Modeling Simple Genetic Algorithms

Michael D. Vose
Department of Computer Science
107 Ayres Hall
The University of Tennessee
Knoxville, TN 37996-1301
vose@cs.utk.edu

Abstract
The infinite- and finite-population models of the simple genetic algorithm are extended and unified. The result incorporates both transient and asymptotic GA behavior. This leads to an interpretation of genetic search that partially explains population trajectories. In particular, the asymptotic behavior of the large-population simple genetic algorithm is analyzed.
Keywords
Asymptotic behavior, Markov chain, steady state distribution.

1. Introduction

Mathematical models have been used as analytical tools in the investigation of genetic algorithms. An early example is Goldberg's work on the minimal deceptive problem (Goldberg, 1987). He used equations for the expected next generation to model the evolutionary trajectory of a two-bit GA under crossover and proportional selection. The results were used to obtain the now familiar type-I and type-II classifications of minimal deceptive problems.

Vose (1990) simplified and extended these equations, incorporating mutation with the recombination of arbitrarily long binary strings. Like Goldberg's, the equations are deterministic and yield what is equivalent to the evolutionary path taken by a GA with infinite population.

Initial observations concerning infinite-population models include the following: Real GAs are based on finite populations and are stochastic, not deterministic. Moreover, the relationship between the trajectory of expectations (which is the path followed by infinite-population models) and the evolution of finite populations in real GAs is unclear.

Confronting the concerns noted above, these issues were explored; the result was twofold. First, an exact Markov-chain model for finite-population GAs was obtained in a form making the interdependency between the finite- and infinite-population models explicit. Second, the trajectory followed by a finite-population GA was related to the evolutionary path predicted by the infinite-population model (Nix & Vose, 1991). The present paper builds on these previous results to further tie finite-population GAs to the infinite-population model.

Coincidentally, Davis independently modeled GAs with Markov chains (Davis, 1991). However, his work focused on whether annealing the mutation rate implies convergence to the global optimum (it does not). Moreover, the model he obtains, while equivalent, does not make evident the manner in which the infinite-population model is an integral component of the transition matrix for the finite-population case.

Roughly speaking, populations correspond to points on a surface and the progression from one generation to the next forms a path leading toward a local fixed point of the infinite-population model. In the infinite-population case, genetic algorithms should not be thought of as global optimizers. Populations move toward the local fixed point in whose basin of attraction they began.

But what about *finite* GAs? Vose proved that for large populations the evolutionary path of a finite-population GA follows very closely, with large probability, and for a long period of time that path predicted by the infinite-population model (Nix & Vose, 1991). Thus, *transient* behavior of a finite-population GA, i. e., the behavior that depends on the initial population, is determined (aside from stochastic effects) by the fixed points of the infinite-population model and their basins of attraction.

When the mutation rate is greater than zero, a finite-population GA typically forms an ergodic Markov chain, visiting every state infinitely often. Hence, after some period of time, a GA will escape *every* local basin. The interesting asymptotic question is: *Where is it likely to be?*[1]

This question is answered by a probability distribution that is converging to the steady-state distribution of the Markov chain. This steady-state distribution concentrates probability near fixed points of the infinite-population model. Hence, the genetic algorithm will escape one local basin only to be temporarily trapped near the fixed point of another. However, the possibility remains that a GA may spend a disproportionate amount of time near some particular fixed point, and, in the long run, that is where it is most likely to be.

The result of this paper is that a large finite-population GA will, with probability close to 1, be asymptotically near that fixed point having the largest basin of attraction. As population size grows, the probability of its spending a nonvanishing proportion of time anywhere else (when observed over increasingly long periods of time) converges to 0. Therefore, short-term behavior is determined by the basin in which the initial population finds itself, and long-term behavior is determined by the fixed point having the largest basin.

This paper concerns the emergent behavior of simple genetic search. The contribution this paper makes to the theory of evolutionary computation is this view of asymptotic behavior. Moreover, it is plausible that the transient behavior (i. e., the short-term path followed by genetic search) in many cases may be correlated with asymptotic behavior. Since asymptotics are determined by the fixed point that has the largest basin, and since a large basin may be more probable as the one containing a random initial population, the connection between the two warrants further study.

