought of as an initial collection of elements P_0 chosen from some *search space* Ω of cardinality n, together with some *transition rule* τ which from P_i will produce another collection P_{i+1}. In general, τ will be iterated to produce a sequence of collections

$$P_0 \xrightarrow{\tau} P_1 \xrightarrow{\tau} P_2 \xrightarrow{\tau} \ldots$$

The beginning collection P_0 is referred to as the *initial population*, the first population (or *generation*) is P_1, the second generation is P_2, and so on. Populations are multisets.

Not all transition rules are allowed. Obtaining a good representation for populations is a first step towards characterizing admissible τ. Define the *simplex* to be the set

$$\Lambda = \{\langle x_0, ..., x_{n-1}\rangle \;:\; \mathbf{1}^T x = 1,\; x_j \geq 0\}$$

An element p of Λ corresponds to a population according to the following rule for defining its components

$p_j =$ the proportion in the population of the j th element of Ω

For example, suppose Ω is $\{0, 1, 2, 3, 4, 5\}$. Then $n = 6$. The population $\{1, 0, 3, 1, 1, 3, 2, 2, 4, 0\}$ is represented by the vector $p = \langle .2, .3, .2, .2, .1, .0 \rangle$ given Table 1.

coordinate	corresponding element of Ω	percentage of P_0
p_0	0	2/10
p_1	1	3/10
p_2	2	2/10
p_3	3	2/10
p_4	4	1/10
p_5	5	0/10

Table 1. Illustration of population vector.

The cardinality of each generation $P_0, P_1, \ldots$ is a parameter r called the *population size*. Hence the proportional representation given by p unambiguously determines a population once r is known. The vector p is referred to as a *population vector*. The distinction between population and population vector will often be blurred. In particular, τ may be thought of as mapping the current population vector to the next.

To get a feel for the geometry of the representation space, the simplex is displayed in figure 1 for n = 2, 3, and 4. The figures depict Λ (indicated with the thicker lines) as a line segment, a triangle, and a solid tetrahedron. The thinner arrows show the coordinate axes of the ambient space (the projection of the coordinate axes are being viewed in the second figure, which is three dimensional, and in the last figure where the ambient space is four dimensional).

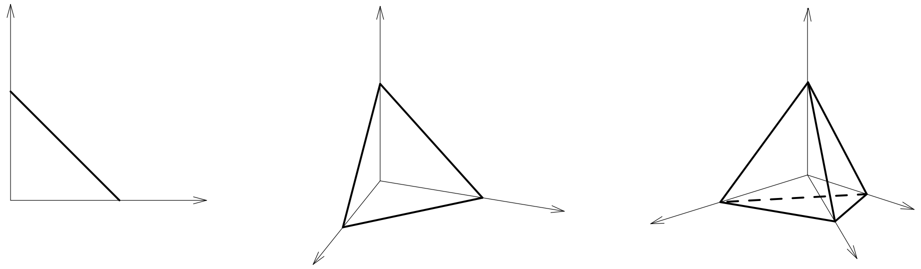

Figure 1. Representation space ($n = 2, 3, 4$).

In general, Λ is a tetrahedron of dimension $n-1$ contained in an ambient space of dimension n. Note that each vertex of Λ corresponds to a unit basis vector of the ambient space; Λ is their convex hull. For example, the vertices of the solid tetrahedron (right most figure) are at the basis vectors $\langle 1, 0, 0, 0 \rangle$, $\langle 0, 1, 0, 0 \rangle$,

$\langle 0,0,1,0 \rangle$, and $\langle 0,0,0,1 \rangle$. Assuming that $\Omega = \{0,1,2,3\}$, they correspond (respectively) to the following populations: r copies of 0, r copies of 1, r copies of 2, and r copies of 3. The center diagram will later be used as a schematic for general Λ, representing it for arbitrary n.

It should be realized that not every point of Λ corresponds to a finite population. In fact, only those rational points with common denominator r correspond to populations of size r. They are the intersection of a rectangular lattice of spacing $1/r$ with Λ,

$$\frac{1}{r}\, X_n^r = \frac{1}{r}\, \{\langle x_0, \ldots, x_{n-1} \rangle \; : \; x_j \in \mathcal{Z},\ x_j \geq 0,\ \mathbf{1}^T x = r\}$$

For example, the points corresponding to $\frac{1}{4}\, X_4^4$ ($n = 4$ and $r = 4$) are the dots in figure 2.

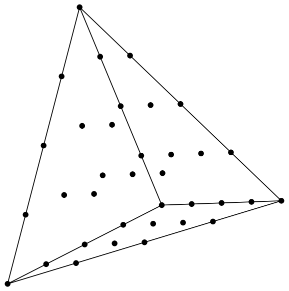

Figure 2. Lattice of populations for $n = 4$ and $r = 4$.

As $r \to \infty$, these rational points become dense in Λ. Since a rational point may represent arbitrarily large populations, a point p of Λ carries little information concerning population size. A natural view is therefore that Λ corresponds to populations of *indeterminate* size. This is but one of several useful interpretations. Another is that Λ corresponds to *sampling distributions* over Ω: since the components of p are nonnegative and sum to 1, p may be viewed as indicating that $i \in \Omega$ is sampled with probability p_i.

In summary, random heuristic search appears to be a *discrete dynamical system* on Λ through the identification of populations with population vectors. That is, there is some transition rule $\tau : \Lambda \to \Lambda$ and what is of interest is the sequence of iterates beginning from some initial population vector p

$$p, \quad \tau(p), \quad \tau^2(p), \quad \ldots$$

This view is incomplete however, because the transitions are in general nondeterministic and not all transition rules are allowed. Next, the stochastic nature of τ will be explained and admissible τ will be characterized.

Because τ is stochastic, the next population vector $\tau(p)$ cannot necessarily be predicted with certainty given the current population vector p. It is most conveniently thought of as resulting from r independent, identically distributed random choices. Let $\mathcal{G} : \Lambda \to \Lambda$ be a *heuristic function* (heuristic for short) which given the current population p produces a vector whose i th component is the probability that the i th element of Ω is chosen (with replacement). That is, $\mathcal{G}(p)$ is that probability vector which specifies the sampling distribution by which the aggregate of r choices forms the next generation. A transition rule τ is admissible if it corresponds to a heuristic function $\mathcal{G}$ in this way. Figure 3 depicts the relationship between p, Λ, Ω, $\mathcal{G}$, and τ through a sequence of generations (the illustration does not correspond literally to any particular case, it depicts how transitions between generations take place in general):

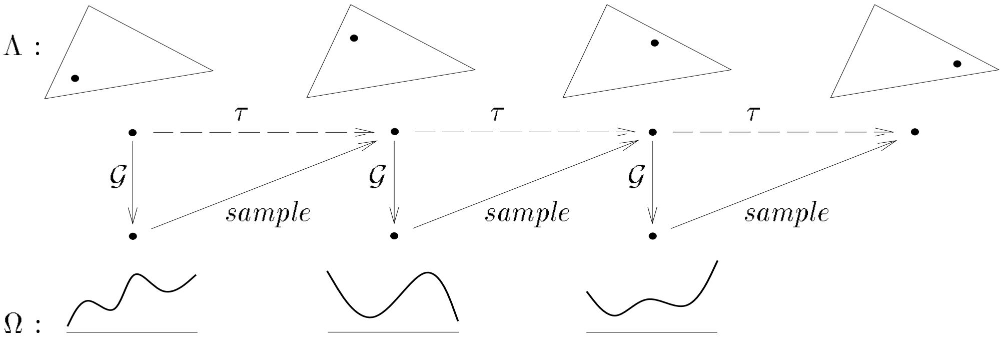

Figure 3. Relationship between p, Λ, Ω, $\mathcal{G}$, and τ.

The triangles along the top row of figure 3 represent Λ, one for each of four generations. Each Λ contains a dot representing a population. These same populations are also represented in the second row with dots; τ maps from one to the next. The transition arrow for τ is dashed to indicate that it is an induced map, computed by following the solid arrows. The third row of dots are images of populations under $\mathcal{G}$. Below each is a curve, suggesting the sampling distribution over Ω which it represents. The line segments in the bottom row represent Ω.

The transition from one generation to the next proceeds a follows. First $\mathcal{G}$ is applied to produce a vector which represents a sampling distribution (curve) over Ω. Next, r independent samples, with replacement, are made from Ω according to this distribution (represented in the diagram by *"sample"*) to produce the next generation.

For example, let $\Omega = \{0, 1, 2, 3\}$ and suppose the heuristic is

$$\mathcal{G}(p) = \langle 0, p_1, 2p_2, 3p_3 \rangle / \sum i\, p_i$$

Let the initial population be $p = \langle .25, .25, .25, .25 \rangle$. Then $\mathcal{G}(p)$ is the sampling distribution $\langle 0, 1/6, 1/3, 1/2 \rangle$, the probability of sampling 0 is 0, of sampling 1 is 1/6, of sampling 2 is 1/3, and of sampling 3 is 1/2. With population size $r = 100$, the transition rule corresponds to making 100 independent samples, with replacement, according to these probabilities.

A plausible next generation is therefore $\tau(p) = \langle 0, .17, .33, .50 \rangle$. Note that the sampling distribution $\mathcal{G}(p)$ used in forming the next generation $\tau(p)$ depends on the current population p. Going one generation further, the new current population is $\tau(p)$ and the sampling distribution for producing the next generation is given by $\mathcal{G}(\tau(p)) \approx \langle 0, .07296, .28326, .64377 \rangle$. It is therefore plausible that the second generation might be $\tau^2(p) = \langle 0, .07, .28, .65 \rangle$.

Note the conceptually dual interpretation of Λ. It serves as both the space of populations and as the space of probability distributions over Ω.

2.4 Dependence On Time

The previous description of random heuristic search is time-homogeneous, that is, neither the population size nor the heuristic depends on time (i.e., on the generation number t).

If, more generally, the population size is a function $r(t)$ of time, or the heuristic is a function $\mathcal{G}(t, \cdot)$ of time, then RHS is said to be *inhomogeneous*. In that case, the heuristic is used to obtain the sampling distribution with which generation $t + 1$ is formed by way of $r(t)$ samples.

In the homogeneous case, random heuristic search is a homogeneous Markov chain over the state space $\frac{1}{r} X_n^r$ since the next state (i.e., population) depends only on the current state, and the dependence is independent of time. In the inhomogeneous case, RHS is still a Markov chain over some subset of Λ, but it is an inhomogeneous chain because the transition from one state to the next, while still a function of the current population, is a function which also depends on t.

3 Examples

This section briefly mentions a few examples to indicate the descriptive power of random heuristic search. The goal is to show the flexibility of RHS as a means to formally describe various search methods.

For some of the methods considered, the heuristic $\mathcal{G}$ will be given explicitly. For others, it will only be indicated how, in principle, $\mathcal{G}$ could be determined. While not exhaustive, or even representative, the examples touched upon below nevertheless demonstrate that a wide variety of search methods are instances of RHS.

3.1 Simulated Annealing

Simulated annealing over a finite domain is an example of inhomogeneous random heuristic search. This is easily seen by identifying the corresponding heuristic.

The population size for simulated annealing is typically $r = 1$, and, given population p (i.e., position p in the search space), the next generation is obtained by the following stochastic procedure:

- Sample q from a neighborhood $N(p)$ of p.
- If $f(q) < f(p)$, where f is the objective function, then the next generation is q.
- Otherwise, the next generation is q with probability

  $$e^{(f(p)-f(q))/T_t}$$

  where T_t is the temperature at generation t.

Since a population contains only a single element of the search space (when $r = 1$), the state space – which is the set of vertices of Λ – is naturally identified with Ω. The corresponding heuristic satisfies

$$\mathcal{G}(t,j)_i = \frac{[i \in N(j)]}{\mid N(j) \mid} \left([f(i) < f(j)] \; + \; [f(i) \geq f(j)]\, e^{(f(j)-f(i))/T_t}\right)$$

for distinct elements i and j of Ω. The case $i = j$ is determined by

$$\mathcal{G}(t,j)_j = 1 - \sum_{i \neq j} \mathcal{G}(t,j)_i$$

3.2 Stochastic Beam Search

Consider a stochastic version of beam search applied to the exploration of a tree. A list $\hat{p}$ of size μ contains nodes and represents the current state. An arbitrary function $f(s, \hat{p})$ – which may, for instance, estimate the likelihood of node s being on a path to the goal, and could, for instance, involve look ahead – determines how "good" node s is with respect to list $\hat{p}$. The list $\hat{p}$ is updated to the next state $\hat{q}$ according to:

- Obtain a sample S of size λ from $\hat{p}$ (sampling of $s \in \hat{p}$ may depend on $f(s, \hat{p})$).
- Let $\hat{p}'$ be the collection of children obtained from expanding elements of S.
- Let $\hat{q}$ be the best μ elements from $\hat{p}'$.

This is summarized by $\hat{q} = \xi(\hat{p})$ where ξ represents the stochastic procedure above.

Since the best μ elements from $\hat{p}'$ are the best μ children of S, the list $\hat{q}$ may be represented by S. Assuming that $\hat{p}$ is similarly represented, the state space for stochastic beam search can be taken to be populations of size λ. Let the representative of $\hat{q}$ – i.e., S – be denoted by q, and let p denote the representative of $\hat{p}$.

While perhaps mysterious, τ determined by $\Pr\{\tau(p) = q\} = \Pr\{\xi(\hat{p}) = \hat{q}\}$ is an instance of RHS representing stochastic beam search. The heuristic $\mathcal{G}$ may be expressed in terms of τ as follows. Since

$$\begin{aligned}\Pr\{i \in \tau(p)\} &= 1 - \Pr\{i \notin \tau(p)\}\\ &= 1 - (1 - G(p)_i)^r\end{aligned}$$

it follows that

$$G(t, p)_i = 1 - (1 - \Pr\{i \in \tau(p) \mid \text{generation } t\})^{1/r}$$

A homogeneous instance of random heuristic search results if μ, λ, f, and the distribution governing the selection of S do not depend on time.

This example, while unsatisfying in the sense that the heuristic was determined only in principle, is important as a prototype for how a search strategy may be shown to be an instance of RHS without explicitly determining the corresponding $\mathcal{G}$.

3.3 Evolutionary Algorithms

The first example below is presented in considerably more detail, though, for reasons of manageability, it is only results rather than underlying reasons that are given (the interested reader is referred to [8,25] for a more general and complete account).

Consider the Simple Genetic Algorithm which moves from one generation to the next as follows:

(1) Obtain two parents by proportional selection.
(2) Mutate (mutation implies change) the parents with rate μ.
(3) Produce the (mutated) parents' child by one-point crossover with rate χ.
(4) Put one child into the next generation.
(5) If the next generation contains less than r members, go to step 1.

Here the search space is the set of all length ℓ strings over the alphabet $\{0, \ldots, c-1\}$. Regarding elements of Ω as c-ary numbers, they are identified with integers in the interval $[0, n-1]$, where $n = c^\ell$. The search space Ω as also naturally identified with the product group

$$\mathcal{Z}_c \times \ldots \times \mathcal{Z}_c$$

The group operation $\oplus$ (i.e., addition modulo c) acts on integers in $[0, n-1]$ via these identifications, and $\otimes$ is used to represent componentwise multiplication modulo c.

Regarding the objective function f as a vector via $f_i = f(i)$, let $F = \mathrm{diag}(f)$. Define the operator $\mathcal{F} : \Lambda \to \Lambda$ by

$$\mathcal{F}(x) = \frac{Fx}{\mathbf{1}^T Fx}$$

Define the matrix M to have i, j th component

$$\frac{(1-\mu)^\ell}{2} \left\{ \eta^{\#i} \left(1 - \chi + \frac{\chi}{\ell - 1} \sum_{k=1}^{\ell-1} \eta^{-\Delta_{i,j,k}} \right) \; + \; \eta^{\#j} \left(1 - \chi + \frac{\chi}{\ell - 1} \sum_{k=1}^{\ell-1} \eta^{\Delta_{i,j,k}} \right) \right\}$$

where $\eta = \mu/((c-1)(1-\mu))$, where $\#x$ denotes the number of nonzero c-ary digits in x, where division by zero at $\mu = 0$ and $\mu = 1$ is to be removed by continuity, and where

$$\Delta_{i,j,k} = \#((c^k - 1) \otimes (c^k - 1) \otimes i) \; - \; \#((c^k - 1) \otimes (c^k - 1) \otimes j)$$

Define permutation matrices σ_j on $\Re^n$ by

$$\sigma_j \langle x_0, \ldots, x_{n-1} \rangle = \langle x_{j \oplus 0}, \ldots, x_{j \oplus (n-1)} \rangle$$

and define the operator $\mathcal{M} : \Lambda \to \Lambda$ by

$$\mathcal{M}(x) = \langle (\sigma_0\, x)^T M \sigma_0\, x, \ldots, (\sigma_{n-1}\, x)^T M \sigma_{n-1}\, x \rangle$$

The Simple Genetic Algorithm's heuristic is

$$\mathcal{G} \quad = \mathcal{M} \circ \mathcal{F}$$

It is well known (and may be verified by direct calculation; simply take $\chi = 0$ above and simplify) that in the case of zero crossover the heuristic has the form

$$\mathcal{G}(x) = \frac{Ax}{\mathbf{1}^T A x}$$

for a suitable matrix A which is positive for nonzero mutation.

As is no doubt clear by contrasting the previous example (stochastic beam search) with this one, establishing that a search strategy is an instance of random heuristic search is, in general, a far easier matter than identifying its heuristic. However, the prototype

$$G(t, p)_i = 1 \; - \; (1 \; - \; \Pr\{i \in \tau(p) \;\mid\; \text{generation } t\})^{1/r}$$

where τ denotes the search strategy's transition rule, implies that many basic types of evolutionary search, including common forms of

- Evolutionary Programming
- Evolutionary Strategies
- Genetic Algorithms
- Genetic Programming

are instances of RHS. The basic requirements are that Ω be finite, and that the transition τ from one generation to the next be Markovian and expressible as the result of r independent, identically distributed random choices (the distribution governing those choices may depend on both the generation number and the current population).

Finiteness is not a serious issue, since limited space and resolution make it a practical reality (for example, in genetic programming it is common to employ a depth bound, and what pass for "real numbers" in Evolutionary Strategies are typically floating point variables of 64 bits or less).

Assuming Markovian transitions, the requirement that τ be expressible as the result of r independent, identically distributed random choices is not a serious issue for many common forms of evolutionary search. For some, like

Genetic Programming for instance, the mechanism producing the next generation is naturally a series of independent identically distributed choices. For others, like $\mu + \lambda$ Evolutionary Strategies, the situation, while considerably less straightforward, may be handled by approximation in the sense that there exists an instance of random heuristic search which approximates, to an arbitrary degree of precision, the actual dynamics.

As illustrated in section 3.2, appropriate choice of representation may help identify a search method as an instance of RHS. In general, Ω need not contain populations rather than strings if for some $r > 1$ there exists a solution $x \in \Lambda$ (which may depend on p and t) to

$$\Pr\{\tau(p) = q\} = r! \prod \frac{x_j^{rq_j}}{(rq_j)!}$$

which holds for all q.

4 Basic Theory

This section is divided into three parts. The first is concerned with the most basic results. The second classifies instances of random heuristic search and introduces fundamental concepts. The third examines transient (i.e., local in time) and asymptotic (i.e., averaged over infinite time) behavior. For simplicity, the exposition will focus on the homogeneous case. For reasons of manageability, it is only results rather than underlying reasons that are given (for related results and more complete details, the interested reader is referred to the citations which appear below).

4.1 First Principles

Given an instance of random heuristic search, perhaps the most fundamental question is: beginning from current population p, what is the probability that the next generation is q? This is the first question to be addressed.

By Stirling's theorem, given $x \in \mathcal{Z}^+$, there exists $0 < \theta < 1$ such that

$$x! = \left(\frac{x}{e}\right)^x \sqrt{2\pi x}\, \exp\left\{\frac{1}{12x + \theta}\right\}$$

Solving this equality for θ defines it as a function of x. The function θ appears in the following theorem (see [23,25,28]).

Theorem 1 *Let p be the current population vector. The probability that population $q \in \frac{1}{r}X_n^r$ is the next population vector is*

$$
\begin{aligned}
& r! \prod \frac{(\mathcal{G}(p)_j)^{rq_j}}{(rq_j)!} \\
= \exp & \left\{ -r \sum q_j \ln \frac{q_j}{\mathcal{G}(p)_j} - \sum \left(\ln \sqrt{2\pi r q_j} + \frac{1}{12rq_j + \theta(rq_j)} \right) + O(\ln r) \right\}
\end{aligned}
$$

where summation is restricted to indices for which $q_j > 0$.

The characterization of random heuristic search as completed in section 2 rests ultimately on sampling Ω, since τ is the induced map in figure 4.

$$\begin{tikzcd}
p \arrow[r, dashed] \arrow[d] & \tau(p) \\
\mathcal{G}(p) \arrow[ur, "\textit{sample } \Omega"'] &
\end{tikzcd}$$

Figure 4. τ as an induced map.

However, since each random vector in the sequence $p, \tau(p), \tau^2(p), \ldots$ depends only on the value of the preceding one, they form a Markov chain with transition matrix

$$
Q_{p,q} = r! \prod \frac{(\mathcal{G}(p)_j)^{rq_j}}{(rq_j)!}
$$

The conceptualization of RHS as given in section 2 may therefore be replaced by an abstraction which makes no reference to sampling Ω at all: from current population p, produce $q = \tau(p)$ with probability $Q_{p,q}$.

As is no doubt clear, the theoretical scaffolding made available by the framework of random heuristic search includes all the machinery of Markov chains. Moreover, any question concerning $\tau(p)$ may be answered in terms of the transition matrix Q, since it defines the stochastic behavior. For example, if the goal of RHS is to produce a population contained in some set S, let ν and Q be the initial population distribution and transition matrix (respectively), except that all entries (rows and columns) corresponding to populations $p \in S$ are omitted. A standard result from Markov chain theory is that the expected number of generations to encounter a member of S is

$$\nu^T(I-Q)^{-1}\mathbf{1}$$

Now that transition probabilities have been determined, it is natural to ask: what is the expected next generation? The answer is given by the following theorem (see [27,28]).

Theorem 2 *Let $\mathcal{E}$ denote the expectation operator.*

$$\mathcal{E}(\tau(p)) = \mathcal{G}(p)$$

Note the conceptually dual interpretation of $\mathcal{G}(p)$. Whereas it previously specified a sampling distribution, it now represents an expected population. Observe that theorem 2 places no conditions on r. It therefore holds independent of population size; $\mathcal{G}$ simultaneously describes the expected next generation for all population sizes.

Theorem 1 in conjunction with theorem 2 provides qualitative information concerning probable next generations. The expression

$$\sum q_j \ln \frac{q_j}{\mathcal{G}(p)_j}$$

is the *discrepancy* of q with respect to $\mathcal{G}(p)$ and is a measure of how far q is from the expected next population $\mathcal{G}(p)$. Discrepancy is nonnegative and is zero only when q is the expected next population. Hence the factor

$$\exp\left\{-r\sum q_j \ln \frac{q_j}{\mathcal{G}(p)_j}\right\}$$

occurring in theorem 1 indicates the probability that q is the next generation decays exponentially, with constant $-r$, as the discrepancy between q and the expected next population increases.

The expression

$$\sum \Big(\ln\sqrt{2\pi r q_j} + \frac{1}{12rq_j + \theta(rq_j)} \Big)$$

measures the *dispersion* of the population vector q. A minimally disperse population q contains r identical population members and corresponds to $q = e_i$ for some i (recall that e_i is the ith column of the identity matrix). The corresponding dispersion is $O(\ln r)$. If $n \geq r$, a maximally disperse population has no duplication (q has r nonzero components which are all $1/r$) and dispersion r. Figure 5 illustrates this for $\ell = 2$, $r = 4$. The size of dots correspond to dispersion; smaller dots have lower dispersion, larger dots have higher dispersion.

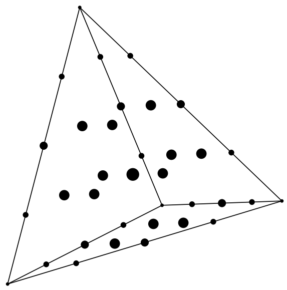

Figure 5. Distribution of dispersion.