This overview began (six paragraphs back) with the phrase "Roughly speaking" for two good reasons. First, some results have intricate mathematical form and are not so directly translated into common language. Second, some results are conditional on mathematical conjectures; the nonlinear mathematics of GAs are complex and are not completely understood.

To present the analysis in a form applicable to a wide variety of GA variants, the next section reviews a broad framework that can be instantiated to various generational-style genetic algorithms. General results are then developed within this framework, to be specialized later to GAs in the conclusion.

1 In an ergodic Markov chain (of the type discussed in this paper), the answers to "Where is it most likely to be?" and "In which state is it most likely to spend time?" coincide as time increases without bound.

2. Random Heuristic Search

Included mainly for the reader's convenience, this section abstracts from Vose (1990) and Vose and Wright (1994a); specialized details may be obtained from those sources. The purpose of this section is to establish the framework in which the analysis of this paper is carried out.

We consider an abstract population-based search algorithm referred to as *Random Heuristic Search* (RHS). Many generational genetic algorithms are special cases of this general framework. An instance of random heuristic search can be thought of as an initial collection of elements P_0 chosen from some search space Ω, together with a stochastic transition rule τ, which from P_i will produce another collection P_{i+1}. In general, τ will be iterated to produce a sequence

$$P_0 \overset{\tau}{\longmapsto} P_1 \overset{\tau}{\longmapsto} P_2 \overset{\tau}{\longmapsto} \cdots$$

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 generated successively until some stopping criterion is reached, when it is hoped that the object of search has been found.

The algorithms comprised by random heuristic search are constrained as to which transition rules are allowed. Characterizing admissible τ will be postponed until after the representation scheme for populations has been reviewed. Let n be the cardinality of Ω and identify the integer i with the ith element of Ω through some fixed enumeration beginning at 0. Define the *simplex* to be the set

$$\Lambda = \{\langle x_0, \ldots, x_{n-1}\rangle : \forall j \cdot x_j \geq 0,\ \mathbf{1}^T x = 1\}$$

where angle brackets (as used above) denote column vectors, and where $\mathbf{1}$ is the n-dimensional vector $\langle 1, \ldots, 1\rangle$. An element p of Λ corresponds to a population according to the rule

$$p_i = \text{the proportion of } i \text{ contained in the population}$$

The cardinality of each generation is a constant 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* (or *descriptor*). For example, the population vector

$$\langle 0.1, 0.2, 0.0, 0.4, 0.3\rangle$$

refers to a search space of cardinality 5 (there are five components). The population described contains 10% of 0 (representing the 0th element of Ω—indexing always begins at 0), 20% of 1 (representing the first element of Ω), 40% of 3 (representing the third element of Ω), and 30% of 4 (representing the fourth element of Ω).

Given the current population P, the next population $Q = \tau(P)$ cannot be predicted with certainty because τ is stochastic; Q results from r independent, identically distributed random choices. Let $\mathcal{G}\colon \Lambda \to \Lambda$ be a function that given the current population vector p produces a vector whose ith component is the probability that i is the result of a random choice. Thus, $\mathcal{G}(p)$ is the probability vector that specifies the distribution from which the aggregate of r choices forms the subsequent generation Q. A transition rule τ is admissible if it corresponds to a heuristic function $\mathcal{G}$ in this way.

In terms of search, P is the starting configuration with corresponding descriptor p, and $\mathcal{G}(p)$ is the *heuristic* according to which the search space is to be explored. The result $Q = \tau(P)$

of that exploration (i.e., r samples chosen according to $\mathcal{G}(p)$) is the next generation. A new heuristic $\mathcal{G}(q)$ (here q is the descriptor of Q) may be invoked to repeat the cycle.

Perhaps the first and most natural question concerning random heuristic search is: What connection is there between the heuristic used and the expected next generation? The answer is that if p is the current population vector and $\mathcal{G}$ is the heuristic, the expected next population vector is $\mathcal{G}(p)$.