The factor

$$\exp\left\{-\sum\Big(\ln\sqrt{2\pi r q_j}+\frac{1}{12rq_j+\theta(rq_j)}\Big)\right\}$$

occurring in theorem 1 indicates the probability that q is the next generation decays exponentially with increasing dispersion. This is related to fluctuations in finite populations induced by sampling; finite populations have a natural tendency under sampling to converge.

The combined effect of the two influences of discrepancy and dispersion is that random heuristic search favors a less disperse population near the expected next generation. In particular, if the current population is near the expected next generation, then the first factor does not contribute a strong bias for change and so the second factor may exert a stabilizing effect on the current population provided it is the less disperse among the alternatives. A further contribution to stasis is provided by the lattice $\frac{1}{r}X_n^r$ of points available to populations for occupation. When $\mathcal{G}(p)$ is nearly the initial population p, the influence of discrepancy favors p as the next generation. The strength of this preference depends upon the distance between p and other lattice points (i.e., alternative populations). This phenomenon is made precise by theorem 1 and the characterization, given in section 2.2, of the finite population state space as $\frac{1}{r}X_n^r$. This same phenomenon was later rediscovered for a particular instance of RHS by van Nimwegen et. al. ([17,18]).

According to theorem 2, the expected next generation from population p is known, but what about the variance? It decreases like $1/r$ (see [22]) and depends upon the distance of $\mathcal{G}(p)$ from a vertex of Λ (see [23,25]).

Theorem 3 *Let $\mathcal{E}$ denote the expectation operator.*

$$\mathcal{E}(\|\tau(p) - \mathcal{G}(p)\|^2) = (1 - \|\mathcal{G}(p)\|^2)/r$$

Theorem 3 points to another influence in support of stasis when the current population is near the expected next generation and in an area of low dispersion. Since $\|\mathcal{G}(p)\|^2 \leq 1$ with equality precisely when $\mathcal{G}(p)$ is at a vertex of Λ, the variance is small in areas of low dispersion. This (i.e., low variance) favors populations near the expected next generation.

A consequence of theorem 3 is that $\tau(p)$ converges in probability to $\mathcal{G}(p)$ as the population size increases. Therefore, τ corresponds to $\mathcal{G}$ in the infinite population case. The following observations can be made (see [7]):

Theorem 4 *The heuristic $\mathcal{G}$ simultaneously answers each of the following questions:*

- *What is the exact sampling distribution describing the formation of the next generation?*
- *What is the expected next generation?*
- *In the limit, as $r \to \infty$, what is the transition function which maps from one generation to the next?*

Moreover, theorems 1, 2, and 3 provide a conceptually simple decomposition of τ into a deterministic *signal* component, and a stochastic *noise* component. Theorem 1 shows, for any r, that $\tau(p)$ is given by a single sample from a multinomial distribution. Associated with the stochastic progression of random heuristic search is the deterministic dynamical system on Λ obtained by iterating $\mathcal{G}$ instead of τ. This is the underlying flow which provides the signal. The message of theorem 2 is that locally (i.e., for a single transition) the expected result of $\tau(p)$ is given by the underlying flow. The message of theorem 3 is that the variance from the flow (i.e., the noise in the sample) is $(1 - \|\mathcal{G}(p)\|^2)/r$.

It is appropriate here to comment on the use of the word “flow” in the previous paragraph. In dynamical systems theory [1], flow is a technical term which does not relate to iterating $\mathcal{G}$, but rather to an extension of that discrete time dynamical system to continuous time by interpolating between successive iterates. While a standard construction might be used to embed a discrete dynamical system in a flow, the domain of the extension differs, in general, from that of the original dynamical system. The use of the word “flow” in this paper is metaphorical, intended to suggest that trajectories (in the infinite population case) are being swept along an evolutionary path under the influence of an underlying current provided by $\mathcal{G}$.

As was noted previously, not every point of Λ corresponds to a finite population; only those rational points with common denominator r correspond to populations of size r. The following theorem makes precise the previous remark that these populations become dense in Λ as $r \to \infty$ (see [23,25]).

Theorem 5 *Let $p, q \in \Lambda$ denote arbitrary population vectors for population size r, and let ξ denote an arbitrary element of Λ. Then*

$$\begin{aligned} \inf_{p \neq q} \|p-q\| &= \sqrt{2}/r \\ \sup_{\xi} \inf_{p} \|\xi - p\| &= O(1/\sqrt{r}) \end{aligned}$$

where the constant (in the "big oh") is independent of the dimension n of Λ.

In the decomposition into signal and noise described above, the signal is *invariant* in the sense that it is independent of the population size ($\mathcal{G}$ does not depend on r). Using the metaphor of the signal exerting a force on a population, the force $\mathcal{G}(p) - p$ acting on p is independent of r (by theorem 1, the influence of r is *external* to $\mathcal{G}$). The lattice spacing within Λ is not, however (theorem 5). When the force is small relative to $\sqrt{2}/r$, discrepancy is minimized by $\tau(p) = p$. In that case, random heuristic search is naturally biased towards treating such populations as if they were fixed points, provided other considerations – like dispersion and noise – do not indicate counter tendencies (theorems 1, 3).

The next result (see [22]) provides a normal approximation to the transition behavior of random heuristic search. In particular, it approximates the fluctuations that occur about a fixed point. Let $q = \mathcal{G}(p)$ and let C be an n by $n-1$ matrix having orthonormal columns perpendicular to $h = \langle \sqrt{q_0}, \ldots, \sqrt{q_{n-1}} \rangle$.

Theorem 6 *For any open subset U of $\mathbf{1}^{\perp}$, the probability that $\tau(p)$ belong to the set $\mathcal{G}(p) + U/\sqrt{r}$ is*

$$(2\pi)^{-(n-1)/2} \int\limits_{C^T \, diag(h)^{-1} U} e^{-y^T y / 2} \, dy \quad + \quad o(1)$$

as r increases.

As will be later explained in some detail (in section 5) the observations made in this section and those that follow apply to random heuristic search *in general*, and speak therefore to both microscopic and macroscopic behavior.

4.2 Fundamental Concepts

Standard terminology from probability theory is used in this section (in the context of Markov chains for example, see [4,13] for the definition of a closed set of states, an absorbing state, etc.).

An instance of random heuristic search is called:

- *Ergodic*, if some some power of the transition matrix Q is positive.
- *Absorbing*, if, in the Markov chain which represents it, every closed set of states contains an absorbing state.
- *Regular*, if whenever C has measure zero, then so does the set $\mathcal{G}^{-1}(C)$.
- *Focused*, if $\mathcal{G}$ is continuously differentiable and p, $\mathcal{G}(p)$, $\mathcal{G}^2(p)$, $\ldots$ converges for every $p \in \Lambda$.
- *Hyperbolic*, if $\mathcal{G}$ is continuously differentiable and its differential $d\mathcal{G}_x$ at x has no eigenvalues of unit modulus when x is a fixed point of $\mathcal{G}$.
- *Normal*, if it is hyperbolic and has a complete Lyapunov function. [3]

If RHS is ergodic, absorbing, regular, focused, hyperbolic, or normal, then both τ and $\mathcal{G}$ are also called ergodic, absorbing, regular, focused, hyperbolic, or normal (respectively). The following observations are, given the previous definitions, standard results from probability theory [4,13].

When RHS is ergodic, every state must be visited infinitely often. Moreover, in that case

$$\pi^T = \lim_{k\to\infty} \nu^T Q^k$$

exists and is independent of the initial population distribution ν. The rows of Q^∞ are each π^T, which is a left eigenvector of Q corresponding to the simple and maximal eigenvalue 1. The pth component of π represents the proportion of time the Markov chain spends in state p (i.e., π is the "steady state distribution"). The steady state distribution π may be extended to a probability measure on Λ as follows:

$$\pi(A) = \sum_{p\in\frac{1}{r}X_n^r} \pi_p\,[p \in A]$$

Here $\pi(A)$ is the probability given to A by the probability measure, and π_p is the pth component of the steady state distribution. Thus for arbitrary $A \subset \Lambda$, the proportion of time that RHS spends in A, averaged over infinitely many generations, is represented by $\pi(A)$.

[3] The paragraph following theorem 7 (below) defines a complete Lyapunov function.

When RHS is absorbing, every initial population has, with probability 1, an evolutionary trajectory which terminates in an absorbing state. Moreover, a steady state distribution

$$\pi^T = \lim_{k\to\infty} \nu^T Q^k$$

exists but is not necessarily independent of the initial population distribution ν. The pth component of π represents the probability that the Markov chain becomes trapped in state p given initial distribution ν. As before, π may be extended to a probability measure on Λ. The extension is denoted by π_ν to make the dependence on ν explicit. Thus for arbitrary $A \subset \Lambda$, the probability that RHS becomes trapped in A, given initial distribution ν, is represented by $\pi_\nu(A)$.

When RHS is regular, if C has positive volume, then so does its expected image (i.e., $\mathcal{G}(C)$). That is, the underlying flow cannot collapse space in any finite number of steps.

When RHS is focused, the trajectory determined by following at each generation what τ is expected to produce will lead to some state ω. By the continuity of $\mathcal{G}$, such points satisfy $\mathcal{G}(\omega) = \omega$ and are therefore called *fixed points*. That is, from every p the underlying flow – or *orbit* – p, $\mathcal{G}(p)$, $\mathcal{G}^2(p)$, $\ldots$ leads to some stagnant location $\omega(p)$ which depends possibly upon p. Moreover, the orbit depends smoothly on p since $\mathcal{G}$ is continuously differentiable.

At a later point the question of speed of convergence will be examined. However, a precise definition of convergence faces several obstacles. The most obvious is that ergodic random heuristic search does not converge, as every state will be visited infinitely often. The naive definition of convergence as time to discover the optimal is generally useless as well. The "no free lunch theorem" [12,31] implies that it is no better, in general, than that achieved by enumeration. The underlying problem here is that the metric of how good RHS is at function optimization is generally worthless to gauge inherent behavior.

Consider, however, that the transition from a population to the next generation is given by $\mathcal{G}$ plus multinomially distributed "noise" (theorem 1). If $\mathcal{G}$ is focused and if the perturbations effected by this noise are not too great, then the initial transient of random heuristic search from initial population p might be characterized by moving towards and spending time in the vicinity of that fixed point $\omega(p)$ to which the underlying flow converges (theorem 8 of the following section partially addresses this phenomenon). This scenario is plausible as the population size grows since the magnitude of the noise decreases with increasing population size (theorems 3, 6).

It is therefore natural to consider the time to convergence of an orbit as an indication of the “settling time” of the initial transient, that is, an approximation of how long it might take for random heuristic search to move from p into the vicinity of $\omega(p)$, assuming the multinomially distributed “noise” is not too great. Even after accepting this concept as an interesting one to pursue, several problems remain. If $\mathcal{G}$ is invertible, then, strictly speaking, the time to convergence of $p, \mathcal{G}(p), \mathcal{G}^2(p), \ldots$ is either zero or infinite depending upon whether p is a fixed point.

The essential point made above is that random heuristic search, under the influence of the underlying dynamical system corresponding to $\mathcal{G}$, may temporarily explore the *vicinity* of $\omega(p)$. This being the case, *approaching* $\omega(p)$ is what matters, and if the concept to be pursued is how the signal component provided by the flow – as opposed to the noise component – relates to this issue, then the most straightforward way to capture the essential idea is to determine, for every δ, the time taken by $p, \mathcal{G}(p), \mathcal{G}^2(p), \ldots$ to come within δ of $\omega(p)$. So as to streamline exposition, the time referred to in the last sentence – which obviously depends on p and δ – will be referred to as “time to convergence”. Note that time to convergence has been defined as a statement regarding the underlying flow of RHS.