According to the law of large numbers, if the next generation's population vector q were obtained as the result of an infinite sample from the distribution described by $\mathcal{G}(p)$, q would match the expectation, and hence $q = \mathcal{G}(p)$. Because this corresponds to random heuristic search with an infinite population, the algorithm resulting from "$\tau = \mathcal{G}$" is called the *infinite population algorithm.*[2]

Many generational GAs currently in use fit into the framework of random heuristic search, provided that only one child of the recombination/mutation process is kept. If the following outline is followed, an instance of random heuristic search is obtained.

1. Generate an initial population x containing r elements from Ω.
2. Produce a child for the next generation using any stochastic function of x.
3. If the next generation is incomplete, repeat step 2.
4. Replace x by the new generation just formed and go to step 2.

The heuristic function corresponding to this instance of random heuristic search is determined by the component equations

$$\mathcal{G}(x)_i = \text{Prob}\{\text{child } i \text{ results from step 2, given the population } x\}$$

An instance of random heuristic search is *focused* if $\mathcal{G}$ is continuously differentiable and for every $p \in \Lambda$ the sequence of iterates

$$p, \quad \mathcal{G}(p), \quad \mathcal{G}^2(p), \quad \ldots$$

converges. In this case, $\mathcal{G}$ is also called *focused.* In terms of search, the latter condition means that following the path that the heuristic function is expected to produce will lead to some state x. By the continuity of $\mathcal{G}$,

$$\mathcal{G}\left(\lim_{l \to \infty} \mathcal{G}^l(p)\right) = \lim_{l \to \infty} \mathcal{G}^{l+1}(p) = x$$

Hence, such states x satisfy $\mathcal{G}(x) = x$ and are called *fixed points.* It turns out that fixed points are particularly relevant to both the transient and the asymptotic behavior of focused random heuristic search. It is an open question as to when a genetic algorithm's heuristic function is focused. The conjecture that it typically is (apart from degenerate situations involving mutation rates greater than 0.5) and in fact has a Lyapunov function is a fundamental conjecture. The validity of this conjecture will be assumed throughout the paper.

To put this into perspective, there is no known nontrivial example where a simple GA is not focused. In the case of a GA over a space of fixed-length binary strings using proportional selection, crossover, and small mutation rates, the conjecture has been proved for fitness functions with low epistasis (Vose & Wright, 1994a).

2 Strictly speaking, τ produces the next generation from the current, while $\mathcal{G}$ produces the *representation* of the expected next generation from the *representation* of the current. The distinction between a population and its representation is conveniently blurred.

3. The Finite-Population Model

This section summarizes and extends Nix and Vose (1991). The purpose of this section is to make explicit the stochastic details of random heuristic search and to develop approximations for probabilities.

Identifying a population with the population vector that represents it, the possible populations of size r correspond to a set S_r of points in the simplex Λ. An *exact model* of random heuristic search is provided by the following Markov chain. The states are given by elements of S_r and the transition probabilities are given by the matrix Q, where

$$Q_{x,x'} = r! \prod_{j=0}^{n-1} \frac{\{\mathcal{G}(x)_j\}^{rx'_j}}{(rx'_j)!}$$

Observe that since the representation of a population by a population vector x is proportional, x may represent populations of unspecified size. As $r \to \infty$, the state space of the Markov chain becomes dense in Λ and the limit of the transition behavior of the Markov chain is a discrete dynamical system on Λ with transition operator $\mathcal{G}$. Thus, both the finite- and infinite-population models are viewed as operating in Λ.

An instance of random heuristic search is *ergodic* if the finite-population model (the Markov chain defined above) is ergodic for every population size. This is always the case for a GA, given positive mutation, and will be assumed throughout the rest of the paper.

Using Stirling's theorem yields

$$Q_{x,x'} = \exp\{-r\alpha_{x,x'} + O(\ln r)\}$$

as $r \to \infty$ where

$$\alpha_{x,x'} = \sum_j x'_j \ln \frac{x'_j}{\mathcal{G}(x)_j}$$

LEMMA 3.1: *For any* $U \subset \Lambda$,

$$\text{Prob}\{\tau(p) \in U\} = \exp\{-r \inf_{q\in U} \alpha_{p,q} + O(\ln r)\}$$

where the constant in the "big O" depends on n.

SKETCH OF PROOF: Appealing to the formula preceding Lemma 3.1,

$$\text{Prob}\{\tau(p) \in U\} = \sum_{q\in U} \exp\{-r\alpha_{p,q} + O(\ln r)\}$$

A lower bound for the right-hand side is

$$\sup_{q\in U} \exp\{-r\alpha_{p,q} + O(\ln r)\}$$

An upper bound is

$$\binom{n+r-1}{n-1} \sup_{q\in U} \exp\{-r\alpha_{p,q} + O(\ln r)\} \quad = O(r^{n-1}) \exp\{-r \inf_{q\in U} \alpha_{p,q} + O(\ln r)\} \qquad \square$$

LEMMA 3.2: *If $\mathcal{G}$ maps Λ into its interior, there exist positive constants α and β depending on $\mathcal{G}$ such that*

$$\alpha \|v - \mathcal{G}(u)\|^2 \leq \alpha_{u,v} \leq \beta \|v - \mathcal{G}(u)\|^2$$

SKETCH OF PROOF: Expanding $\ln z$ in powers of $1 - 1/z$ leads to the following asymptotic relation as $u \to v$:

$$v \ln \frac{v}{u} = v - u + \frac{(v-u)^2}{v} \left\{ \frac{1}{2} + o(1) \right\}$$

Hence

$$\alpha_{u,v} = \left(\frac{1}{2} + o(1) \right) \sum_i \frac{(v_i - \mathcal{G}(u_i))^2}{v_i}$$

Since continuous $\mathcal{G}$ maps compact Λ into its interior, the denominators in the preceding sum are bounded away from zero as $v \to \mathcal{G}(u)$. This proves the result as $v \to \mathcal{G}(u)$. The restriction (that $v \to \mathcal{G}(u)$) is removed by compactness of Λ and the observation that $\alpha_{u,v} = 0 \iff v = \mathcal{G}(u)$. □

Combining Lemmas 3.1and 3.2 yields Theorem 3.3.

THEOREM 3.3: *Let* d *be the Euclidean distance from $\mathcal{G}(p)$ to U. If $\mathcal{G}$ maps Λ into its interior,*

$$\text{Prob}\{\tau(p) \in U\} = e^{-rO(d^2)+O(\ln r)}$$

where the constant in the "big O" depends on n *and $\mathcal{G}$.*

In the case of the simple genetic algorithm, it suffices for Theorem 3.3 that mutation is positive. The following lemma is a general result of probability theory concerning the characteristic function of a multinomial distribution (see Renyi, 1970). Let the current population vector p be fixed and let $q = \mathcal{G}(p)$. Define the random vector η by the component equations

$$\eta_j = \frac{\tau(p)_j - q_j}{\sqrt{q_j/r}}$$

LEMMA 3.4: *Let $\mathcal{E}$ denote expectation. Then*

$$\lim_{r \to \infty} \mathcal{E}(e^{\sqrt{-1}\eta^T t}) = \exp\left\{ -\frac{1}{2} \left(\sum t_j^2 - \left(\sum t_j \sqrt{q_j} \right)^2 \right) \right\}$$

The next result is related to approximating the transition probabilities of RHS by a multinormal distribution.

LEMMA 3.5: *Let ξ be an* (n − 1)*-dimensional random vector with density*

$$\rho(y) = (2\pi)^{-(n-1)/2}\, e^{-y^T y/2}$$

and let C *be an* n *by* (n − 1) *matrix having orthonormal columns perpendicular to* h = $\langle \sqrt{q_0}, \dots \sqrt{q_{n-1}} \rangle$. *Then η converges in distribution to* Cξ *as* r $\to \infty$.