Difficulties remain. Perhaps the most obvious is that the time to convergence depends upon the initial population, and, given fixed δ, there is nothing to prevent the existence of a sequence of initial populations along which the time to convergence diverges to infinity. For example, consider any instance of focused random heuristic search such that u and v are distinct attracting fixed points, and let $s(t) = tu + (1 - t)v$. Let t^* be the supremum of $t \in [0, 1]$ such that $\omega(s(t)) = v$. If the time to convergence to v were bounded, say by k, then by the uniform continuity of $\mathcal{G}^k$ (it is continuous and Λ is compact) it follows that $\mathcal{G}^k(s(t^*))$ is mapped within δ of v, and hence converges to v (for suitably small δ) since v is an attractor. But this contradicts that t^* was the supremum because the same continuity argument would imply the flow from $s(t^* + \varepsilon)$ converges to v for some $\varepsilon > 0$. Therefore, given fixed δ, the time to convergence cannot, in general, be uniformly bounded.

However, the possibility remains that time to convergence could be uniformly bounded for “most” initial populations. Let a probability density ϱ be given, and for any set A define the probability that the initial population is contained in A as

$$\int_A \varrho \, d\lambda$$

where λ is surface measure. A natural definition of “most” is a set of probability at least $1 - \varepsilon$ for small ε. It is at this point that the current exposition stresses the generality of the methods employed in [24]. They support surface measure on any manifold invariant under $\mathcal{G}$ – not just Lebesgue measure on Λ – as defining the meaning of “most”.[4]

A position has now been reached where a reasonable definition can be formulated: *Logarithmic convergence* of RHS is a statement about the flow induced by $\mathcal{G}$, and is defined to mean that for every probability density ϱ and every $\varepsilon > 0$, there exists a set A of probability at least $1 - \varepsilon$ such that if the initial population p is in A then the number of generations k required for $\|\mathcal{G}^k(p) - \omega(p)\| < \delta$ is $O(-\log \delta)$, for any $0 < \delta < 1$.

Let $\Im$ be the set of fixed points of $\mathcal{G}$. Note that $\Im$ contains the absorbing states (if there are any) of the Markov chain representing random heuristic search. When RHS is hyperbolic, $\Im$ is finite (see [24]). Moreover, a standard observation from dynamical systems theory [1] is that near a fixed point ω the heuristic $\mathcal{G}$ is locally well approximated by the linear transformation $d\mathcal{G}_\omega$ (regarding ω as the origin) which is a contraction on some linear space $\mathcal{L}$ and an expansion on $\mathcal{L}^\perp$ (for some suitable choice of inner product and corresponding norm; eigenvectors having corresponding eigenvalues within the unit disk are within the contracting linear space, eigenvectors having corresponding eigenvalues exterior to the unit disk are in the orthogonal space).

A discrete form of Lyapunov's theorem is given by the following (see [28]).

Theorem 7 *If $\Im$ is finite and ϕ is a continuous function satisfying*

$$x \neq \mathcal{G}(x) \implies \phi(x) > \phi(\mathcal{G}(x))$$

then iterates of $\mathcal{G}$ converge.

The function ϕ occurring above is called a *Lyapunov function.* The condition on ϕ given in the proposition may be taken as $x \neq \mathcal{G}(x) \implies \phi(x) < \phi(\mathcal{G}(x))$ since it is actually the monotone behavior of ϕ along orbits that matters. When ϕ assigns distinct values to distinct fixed points, it is called a *complete Lyapunov function.*

Since normal heuristics are hyperbolic, $\Im$ is finite, and therefore theorem 7 implies that normal heuristics are focused. Normal heuristics are also open; an arbitrarily small smooth perturbation of a normal heuristic remains normal. Moreover, similar normal heuristics have similar flows (see [25,28]).

When it makes sense to solve the fixed point equation $\mathcal{G}(x) = x$ outside of Λ, as for instance in the case of the simple genetic algorithm where the fixed point equation can be considered over complex space (see [3,7]), then fixed points near but not within Λ may influence the behavior of RHS (see [23,25]).

[4] The statement of results in [24] was not as general as the proof allowed.

The principle involved has been encountered before: By the continuity of the flow, regions in Λ near a fixed point – whether or not the fixed point is within Λ – have a signal component which does not exert strong pressure for change. In such regions, the expected next generation is nearly the initial population (theorem 2). The lattice of points available to populations for occupation contributes to stasis; because populations are constrained to $\frac{1}{r}X_n^r$, discrepancy favors the current population as the next generation in regions where the flow has stalled (theorem 1). The natural preference of random heuristic search for states having low dispersion may have a stabilizing effect on the current population provided it is the less disperse among the alternatives (theorem 1). Moreover, the noise is smaller in such areas of low dispersion (theorem 3).[5]

As pointed out by Rowe [14], fixed points are not the only regions where the phenomenon described above may be manifest. He gives an example where $\mathcal{G}$ is nearly the identity within the unstable manifold of an unstable fixed point. Since the flow has therefore stalled at lattice points near that unstable manifold, it is the *entire manifold* – not just the fixed point – which impacts the behavior of RHS. More generally, what matters is that the flow has stalled, and that may occur in areas not necessarily associated with fixed points (or with unstable/stable manifolds, for that matter).

4.3 Transient And Asymptotic Behavior

The following theorem (see [10]) shows as r increases that, with probability converging to 1, the transient behavior of a population trajectory converges to the flow, and the initial transient occupies an increasing amount of time.

Theorem 8 *Given $k > 0$, $\varepsilon > 0$ and $\gamma < 1$, there exists N such that with probability at least γ and for all $0 \le t \le k$*

$$r > N \implies \|\tau^t(x) - \mathcal{G}^t(x)\| < \varepsilon$$

Theorem 8 indicates that as r increases, a trajectory from p follows a transient trajectory towards a fixed point by approximately following the flow. In particular, if p is near the stable manifold of an unstable fixed point, the initial transient is characterized by moving towards that unstable fixed point.

[5] These mechanisms, as well as those described in section 4.3.1 as inducing punctuated equilibria, have been the subject of public presentations at: The Sixth International Conference on Genetic Algorithms (1995), EvCA'96 sponsored by the Russian Academy of Sciences (1996), IMA Workshop on Evolutionary Algorithms (1996).

The next theorem (see [10,22]) provides a partial answer to the asymptotic question of where RHS is predominantly spending time.

Theorem 9 *If $\mathcal{G}$ is focused and ergodic, then for every $\varepsilon > 0$ and every open set U containing $\Im$, there exists N such that*

$$r > N \implies \pi(U) \; > \; 1 - \varepsilon$$

If $\mathcal{G}$ is absorbing, then $\pi_\nu(\Im) = 1$ for all ν.

Assuming $\mathcal{G}$ is either absorbing or else focused and ergodic, theorem 9 indicates that as r increases, population trajectories predominately spend time near $\Im$ asymptotically. The next theorem (see [24]) partially addresses how quickly orbits approach a fixed point.

Theorem 10 *If $\mathcal{G}$ is regular, focused, and hyperbolic, then $\mathcal{G}$ is logarithmically convergent.*

4.3.1 Punctuated equilibria

Assuming $\mathcal{G}$ is ergodic, regular, focused, and hyperbolic, the view of RHS behavior that emerges is the following (the absorbing, regular, focused, and hyperbolic case is similarly characterized, except that once an absorbing state has been encountered there can be no further change).

As r increases, and then with probability converging to 1, the initial transient of a population trajectory converges to following the flow determined by $\mathcal{G}$, and that transient occupies an increasing time span (theorem 8). Consequently, populations will predominately appear near some fixed point ω of $\mathcal{G}$ (theorem 9), since, by logarithmic convergence, orbits approach fixed points relatively quickly (theorem 10).

This appears in contrast to the fact that ergodic RHS visits every state infinitely often, and is reconciled by *punctuated equilibria* (see [24,27]): Random events will eventually move the system to a population x' contained within or near the stable manifold (with respect to the underlying dynamical system corresponding to $\mathcal{G}$) of a different fixed point ω'. Since random heuristic search is Markovian, the anticipated behavior follows the flow to reach a new temporary stasis in the vicinity of ω'. This cycle of a period of relative stability followed by a sudden change to a new dynamic equilibrium, commonly called *metastability*, is the picture provided by the previous results. The time spent in dynamic equilibrium near a fixed point will be referred to as an epoch.

As has already been explained (see the discussion at the end of section 4.2), metastability is, among other things, a natural consequence of the ergodicity of the Markov Chain, and the interplay between the flow and the lattice available to finite populations for occupation. This mechanism inducing epochal behavior was later rediscovered for a particular instance of RHS in [17,18].

The relationship of logarithmic convergence (theorem 10) to metastability is clarified by reviewing the previous discussion in light of the existence of unstable fixed points and fixed points not within Λ (see [3,23,25]). For focused and hyperbolic RHS, Λ is a finite disjoint union of basins of attraction of fixed points. Although the stable manifolds of unstable fixed points have measure zero, they are interesting because small populations might not be within the basin of attraction of any stable fixed point. Moreover, since the stable manifolds of unstable fixed points have probability zero with respect to every probability density over Λ, it might seem that the logarithmic convergence of RHS does not speak to them.

That is not true, however. Logarithmic convergence is a statement about the underlying flow, and the flow being considered may be taken to be that within the stable manifold $\mathcal{B}$ of an unstable fixed point: the probability density ϱ may be taken over $\mathcal{B}$, the set A may be taken within $\mathcal{B}$, and the integration $\int_A \varrho \, d\lambda$ may be performed with respect to surface measure on $\mathcal{B}$.

It further clarifies matters to realize that whereas the flow within the stable manifold of an unstable fixed point or of a fixed point not within Λ is relatively unrestricted, finite populations are not. As pointed out in 2.2, only elements of a finite lattice of points in Λ are available to finite populations for occupation. Moreover, the lattice has measure zero with respect to every probability density over $\mathcal{B}$, which again suggests that logarithmic convergence of RHS does not speak to those regions of Λ most relevant; i.e., the populations themselves.

However, consider a small neighborhood U of a lattice point. By continuity of the flow, the transient behavior from the lattice point as given by the flow is nearly the transient behavior from any set $A \subset U$ of positive probability with respect to surface measure on *any* stable manifold $\mathcal{B}$ of *any* fixed point. In particular, this continuity together with logarithmic convergence and theorem 8 implies that the flow supports an initial transient of RHS which moves towards the unstable fixed point of lowest dimension[6] having stable manifold near the lattice point (simply consider theorem 10 on the stable manifold $\mathcal{B}$ of lowest dimension which intersects U in some set A of positive probability with respect to surface measure on $\mathcal{B}$); there is a predisposition to visit fixed points in order of increasing dimension. In the context of genetic algorithms, this predisposition has been expressed in terms of visiting fixed points in order of increasing fitness, though in a much less precise and far more heuristic fashon [27]. It was later rediscovered for a particular instance of RHS in [17,18].

[6] The dimension of a fixed point is the dimension of its stable manifold.

The bias of random heuristic search to visit fixed points in order of increasing dimension does not necessarily imply that fixed points of higher dimension (with a larger number of attracting dimensions) are more likely to be visited. Expressed quantitatively in [10], as r decreases the lattice $\frac{1}{r} X_n^r$ of allowable values for population vectors becomes increasingly coarse, as fewer points become available for occupation. Search is conducted in lower dimensional faces of Λ, which constrains the system's ability to follow the signal. The restriction of the heuristic to these low dimensional faces approximates the effective signal, and it is possible that the fixed points of high dimension are not visited, being nowhere close to the low dimensional faces of Λ which can be occupied. Among accessible fixed points, those of higher dimension may be relatively more stable if they have fewer independent unstable directions lying in the low dimensional faces of Λ explored by RHS.

The phenomenon of punctuated equilibria is not confined to the finite population case (though it may be more prevalent there due to the influences peculiar to the finite population case which support its emergence, like, for instance, the ergodicity of the Markov chain and the lattice of points available to populations for occupation). The flow itself – which is followed exactly in the infinite population case – is able to support metastability when there are a number of fixed points of various dimensions. This follows from the continuity referred to above, and is illustrated in figure 6.

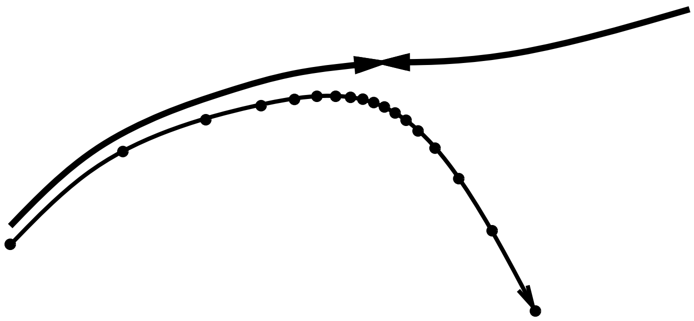

Figure 6. Flow near an unstable fixed point.