SKETCH OF PROOF: It suffices to show

$$\lim_{r \to \infty} \mathcal{E}(e^{\sqrt{-1}\eta^T t}) = \mathcal{E}(e^{\sqrt{-1}(C\xi)^T t})$$

The right-hand side is the characteristic function of a multinormal (see Renyi, 1970) at $C^T t$,

$$\mathcal{E}(e^{\sqrt{-1}\xi^T(C^T t)}) = \exp\left\{-\frac{1}{2}t^T C C^T t\right\}$$

By Lemma 3.4, the left-hand side is

$$\exp\left\{-\frac{1}{2}t^T(I - hh^T)t\right\}$$

Thus it suffices that $CC^T = I - hh^T$. This follows easily by considering how both sides map the basis consisting of h and the columns of C. □

THEOREM 3.6: *For any open subset U of $\mathbf{1}^{\perp}$,*

$$\text{Prob}\{\tau(p) \in \mathcal{G}(p) + U/\sqrt{r}\} \longrightarrow (2\pi)^{-(n-1)/2} \int_{C^T \, diag\,(h)^{-1} U} e^{-y^T y/2}\, dy$$

as $r \to \infty$.

SKETCH OF PROOF: Using Lemma 3.5,

$$\begin{aligned}
& \text{Prob}\{\, \tau(p) \in \mathcal{G}(p) + U/\sqrt{r}\,\} \\
= \; & \text{Prob}\{\, \sqrt{r}(\tau(p) - q) \in U\,\} \\
= \; & \text{Prob}\{\, diag\,(h)\eta \in U\,\} \\
= \; & \text{Prob}\{\, \eta \in diag\,(h)^{-1}U\,\} \\
\to \; & \text{Prob}\{\, C\xi \in diag\,(h)^{-1}U\,\} \\
\leq \; & \text{Prob}\{\, \xi \in C^T diag\,(h)^{-1}U\,\} \\
\leq \; & \text{Prob}\{\, C\xi \in (I - hh^T)\, diag\,(h)^{-1}U\,\} \\
= \; & \text{Prob}\{\, C\xi \in diag\,(h)^{-1}U\,\}
\end{aligned}$$

The last equality follows from the fact that $I - hh^T$ is the identity on $h^{\perp}$. □

4. The Fixed Point Graph

This section introduces a Markov chain $\mathcal{C}_r$ which will be proved to approximately model the behavior of random heuristic search in a later section.

A fixed point ω of $\mathcal{G}$ is called *hyperbolic* if the differential $d\mathcal{G}_\omega$ has no eigenvalues on the unit circle (of the complex plane). A focused instance of random heuristic search is called *hyperbolic* if all fixed points (of $\mathcal{G}$ in Λ) are hyperbolic. A hyperbolic fixed point ω is classified according to whether all of the eigenvalues of the differential $d\mathcal{G}_\omega$ are contained in the interior of the unit disk. If so, ω is called *stable*. When eigenvalues are exterior to the unit disk, ω is *unstable*.

The remainder of this paper deals with hyperbolic, ergodic random heuristic search. It is currently not known whether the instances of random heuristic search represented by GAs are usually hyperbolic. One reason is that $\mathcal{G}$ has not been explicitly determined for many GA variants. However, in the case of a simple GA (using proportional selection, crossover, and mutation) over a space of fixed-length binary strings, it has been proved that generically there are finitely many fixed points (Wright & Vose, 1994b).[3] This result is

3 A situation is *generic* if it holds for the parameters in some dense open set. In this paper, the parameters in question are the fitness values.

encouraging, since hyperbolicity implies the number of fixed points is finite. Moreover, all empirical evidence—direct calculation of the spectrum of $d\mathcal{G}_\omega$—supports the conjecture that hyperbolicity is generic (in the case of a simple binary GA using proportional selection, crossover, and positive mutation).[4]

Given an object x with associated numerical quantities, let $|x|$ denote their sum. This provides a flexible and general concept of "cost" or "size" that is notationally convenient. When x is a vector, $|x|$ denotes the sum of its coordinates, which agrees with the ℓ_1 norm if x has nonnegative components.