The bold curves in figure 6 represent a stable manifold flowing into an unstable fixed point of dimension one. The thin line depicts the flow nearby the stable manifold, and the dots represent an infinite population trajectory. Since the unstable fixed point *is* a fixed point, the flow must slow in its vicinity (by continuity). Thus populations appear to be stable, for a while, as the orbit approaches and leaves the fixed point ... only to approach, perhaps, another unstable fixed point, though of dimension two, whereupon another temporary stasis is experienced, and so on. This scenario of metastability wherein population trajectories may visit fixed points in order of increasing dimension is supported by the continuity of the underlying flow.

4.3.2 Meta-level Chain

Given that random heuristic search is adept at locating regions in the vicinity of fixed points of $\mathcal{G}$ (theorems 8, 9, 10; see also [23,25]), the transition probabilities from one such region to another are significant; random heuristic search could be modeled by a Markov chain over the fixed points. If the transition probabilities from temporary stasis in the vicinity of one fixed point to temporary stasis near another can be determined, then some aspects of the punctuated equilibria could in principle be analyzed.

The goal of constructing a meta-level Markov chain as described in the previous paragraph has been partially achieved in the large population case, insofar as steady state behavior is concerned, subject to the condition that $\mathcal{G}$ is normal and maps Λ to its interior (the interested reader is referred to [22,25] for a more complete account).

Let $\rho = x_0, \ldots, x_k$ be a sequence of points from Λ, referred to as a *path* of length k from x_0 to x_k. Define the *cost* of ρ as

$$\mid \rho \mid = \alpha_{x_0,x_1} \; + \; \cdots \; + \; \alpha_{x_{k-1},x_k}$$

where

$$\alpha_{u,v} = \sum v_j \ln \frac{v_j}{\mathcal{G}(u)_j}$$

Let the stable fixed points of $\mathcal{G}$ in Λ be $\{\omega_0, \ldots, \omega_w\}$ and define

$$\rho_{\omega_i,\omega_j} = \inf \{ \mid \rho \mid : \rho \text{ is a path from } \omega_i \text{ to } \omega_j \}$$

Let $\mathcal{C}$ be a Markov chain defined over $\{1, \ldots, w\}$ with $i \to j$ transition probability (for $i \neq j$) given by

$$\mathcal{C}_{i,j} = \exp\{-r \, \rho_{\omega_i,\omega_j} + o(r)\}$$

As r increases, and then up to uncertainly in the $o(r)$ terms, the desired Markov chain is $\mathcal{C}$ in the sense that the steady state distribution of random heuristic search converges to that of $\mathcal{C}$.

As noted in section 4.2, the Markov chain $\mathcal{C}$ cannot possibly be appropriate for small r because unstable, complex, and stable fixed points outside Λ make no contribution to $\mathcal{C}$. Moreover, as pointed out by Rowe (see the discussion at the end of section 4.2), entire manifolds may have relevance. More generally, what matters is that the flow has stalled, and that may occur in areas not necessarily associated with fixed points or with stable or unstable manifolds. Nevertheless, the form of the transition probabilities above is instructive. The likelihood of a transition from i to j is determined by the minimal cost path from ω_i to ω_j where a path incurs cost to the extent that it is made up of steps which end at a place differing from where $\mathcal{G}$ maps their beginning.

As the population size increases, the steady state distribution of RHS concentrates probability near $\Im$ (theorem 9), which for normal random heuristic search is a finite set. Ergodic RHS will escape the vicinity of one fixed point only to temporarily spend time in the vicinity of another. However, a disproportionate amount of time may be spent near some particular fixed point. Under suitable conditions, random heuristic search will, with probability approaching one, be asymptotically near that fixed point having "largest" basin of attraction; as population size grows, the probability of it spending a nonvanishing proportion of time anywhere else converges to zero.

Define the *fixed point graph* to be the complete directed graph on vertices $\{0, \ldots, w\}$ with edge $i \to j$ (for $i \neq j$) having weight ρ_{ω_i,ω_j}. Define a *tributary* to be a tree containing every vertex such that all edges point towards its root. Let Tree_k be the set of tributaries rooted at k, and for $t \in \text{Tree}_k$ let its cost $|\,t\,|$ be the sum of its edge weights.

A *steady state solution* for an ergodic Markov chain with transition matrix A refers to any solution x of the steady state equation $x^T = x^T A$. The steady state distribution of the Markov chain is obtained simply by dividing x by $\mathbf{1}^T x$. The Markov chain $\mathcal{C}$ has steady state solution

$$x = \langle \sum_{t \in \text{Tree}_0} e^{-r\,(|t|+o(1))}\;, \ldots, \sum_{t \in \text{Tree}_w} e^{-r\,(|t|+o(1))} \rangle$$

Theorem 11 *If there exists a unique minimum cost tributary rooted at some vertex k', then, as r increases, the steady state distribution of $\mathcal{C}$ – and that of ergodic, normal random heuristic search as well – converges to point mass at k'.*

In this case, $\omega_{k'}$ is said to have the "largest" basin of attraction.

5 Hierarchical Models

This section considers the interpretation of random heuristic search as taking place on equivalence classes. One might observe that there is nothing to do, because the search space Ω can simply be taken to be a collection of equivalence classes. While trivially true, the observation is nevertheless important.

Random Heuristic Search is a *general* framework which allows *any* finite set as the search space. Preconceived notions of "microscopic" vs "macroscopic" or "genotype" vs "phenotype" are irrelevant to the scope, power, and application of the paradigm.

At the risk of belaboring what is patently obvious, choosing Ω to be a space of "phenotypes" – which, by the way, is simply a set of equivalence classes – brings the full force of the theory of RHS to bear at what one might call the "macroscopic" level.

If, however, an instance of random heuristic search is already defined, the interesting question is whether that instance is compatible with a given equivalence relation. Put another way: given a microscopic definition of RHS, is a macroscopic model compatible with it?

The issue of compatibility may perhaps best be illustrated by discussing an abstract example. Let τ be an instance of RHS over search space Ω. Let $\equiv$ be an equivalence relation on Ω, and for $p \in \Omega$ let $[p]$ denote the equivalence class containing p.[7] Suppose further that $\tilde{\tau}$ is an instance of RHS having the equivalence classes as its search space.

Given $p \in \Omega$, one may be interested in some aspect of the sequence

$$p,\ \tau(p),\ \tau(\tau(p)),\ \ldots$$

Suppose the investigation is to be carried out by considering $\tilde{\tau}$ instead, i.e., by focusing attention solely on

$$[p],\ \tilde{\tau}([p]),\ \tilde{\tau}(\tilde{\tau}([p])),\ \ldots$$

If, for general p, a conclusion based on the behavior of $[p], \tilde{\tau}([p]), \tilde{\tau}(\tilde{\tau}([p])), \ldots$ applies to $p, \tau(p), \tau(\tau(p)), \ldots$ then it must also apply – without any change whatsoever – to $q, \tau(q), \tau(\tau(q)), \ldots$ whenever $[q] = [p]$. In other words, valid conclusions cannot distinguish between members of an equivalence class. The following question therefore arises: does the aspect of interest depend upon the initial population p in any way? If so, then it had better be the case that, with respect to the aspect of interest, members of an equivalence class are indistinguishable.

Note that the situation described above depends on τ (since $p, \tau(p), \tau(\tau(p)), \ldots$ is the object of interest) and upon the equivalence relation (since valid conclusions cannot distinguish between equivalent members) but is independent

[7] Previous usage of $[expr]$ to denote an indicator function will be maintained; the type of the argument to $[\cdot]$ will disambiguate possible meanings.

of $\tilde{\tau}$ in the sense that, however it may be defined, only properties shared by members of an equivalence class can be deduced.

Of course, $\tilde{\tau}$ needs to be defined such that properties of $[p], \tilde{\tau}([p]), \tilde{\tau}(\tilde{\tau}([p])), \ldots$ are relevant. Towards that end, one may desire a relationship between τ and $\tilde{\tau}$ similar to

$$[\tau(p)] = \tilde{\tau}([p])$$

In that case, a hierarchical relationship exists between them in that the following diagram commutes

$$\begin{tikzcd}
p \arrow[r] \arrow[d] & \tau(p) \arrow[d] \\
{[p]} \arrow[r] & \tilde{\tau}([p])
\end{tikzcd}$$

Thus the trajectory of an equivalence class under $\tilde{\tau}$ is the equivalence class of a trajectory under τ. Without a relationship of this kind, there is no guarantee that the equivalence class of a future generation, namely $[\tau^k(p)]$, bears any relationship to that predicted by $\tilde{\tau}$, namely $\tilde{\tau}^k([p])$.

In other words, if the goal of introducing $\tilde{\tau}$ is to provide a coarse-grained model of τ over a simplified search space of reduced complexity in which many states have been collapsed or aggregated together, then the commutativity – in some sense – of the diagram is required in order that the model reflect the search behavior of τ. Otherwise, without one reflecting the other, there is no guarantee that the "model" $\tilde{\tau}$ has any relevance to τ.

The general theory of random heuristic search, as well as the remarks above, may be brought to bear on the model $\tilde{\tau}$ since it is an instance of RHS. In particular, an equivalence relation $\equiv'$ might be defined over its search space and a coarse-grained model τ' of $\tilde{\tau}$ might be introduced, leading to a commutative diagram of the sort

$$\begin{tikzcd}
p \arrow[r] \arrow[d] & \tau(p) \arrow[d] \\
{[p]} \arrow[r] \arrow[d] & \tilde{\tau}([p]) \arrow[d] \\
{[[p]]'} \arrow[r] & \tau'([[p]]')
\end{tikzcd}$$

where $[[p]]'$ indicates the equivalence class of $[p]$ with respect to $\equiv'$. In this manner a hierarchy of models of varying granularity, form fine-grained models which capture complete information, to coarse approximations, which only attempt to track particular statistics, may be constructed.

The first part of this section concerns the issues discussed above. Its main results are conditions under which random heuristic search can be viewed as taking place on equivalence classes in a hierarchical manner. That is, it is concerned with consistency and commutativity.

The second part of this section briefly considers the suitability of random heuristic search over equivalence classes as a framework for approximate models in which no analogue of the hierarchical relationship $[\tau(p)] = \tilde{\tau}([p])$ necessarily holds.

To put this and the following sections in perspective, a few observations can be made. First, the idea of moving to equivalence classes for the purpose of simplifying or analyzing behavior is hardly new. In mathematics, for example, the use of quotient spaces dates back nearly a century (see [2] for a general discussion of quotient spaces corresponding to a function f and its equivalence relation $E(f)$).

As to the application of equivalence classes to genetic algorithms, Holland [6] was perhaps the first. His schemata result from the equivalence relation $E(f)$ of suitably chosen f related to patterns occurring in chromosomes. Choosing f to be fitness, or related to fitness, results in examples $E(f)$ of a different character. Rabinovich and Wigderson have analyzed GA dynamics in terms of the corresponding quotient, i.e., in terms of fitness distributions [11]. Whereas ad hoc statistics of fitness distributions (online performance, offline performance, etc.) have historically been used as indicators of GA performance, classical statistics (mean, variance, skewness, excess) have been used for the purpose of modeling evolutionary trajectories [16].

Therefore, the point here is not to introduce the field of genetic algorithms to the concept of equivalence classes – as noted above that has been done before, the most notable examples being schema, and fitness distributions. The point is rather to give a coherent general account of quotients as they relate to the abstract framework of random heuristic search, and to explicate relevant consequences, interpretations, and interrelationships of a given instance of random heuristic search to natural interpretations of it in a quotient. For reasons of space, theorems in the following sections are simply stated. The interested reader is referred to [25,26] for details.

5.1 Equivalence

Because Ω can naturally be regarded as a subset of Λ through the correspondence

$$i \in \Omega \quad \longleftrightarrow \quad e_i \in \Lambda$$

an equivalence relation on Ω may be regarded as applying to the unit basis vectors of $\Re^n$ (i.e., the vertices of Λ) by

$$e_i \equiv e_j \quad \Longleftrightarrow \quad i \equiv j$$

This relation on the vertices of Λ is extended to all $x, y \in \Lambda$ by

$$x \equiv y \quad \Longleftrightarrow \quad \forall\, t\,.\sum [i \equiv t]\, x_i \;=\; \sum [i \equiv t]\, y_i$$

The practice of using $\equiv$ for an equivalence relation on both Ω and Λ, as above, will be continued, since context makes the meaning clear. Moreover, $\equiv$ can without modification be regarded as an equivalence relation on all of $\Re^n$, since the definition above applies to any $x, y \in \Re^n$.

Let $\Lambda/\!\equiv$ denote the set of equivalence classes of $\equiv$ in Λ, and let $\Omega/\!\equiv$ denote the set of equivalence classes of $\equiv$ in Ω. The notation $[a]$ will be used to denote the equivalence class of a; thus $[a] \in \Omega/\!\equiv$ when $a \in \Omega$, and $[a] \in \Lambda/\!\equiv$ when $a \in \Lambda$.

Equivalence can be expressed in terms of the linear operator

$$\Xi \,:\, \Re^{|\Omega|} \longrightarrow \Re^{|\Omega/\equiv|}$$

having matrix

$$\Xi_{[i],j} = [i \equiv j]$$

where the rows are indexed by elements of $\Omega/\!\equiv$ and the columns are indexed by Ω. Note that $\Xi x = \Xi y$ if and only if for all i

$$\sum [i \equiv j]\, x_j = \sum [i \equiv j]\, y_j$$

Therefore, elements $x, y \in \Lambda$ are equivalent precisely when $\Xi x = \Xi y$. Since equivalence corresponds to having the same image under Ξ, the equivalence classes must be preimages under Ξ,

$$\Lambda/\!\equiv\, = \{\, \Lambda\ \cap \Xi^{-1} x \ : \ x \in \Xi\Lambda\}$$

Theorem 12 *Elements of* $\Lambda/\!\equiv$ *are convex, compact sets.*

Let $T \subset \Lambda$ be a collection of equivalence class representatives. That is, let T be minimal with respect to containment such that

$$\Lambda \quad \subset \quad \bigcup_{t \in T} [t]$$

Note that T represents $\Lambda/\!\equiv$ through the correspondence

$$t \longleftrightarrow [t]$$

Given any collection T of equivalence class representatives, the map

$$\Xi \ : T \ \longrightarrow \ \Xi\Lambda$$

is an isomorphism. Hence $\Lambda/\!\equiv$, which is represented by T, may be identified with $\Xi\Lambda$. Note further that

$$(\mathbf{1}^T \Xi)_j \; = \sum_{[i] \in \Omega/\equiv} [i \equiv j] \; = \; 1$$

Hence the image of Λ under Ξ consists of non-negative vectors of dimension $\mid \Omega/\!\equiv \mid$ which sum to 1. It follows that $\Xi\Lambda$, which has been identified with $\Lambda/\!\equiv$, represents the state space for random heuristic search over the search space $\Omega/\!\equiv$. The set $\Xi\Lambda$ is called the *quotient representation space*, Ξ is called the *quotient map*, and $\Omega/\!\equiv$ is called the *quotient search space.*

Now that basic objects (the quotient map, the quotient search space, and the quotient representation space) have been introduced and the correspondences

$$T \ \longleftrightarrow \ \Lambda/\!\equiv \ \longleftrightarrow \ \Xi\Lambda$$

have been established, the question of how a map on Λ may act naturally on the quotient space will be considered.

Given a stochastic function h on Λ, define stochastic $h_T : T \longrightarrow T$ in accordance with

$$\Pr\{h_T(t) = t'\} = \Pr\{h(t) \equiv t'\}$$

If h is deterministic, the definition reduces to

$$h_T(t) = t' \in T \;\text{ such that }\; h(t) \in [t']$$

The map h_T is equivalent to a map $\tilde{h}$ on the quotient space by $\tilde{h}(\Xi t) = \Xi h_T(t)$ for $t \in T$. As expected, $\tilde{h}$ depends on the choice T of representatives. That is, there is no reason to expect any natural relationship exists between h and $\tilde{h}$. Whereas the hierarchical relationship

$$[h(t)] = \tilde{h}([t])$$

holds in the deterministic case – by definition – for $t \in T$, there is no guarantee it holds for elements not in T. When h is nondeterministic, the relationship may fail altogether. However, a strict interpretation of the hierarchical relationship in the context of stochastic functions is neither necessary nor desirable. Given functions h and g, to say "as stochastic functions, $h = g$" is to indicate that

$$\Pr\{h(x) = y\} = \Pr\{g(x) = y\}$$

for all x and y. It is true in the nondeterministic case that, as stochastic functions,

$$[h(t)] = \tilde{h}([t])$$

provided $t \in T$. As in the deterministic case, there is no guarantee this relationship holds for elements not in T.

The stochastic function h is said to be *compatible with* $\equiv$ if

$$x \equiv y \implies \forall t \in T \,.\, \Pr\{h(x) \in [t]\} = \Pr\{h(y) \in [t]\}$$

When h is deterministic, this reduces to $x \equiv y \Longrightarrow h(x) \equiv h(y)$.

Theorem 13 *In order, for every $t \in T$, that the distribution of $\tilde{h}(\Xi t)$ be independent of the collection T of equivalence class representatives, it is necessary and sufficient that h is compatible with $\equiv$. When h is deterministic, $\tilde{h}$ is completely determined by the following commutative diagram.*

$$\begin{tikzcd}
x \arrow[r] \arrow[d, "\Xi"'] & h(x) \arrow[d, "\Xi"] \\
\Xi x \arrow[r, "\tilde{h}"] & \Xi h(x)
\end{tikzcd}$$

Given h compatible with $\equiv$, the function $\tilde{h}$ is referred to as the *quotient* of h (with respect to $\equiv$).[8] To simplify exposition, $\Omega/\!\equiv$ will be denoted by $\tilde{\Omega}$, and the image of $x \in \Lambda$ under the quotient map will be denoted by $\tilde{x}$.

Theorem 13 has the consequence for random heuristic search that $\tilde{\mathcal{G}}$ is well defined by the hierarchical relationship $[\mathcal{G}(p)] = \tilde{\mathcal{G}}([p])$ if and only if $\mathcal{G}$ is compatible with $\equiv$. The situation for τ is essentially the same, though as the instance τ of random heuristic search is defined with respect to its heuristic $\mathcal{G}$, so the instance $\tilde{\tau}$ should be defined with respect to its heuristic $\tilde{\mathcal{G}}$. It is therefore not at all clear that the definition of $\tilde{\tau}$ by way of τ_T – even if it is independent of T – is compatible with definition by way of its heuristic $\tilde{\mathcal{G}}$. The next theorem resolves this issue.

Theorem 14 *An instance τ of RHS is compatible with $\equiv$ if and only if its heuristic $\mathcal{G}$ is. Moreover, in that case*

$$\mathcal{G}^k(p) \equiv q \iff \tilde{\mathcal{G}}^k(\tilde{p}) = \tilde{q}$$

and

$$Pr\{\tilde{\tau}(\tilde{p}) = \tilde{q}\} = r! \prod_{j \in \tilde{\Omega}} \frac{\tilde{\mathcal{G}}(\tilde{p})_j^{r\tilde{q}_j}}{(r\tilde{q}_j)!}$$

for all $p, q \in \Lambda$ and $k > 0$. If $p, q \in \frac{1}{r}X_n^r$ then

$$Pr\{\tau^k(p) \equiv q\} = Pr\{\tilde{\tau}^k(\tilde{p}) = \tilde{q}\}$$

for all $k > 0$.

[8] When h is nondeterministic, $\tilde{h}$ is only determined up to distribution.

The basic framework is now in place for interpreting random heuristic search as operating on equivalence classes. The consequence of compatibility is that one does not need to know the detailed system state to obtain the dynamics of the quotient. In particular, fixed points x of $\mathcal{G}$ correspond to fixed points Ξx of $\tilde{\mathcal{G}}$. As a trajectory $\tau(t)$, $\tau^2(t)$, $\tau^3(t)$, $\ldots$ relates to fixed points of $\mathcal{G}$, so $\Xi\,\tau(t)$, $\Xi\,\tau^2(t)$, $\Xi\,\tau^3(t)$, $\ldots$ relates to fixed points of $\tilde{\mathcal{G}}$. Moreover, the previous theorem shifts the focus from τ to $\mathcal{G}$. Since compatibility of the heuristic suffices, the following result may be useful when $\mathcal{G}$ is expressed as a composition of functions on Λ.

Stochastic functions h and g are called *independent* provided that, for all w, x, y, z,

$$\Pr\{g(w) = x \,\wedge\, h(y) = z\} = \Pr\{g(w) = x\}\,\Pr\{h(y) = z\}$$

In particular, deterministic functions are independent.

Theorem 15 *If stochastic functions g and h map Λ to Λ, are independent, and are compatible with $\equiv$, then $\tilde{g}$ and $\tilde{h}$ are independent, $g \circ h$ is compatible with $\equiv$, and, as stochastic functions, $(g \circ h)\tilde{} = \tilde{g} \circ \tilde{h}$.*

5.2 Approximate Models

In situations where $\mathcal{G}$ is compatible with a nontrivial equivalence relation, one might be interested in $\tilde{\tau}$ or in $\tilde{\mathcal{G}}$ as an alternative to τ or $\mathcal{G}$. Objects are simpler in the quotient for the reason that $\tilde{\Omega}$ is smaller than Ω.

In situations where $\mathcal{G}$ is not compatible with the equivalence relation (and, by theorem 14, neither is τ), the dauntless may nevertheless choose to proceed at the peril of sacrificing any expectation that the equivalence class of a future generation bears any relationship – besides serendipitous – to that predicted by $\tilde{\tau}$.

Depending upon one's goals, that might be appropriate. Certainly $\tilde{\mathcal{G}}$ is perfectly well defined with respect to any choice T of equivalence class representatives, whether or not it happens to be compatible with the underlying equivalence relation. And, given any definition of $\tilde{\mathcal{G}}$ on the quotient space, one may consider the instance of random heuristic search over $\tilde{\Omega}$ having $\tilde{\mathcal{G}}$ as its heuristic.

Whereas the freedom allowed by the approach described in the previous paragraph (i.e., define $\tilde{\mathcal{G}}$ based on a choice for T, then take $\tilde{\tau}$ corresponding to $\tilde{\mathcal{G}}$) provides flexibility and hope of obtaining a reasonable fit by judicious choice, the hierarchical relationship may vanish – even in expectation! One could wind up in the situation of having a simple model about which nothing has been proved *except internally*; the resulting model is an instance of RHS, so the general theory of random heuristic search may be brought to bear *on the model* ...but the degree to which the model represents τ is another matter altogether!

When proof is an irrelevant concept, as when empirically validating a model by way of anecdotal examples, the outcome described above is of no consequence. Moreover, estimating $\tilde{\mathcal{G}}$ – rather than defining it with respect to T – may provide further simplification. If confidence in the model is desired, one may resort to empirical means, assuming the model's complexity is not a computational barrier.

As far as choosing T is concerned, the elements of $\Lambda/\!\equiv$ are convex compact sets (theorem 12), and so the average of $[t]$ is a natural candidate to represent $[t]$. One might alternatively pick a maximal element of $[t]$ with respect to entropy, for instance, as a representative (models employing some sort of maximum entropy assumption are not uncommon; see, for example, [11,15,17,18]). These two possibilities coincide, however.

An element $x \in \Re^n$ is said to be *dominated by* $\equiv$, denoted $x \prec\, \equiv$, provided

$$i \equiv j \quad\Longrightarrow\quad x_i = x_j$$

Theorem 16 *If $\Xi x = \Xi y$ and $x \prec\, \equiv$, then the entropy of x is greater than or equal to that of y.*

Theorem 17 *Let T be the set of equivalence class representatives given by averaging,*

$$T = \{\frac{1}{\lambda([x])}\int\limits_{[x]} y\, d\lambda(y) \;:\; x \in \Lambda\}$$

Then the representative $t \in T$ of $[x]$ has i th component

$$t_i = \frac{(\Xi x)_{[i]}}{\mid [i] \mid}$$

In particular, $t \prec\, \equiv$.

Combining theorems 16 and 17, it may be concluded that equivalence class representatives given by averaging have maximum entropy. This choice for T is convenient because it allows a simple characterization of $\tilde{\mathcal{G}}$.

Theorem 18 *If equivalence class representatives are chosen by maximum entropy, then*

$$\tilde{\mathcal{G}} = \Xi \circ \mathcal{G} \circ D \Xi^T$$

where D is the square diagonal matrix having ii th entry $\mid [i] \mid^{-1}$.

Another consideration in choosing T is invariance. Suppose there exists a set of representatives such that $\mathcal{G} : T \longrightarrow T$. In the case where T is chosen by maximum entropy, this is equivalent to the condition that

$$t \prec \equiv \implies \mathcal{G}(t) \prec \equiv$$

Since the hierarchical relationship

$$[\mathcal{G}(t)] = \tilde{\mathcal{G}}([t])$$

holds for $t \in T$, a consequence of invariance is the following.

Theorem 19 *If T is invariant under $\mathcal{G}$, then*

$$[\mathcal{G}^k(t)] = \tilde{\mathcal{G}}^k([t])$$

for all k, provided $t \in T$. Moreover, the local dynamics of τ as viewed in the quotient space – i.e., $\Xi\,\tau(t)$, $\Xi\,\tau^2(t)$, $\Xi\,\tau^3(t)$, $\ldots$ – is attracted to the local dynamics of $\tilde{\mathcal{G}}$ as population size increases, for population trajectories beginning in T.

As far as choosing $\equiv$ is concerned (assuming compatibility and invariance are not considerations), its definition depends on the main points of interest. For example, it may be natural, in the context of function optimization, to equivalence class based on fitness.

6 Example

The purpose of this section is not the analysis of a previously unexamined system. The point is rather to illustrate the theory presented in this paper by way of a concrete application. The example of this section – royal road functions – has been considered before [9,17].

The results presented in previous sections point towards fixed points as important objects. However, finding them is not necessarily trivial. In the case of the simple genetic algorithm (see section 3.3), the heuristic has the form

$$\mathcal{G} = \mathcal{M} \circ \mathcal{F}$$

and whereas the fixed points of $\mathcal{M}$ and $\mathcal{F}$ are known separately, those for the composition are not (see [25,27,29]).

One might consider "approximating" $\mathcal{G}$ by assuming zero crossover. In that case, the heuristic takes the form

$$\mathcal{G}(x) = \frac{Ax}{\mathbf{1}^T Ax}$$

for a matrix A which, given nonzero mutation, is positive. This is a well-known result which reduces several key concepts to more or less standard concepts from linear algebra. In particular, the fixed points of $\mathcal{G}$ are eigenvectors of A; apart from magnitude, $\mathcal{G}$ is simply matrix multiplication. Moreover, $\mathcal{G}$ is focused if A has a simple maximal eigenvalue (which is the case by Perron-Frobenius theory because A is positive [5]). In fact, the sequence $\mathcal{G}(p)$, $\mathcal{G}^2(p)$, $\mathcal{G}^3(p)$, ... is essentially the power method for calculating the corresponding positive eigenvector [30].

Giving no thought to compatibility issues, one may seek to further reduce complexity by passing to a simplified model based on fitness (see, for example, [17,18]). That is, consider the state space to be the possible fitness distributions which populations could take on. Given fitness function f, let its range be $\{y_0, \ldots, y_k\}$. Then a population $p \in \Lambda$ has fitness distribution $\tilde{p}$ defined by the component equations

$$\tilde{p}_i = \sum_{j \in \Omega} [f(j) = y_i]\, p_j$$

The situation just described is simply a case of quotients as described in section 5.1. Let the equivalence relation $\equiv$ be defined on Ω by

$$x \equiv y \Longleftrightarrow f(x) = f(y)$$

Let $S = \{s_0, \ldots, s_k\}$ be a set of equivalence class representatives such that $f(s_i) = y_i$. Renaming the $[s_i]$ th row of Ξ with i,

$$(\Xi p)_i = \sum_j [s_i \equiv j]\, p_j$$

$$= \sum_j [y_i = f(j)]\, p_j$$

Thus $\Xi p = \tilde{p}$, which, since the quotient representation space and the space of fitness distributions coincide, justifies the notation $\tilde{p}$ to denote the fitness distribution of p.

The following theorems (theorems 20 and 21) present preliminary results of a general nature which relate to the issue of compatibility in the context of population-based genetic algorithms (see [25,26]).

Theorem 20 *If the fitness f is dominated by $\equiv$, then the proportional selection, ranking selection, and tournament selection schemes are compatible with $\equiv$.*

When equivalence is defined with respect to fitness, as it is for the example of this section (i.e., $x \equiv y \iff f(x) = f(y)$), theorem 20 implies the equivalence relation is compatible with several commonly used selection schemes. The situation for mutation is not as simple.

An equivalence relation $\equiv$ is called *uniform with respect to translation* provided that for all $i, j, h, k \in \Omega$,

$$i \equiv j \implies |\,(i \oplus [h]) \cap [k]\,| = |\,(j \oplus [h]) \cap [k]\,|$$

That is, the cardinality of the intersection of the equivalence class of k with the translate by j of the equivalence class of h depends on the class of j rather than the particular value of j.

The next theorem is a sufficient, though not necessary, condition for the mutation scheme to be compatible with $\equiv$. The mutation distribution it refers to is the vector μ defined by

$$\mu_i = \Pr\{j \text{ mutates to } j \oplus i\}$$

Theorem 21 *If the mutation distribution μ is dominated by $\equiv$, and if $\equiv$ is uniform with respect to translation, then the mutation scheme is compatible with $\equiv$.*

In order to investigate compatibility further, details concerning $\mathcal{G}$ are required. Let the search space be $\mathcal{Z}_2^\ell$ (as in section 3.3, but with $c = 2$) and consider the class of degenerate Royal Road functions, which have the following form (see [9] for the general case). Let $\mathbf{1} = a_0 \oplus \cdots \oplus a_k$ where $a_i \otimes a_j = a_i[i = j]$. The fitness of x is given by

$$f(x) = \sum [a_i = x \otimes a_i]\, g_i$$

where g is some positive real vector. A particularly simple parametrized set of examples is given by $\ell = NK$, $g = \mathbf{1}$, and $a_i = 2^{iK}(2^K - 1)$. The positive integer parameters N and K correspond to a decomposition of the optimal string, $\mathbf{1}$, into N blocks of K contiguous 1s. An arbitrary string x has fitness equal to the number of blocks in common with $\mathbf{1}$. The range of f is therefore $\{0, \ldots, N\}$, hence y_i may be taken to be i and s_i may be taken to be $2^{iK} - 1$ (the paragraphs preceding theorem 20 introduce y_i and s_i).

Letting $\mathcal{G}$ be the heuristic for the simple genetic algorithm with proportional selection, zero crossover, positive mutation, and fitness function f (as described above, with parameters N and K) refines the instance of random heuristic search represented by the example of this section (this same example is treated in [17,18]). For the case $K = 1$, the analysis has an entirely different character, and while not difficult, will not be pursued here. Assume therefore that $K > 1$.

The equivalence relation $\equiv$ is not uniform with respect to translation, as is easily seen by the definition via the choice $h = k = \mathbf{1}$, $i = 0$, $j = \sum_u 4^u$. While not proof, this raises the suspicion that mutation is not compatible with $\equiv$. It is easily seen that the suspicion is actually the case; a population consisting entirely of i is equivalent to one consisting entirely of j, but the probability of the first producing – via mutation – a subsequent generation containing $\mathbf{1}$ is exponentially less than the probability of the second producing a subsequent generation containing $\mathbf{1}$ (in the first case all bits of a string must mutate, in the second case only half).

The example of the previous paragraph does more than show mutation is incompatible with $\equiv$ (that is, all strings with a given fitness cannot be treated as equivalent with respect to the dynamics of mutation), it shows that τ – which encompasses selection as well as mutation – is also incompatible, and hence (by theorem 14) so is $\mathcal{G}$.

A situation has now been arrived at where an equivalence relation $\equiv$ is defined over a search space Λ, its corresponding quotient map Ξ and quotient space $\tilde{\Lambda} = \Xi\Lambda$ are thereby defined, an instance τ of random heuristic search has been identified with its corresponding heuristic $\mathcal{G}$ (parametrized by N and K), ...but there is no natural well defined notion for either $\tilde{\mathcal{G}}$ or $\tilde{\tau}$, because both $\mathcal{G}$ and τ are incompatible with $\equiv$.

Following Rabinovich and Wigderson [11], let T be the set of equivalence class representatives corresponding to maximum entropy. By theorem 17, the representative $t \in T$ of $[x]$ has s_i th component

$$t_{s_i} = (\Xi x)_i \begin{pmatrix} N \\ i \end{pmatrix}^{-1} (2^K - 1)^{i-N}$$

and $t_i = t_j$ whenever $i \equiv j$. This choice of T corresponds to an assumption that the bit values in unaligned blocks are uniformly represented (random).

Since $\tilde{\mathcal{G}}$ is determined by $\tilde{\mathcal{G}}(\Xi t) = \Xi \mathcal{G}(t)$ for $t \in T$, the hierarchical relationship

$$[\mathcal{G}(t)] = \tilde{\mathcal{G}}([t])$$

holds – by definition – for $t \in T$, ...but it is hopeless (since $\mathcal{G}$ is incompatible with $\equiv$) to expect it will hold for elements which are not equivalence class representatives (i.e., elements for which the bit values in unaligned blocks are not random). One would expect, even if beginning at an initial population $t \in T$, that the hierarchical relationship would vanish after one application of τ.

If, however, randomness (i.e., maximum entropy) were preserved in expectation, then T would be invariant under $\mathcal{G}$. Appealing to theorem 19, the dynamics of τ as viewed through fitness distributions – i.e., $\Xi\, \tau(t)$, $\Xi\, \tau^2(t)$, $\Xi\, \tau^3(t)$, ... – would be attracted to the dynamics of $\tilde{\mathcal{G}}$ as population size increases, for population trajectories beginning in T. [9]

That is not the case, however. Given fixed positive mutation, the dynamics for τ is not attracted to the dynamics for $\tilde{\mathcal{G}}$ in any meaningful sense, because whereas selection preserves randomness of unaligned blocks, mutation does not. For example, consider the population $t \in T$ containing only copies of $\mathbf{1}$. The next generation is expected to contain strings of fitness zero, but all such strings do not occur with equal probability; 0 is exponentially less likely to occur than $\sum 4^i$. Hence maximum entropy is not preserved.

From the perspective of modeling, it is of little concern that exact theoretical coupling between τ and $\tilde{\tau}$ (or between $\mathcal{G}$ and $\tilde{\mathcal{G}}$) does not exist. It is still of interest to pursue $\tilde{\mathcal{G}}$ as an approximate model and to investigate the sense in which it approximates.

The situation for selection is altogether different from that for mutation. Because selection satisfies $t \prec \equiv \implies \mathcal{F}(t) \prec \equiv$, it follows that $\mathcal{G} : T \longrightarrow T$ when mutation is zero. By theorem 19, the dynamics of τ as viewed through fitness distributions is therefore attracted to the dynamics of $\tilde{\mathcal{G}}$ as population

[9] While not worked through in generality, the invariance principle (in this case, the preservation of entropy) was implicit in the analysis of Rabinovich and Wigderson.

size increases, provided mutation is zero and population trajectories begin at members of T.

However, more is true. Since selection is compatible with $\equiv$ (theorem 20), $\tilde{\mathcal{F}}$ is well defined independent of T (theorem 13), and the hierarchical relationships

$$\begin{aligned} [\mathcal{G}^k(x)] &= \tilde{\mathcal{G}}^k([x]) \\ \Pr\{\tau^k(p) \equiv q\} &= \Pr\{\tilde{\tau}^k(\tilde{p}) = \tilde{q}\,\} \end{aligned}$$

hold in the zero mutation case for all k and every initial population (theorem 14). By theorem 18 (and using the fact that $\mathbf{1}^T\Xi = \mathbf{1}^T$),

$$\tilde{\mathcal{G}}(\tilde{t}) = \frac{B\tilde{t}}{\mathbf{1}^T B\tilde{t}}$$

where $B = \Xi A D \Xi^T$. Given zero mutation this simplifies to

$$\tilde{\mathcal{F}}(\tilde{x}) = \frac{\operatorname{diag}(\langle 0,\ldots,N\rangle)\tilde{x}}{\langle 0,\ldots,N\rangle^T\tilde{x}}$$

Since $\mathcal{G}$ is a continuous function of mutation, so to is $\tilde{\mathcal{G}}(\Xi t) = \Xi\mathcal{G}(t)$. Hence, for small mutation, the local dynamics of $\tilde{\mathcal{G}}$ is nearly that of $\tilde{\mathcal{F}}$ (continuity), which is the image under Ξ of the local dynamics of $\mathcal{F}$ (theorem 14), which is nearly the image under Ξ of the local dynamics of $\mathcal{G}$ (continuity), which coincides with that of τ as viewed through fitness distributions as population size increases (theorem 4). Therefore, there is theoretical reason to hope that $\tilde{\mathcal{G}}$ approximately models trajectories through fitness distribution space:

Theorem 22

- *As the mutation rate decreases, the local dynamics of τ as viewed through fitness distributions converges to that of $\tilde{\tau}$.*
- *As the population size increases and the mutation rate decreases, the local dynamics of τ as viewed through fitness distributions converges to that of $\tilde{\mathcal{G}}$.*

The above theorem speaks to local (i.e., time bounded) dynamics. What about global dynamics? What can be said concerning fixed points and their stable and unstable manifolds as the mutation rate increases from zero?

The matrix $\operatorname{diag}(\langle 0,\ldots,N\rangle)$ has distinct eigenvectors, which correspond to the fixed points of $\mathcal{F}$; these are the vertices of $\tilde{\Lambda}$. As has been explained in [28], $\tilde{\mathcal{F}}$ is a normal heuristic. When it is regarded as acting on the sphere, call it $\mathcal{F}'$ in that context,

$$\mathcal{F}'(x) = \frac{\text{diag}(\langle 0,\ldots,N\rangle)x}{\|\text{diag}(\langle 0,\ldots,N\rangle)x\|}$$

its global dynamics are continuous; for small smooth perturbations, normality is preserved, the number and dimensions of fixed points are preserved, and their locations and stable and unstable manifolds vary continuously. However, the global dynamics on $\tilde{\Lambda}$ is, technically speaking, a different story. The addition of positive mutation, however small, changes the number of fixed points from N to 1; this is a simple consequence of Perron-Frobenius theory: there is a unique positive eigenvector of B in $\tilde{\Lambda}$ (since the matrix B is positive) and all of $\tilde{\Lambda}$ is contained within its basin of attraction [5].

What is happening here is that the global dynamics on the sphere is varying continuously, but fixed points – except for the one represented by the eigenvector corresponding to the maximal eigenvalue of B – are moving from the vertices of $\tilde{\Lambda}$ into the exterior of $\tilde{\Lambda}$ taking their stable manifolds with them. Although all but one fixed point leaves $\tilde{\Lambda}$, they still exert an influence on trajectories within $\tilde{\Lambda}$ by way of the continuity of the flow.

Since, for small mutation, $\tilde{\mathcal{G}}$ is a normal and regular heuristic, the general theory of random heuristic search provides a unified understanding of the mechanisms that control the dynamics and determine the quantitative and qualitative nature of $\tilde{\tau}$.

Qualitatively, one would expect to observe punctuated equilibria, even in regions where fitness is not locally optimal.[10] Moreover, periods of stasis in population fitness distributions are identified near the flow's fixed points whether or not they are contained within $\tilde{\Lambda}$ (see the discussion at the end of section 4.2). The following observations can be made about such regions:

- They are, for small mutation, near vertices of $\tilde{\Lambda}$, and are areas of low dispersion.
- They are regions where the force, $\tilde{\mathcal{G}}(\tilde{p}) - \tilde{p}$, is weak.
- They are regions where the noise, $\mathcal{E}(\|\tilde{\tau}(\tilde{p}) - \tilde{\mathcal{G}}(\tilde{p})\|^2)$, is weak.

As discussed in section 4.3.1, one expects to observe alternation between periods of stasis and a sudden change to a new dynamic equilibrium. This punctuated equilibria results from mechanisms fairly well understood in the theory of random heuristic search: the interplay between the flow and the lattice available to finite populations for occupation, the continuity of the underlying flow which supports population trajectories visiting fixed points in order of increasing dimension, the depressed dispersion, signal, and noise, and the ergodicity and logarithmic convergence of the heuristic.

[10] A specific example of this phenomenon, though in a different context, is given in [19].

One expects spatial fluctuations during an epoch to be approximately Gaussian (theorem 6) and the variance to scale inversely with the population size (theorems 3, 6). The spatial location of an epoch is not expected to change significantly as the population size varies, since it is determined by the dynamics of $\tilde{\mathcal{G}}$ (by theorem 1, the influence of population size is external to $\tilde{\mathcal{G}}$). However, population size is expected to impact its duration as well as the probability, both local in time and averaged over infinite time, of it being encountered (theorems 3, 8, 9, 11). From an asymptotic perspective, the meta-level chain indicates increasing dominance, as population size increases, of the epoch represented by the eigenvector corresponding to the maximal eigenvalue of B (theorem 11). From a transient perspective, the systems ability to follow the flow increases with population size (theorems 5, 8). Whereas many of these conclusions are reached in [17,18] for the specific example considered in this section, the conclusions here are seen to be consequences of the general theory of random heuristic search.

7 Conclusion

Parts of the theory of random heuristic search were illustrated in the previous section, though only in a qualitative and superficial way. The detailed information provided by theorem 1

$$\tilde{Q}_{\tilde{p},\tilde{q}} = r! \prod_{j \in \tilde{\Omega}} \frac{\tilde{\mathcal{G}}(\tilde{p})_j^{r\tilde{q}_j}}{(r\tilde{q}_j)!}$$

was not even touched (here $r = N + 1$). An analysis of $\tilde{\tau}$ based on

$$\nu^T (I - \tilde{Q})^{-1} \mathbf{1}$$

along the lines suggested in section 2.2 could be performed. Whereas the triviality of the example – it is essentially linear – would enable a fairly accurate quantitative analysis in terms of $d\tilde{\mathcal{G}}_x$ at eigenvectors x, the computational expense of computing eigenvectors compares with matrix inversion (for a treatment from that perspective, see [18]). With respect to theoretical analysis of the example, the advantage of $\tilde{\tau}$ over τ is unclear.

The reader interested in more details, further results, and analysis as applied to genetic algorithms is referred to [25].

References

[1] E. Akin, *The General Topology of Dynamical Systems* (American Mathematical Society, 1993).

[2] M. Arbib and E. Manes, *Arrows, Structures, And Functors, the categorical imperative* (Academic Press, New York, 1975).

[3] M. Eberlein, The GA Heuristic Generically has Hyperbolic Fixed Points, Ph. D. Dissertation, The University of Tennessee, 1996.

[4] W. Feller, *An Introduction to probability Theory and Its Applications* (Wiley, New York, 1968).

[5] F. R. Gantmacher, *Matrix Theory* (Chelsea, 1997).

[6] J. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor, 1975).

[7] J. Juliany and M. D. Vose, The Genetic Algorithm Fractal, *Evolutionary Computation* **v. 2**, **n. 2**, (1994) 165–180.

[8] G. Koehler, S. Bhattacharyya, and M. D. Vose, General Cardinality Genetic Algorithms, *Evolutionary Computation*, **v. 5**, **n. 4**, (1997) 439–459.

[9] M. Mitchell, J. Holland, and S. Forrest, When will a genetic algorithm outperform hill climbing?, in: Cowan, Tesauro, and Alspector, eds., *Advances in Neural Information Processing Systems 6* (Morgan Kaufmann, San Mateo, CA., 1994).

[10] A. Nix and M. D. Vose, Modeling Genetic Algorithms With Markov Chains, *Annals of Mathematics and Artificial Intelligence*, **5** (1992) 79–88.

[11] Y. Rabinovich and A. Wigderson, An Analysis of a Simple Genetic Algorithm, in: Belew and Booker eds., *Proceedings of the Fourth International Conference on Genetic Algorithms* (Morgan Kaufmann, 1991) 215–221.

[12] N. Radcliffe and P. Surry, Fundamental Limitations on Search Algorithms: Evolutionary Computing in Perspective, in: *Lecture Notes in Computer Science, 1000* (Springer-Verlag, New York, 1995) 275–291.

[13] A. Renyi, *Probability Theory* (North-Holland, Amsterdam, 1970).

[14] Rowe, Population fixed-points for functions of unitation, in: *Foundations Of Genetic Algorithms 5* (Morgan Kaufmann, to appear).

[15] J. Shapiro and A. Prugel-Bennett, Maximum Entropy Analysis of Genetic Algorithm Operators, in: *Lecture Notes in Computer Science, 993* (Springer-Verlag, Berlin, 1995) 14–24.

[16] J. Shapiro, A. Prugel-Bennett, and M. Rattray, A Statistical Mechanical Formulation of the Dynamics of Genetic Algorithms, in: *Lecture Notes in Computer Science, 865* (Springer-Verlag, Berlin, 1994) 17–27.

[17] E. van Nimwegen, J. Crutchfield, and M. Mitchell, Finite Populations Induce Metastability in Evolutionary Search, *Phys. Lett. A* **v. 229**, (1997) 144–150.

[18] E. van Nimwegen,J. Crutchfield, and M. Mitchell, Statistical Dynamics of the Royal Road Genetic Algorithm, *Theoretical Computer Science*, this issue.

[19] M. D. Vose, A Closer Look At Mutation In Genetic Algorithms, *Annals of Mathematics and Artificial Intelligence*, **10**, (1994) 423–434.

[20] M. D. Vose, Formalizing Genetic Algorithms, in: *Proc. IEEE workshop on Genetic Algorithms, Neural Nets, and Simulated Annealing applied to problems in Signal and Image Processing*, May 1990, Glasgow, U.K.

[21] M. D. Vose, Modeling Alternate Selection Schemes For Genetic Algorithms, in: Koppel and Shamir eds., *Proceedings of BISFAI '95*, (1995) 166–178.

[22] M. D. Vose, Modeling Simple Genetic Algorithms, *Evolutionary Computation,* **v. 3**, **n. 4**, (1995) 453–472.

[23] M. D. Vose, What are Genetic Algorithms? a mathematical perspective, in: Davis, De Jong, Davis, Vose, Whitley eds., *Evolutionary Algorithms, Vol. 111* (Springer-Verlag, New York, 1999) 251–276.

[24] M. D. Vose, Logarithmic Convergence of Random Heuristic Search, *Evolutionary Computation*, **v. 4**, **n. 4**, (1996) 395–404.

[25] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory* (MIT press, 1999).

[26] M. D. Vose, Random Heuristic Search: applications to gas and functions of unitation, *University of Tenness Technical Report* ut-cs-98-402.

[27] M. D. Vose and G. Liepins, Punctuated Equilibria In Genetic Search, *Complex Systems,* **5** (1991) 31–44.

[28] M. D. Vose and A. Wright, Simple Genetic Algorithms with Linear Fitness, *Evolutionary Computation,* **v. 2**, **n. 4**, (1994) 347–368.

[29] M. D. Vose and A. Wright, The Walsh Transform and the Theory of the Simple Genetic Algorithm, in: S. Pal and P. Wang eds., *Genetic Algorithms for Pattern Recognition* (CRC Press, Boca Raton, 1996) 25–43.

[30] J. H. Wilkinson, *The Algebraic Eigenvalue Problem* (Oxford University Press, London, 1965).

[31] D. Wolpert and W. Macready, No Free Lunch Theorems for Search, Santa Fe Institute Technical Report, (1994) SFI-TR-95-02-010.

[32] A. Wright and M. D. Vose, Finiteness of the Fixed Point Set for the Simple Genetic Algorithm. *Evolutionary Computation*, **v. 3**, **n. 3**, (1995) 299–309.