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. The points $x_1, \ldots, x_{k-1}$ are referred to as *interior points* of ρ. The *cost* of ρ is

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

Let the stable fixed points of $\mathcal{G}$ in Λ be $\{\omega_0, \ldots, \omega_w\}$. In Vose (1993a), where the main ideas of this paper were sketched, the *fixed point graph* was defined to be the complete directed graph $\Im$ on vertices $\{0, \ldots, w\}$ with edge $i \rightarrow j$ (for $i \neq j$) having weight

$$\inf_{x \in \{y:\ \lim \mathcal{G}^k(y) = \omega_j\}} \sum_{k=0}^{n-1} x_k \ln \frac{x_k}{(\omega_i)_k}$$

This weight is the minimum relative entropy (or Kullback-Lieber information) of U_j relative to ω_i. While these weights are an approximation, it turns out that the appropriate definition of edge weight is more complicated.[5] The proper choice of weight for edge $i \rightarrow j$ is

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

A *tributary* of a complete directed graph is a spanning intree.[6] 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 Markov chain is represented by a complete directed graph over $\{0, \ldots, w\}$ when the $i \rightarrow j$ edge for $i \neq j$ is labeled by a weight encoding the i,jth entry of its transition matrix (the $i = j$ entries will be ignored because they can be inferred from the fact that row sums are 1). There is a beautiful connection between a Markov chain's steady-state distribution and the tributaries of a graphical representation of this kind (see Vose, 1993b, for a discussion and proof of the following):

THEOREM 4.1: *Let A be a transition matrix for a Markov chain with states $\{0, \ldots, w\}$, and let the corresponding graphical representation have edge $i \rightarrow j$ weighted by $-\ln A_{i,j}$ for $i \neq j$. A solution x to the steady-state equation $x^T A = x^T$ has components*

$$x_k = \sum_{t \in \text{Tree}_k} \exp\{-|t|\}$$

A *steady-state solution* refers to any solution x of the steady-state equation. The steady-state distribution of the Markov chain, assuming it exists, is obtained by dividing x by $\mathbf{1}^T x$. Note that Theorem 4.1 makes sense even if $A_{i,j} = 0$ through the convention that $0 = \exp\{-\infty\} = \exp\{\ln 0\}$.

4 This conjecture has recently been proved by Mary Eberlein and Michael D. Vose.

5 Or at least it appears to be more complicated; the two may be interchangeable for many purposes.

6 I.e., a tree containing every vertex and such that all edges point toward the root.

Suppose a graph G is given, similar to that referred to in the paragraph preceding Theorem 4.1, where $g_{i,j}$ labels its $i \rightarrow j$ edge (for $i \neq j$). Consider the Markov chain C_r, parametrized by r, with $i \rightarrow j$ transition probability

$$\exp\{-r\, g_{i,j} + o(r)\}$$

It follows from Theorem 4.1 that C_r has steady-state solution

$$x = \left\langle \sum_{t \in \mathrm{Tree}_0} e^{-r(|t|+o(1))}, \ldots, \sum_{t \in \mathrm{Tree}_{\varpi}} e^{-r(|t|+o(1))} \right\rangle^T$$

where the Tree_k are computed with respect to G. Suppose further that G has a unique minimum cost tributary rooted at k' and having cost c. In this case, the steady-state distribution of C_r converges as $r \rightarrow \infty$ to point mass at k' since

$$\begin{aligned} x_k\, e^{r(c+o(1))} &= \sum_{t \in \mathrm{Tree}_k} e^{-r(|t|-c+o(1))} \\ &= o(1) \end{aligned}$$

if $k \neq k'$ (because then $|t| > c$), and otherwise

$$\begin{aligned} x_{k'}\, e^{r(c+o(1))} &= 1 + \sum_{t \in \mathrm{Tree}_{k'}} [|t| > c]\, e^{-r(|t|-c+o(1))} \\ &= 1 + o(1) \end{aligned}$$

By choosing $G = \Im$, a Markov chain C_r is thereby defined up to $o(r)$ terms as above. It is conjectured, though as yet unproven, that an arbitrarily small perturbation of $\mathcal{G}$ (effected in the case of the simple GA by an arbitrarily small perturbation of crossover, mutation, or fitness) will guarantee that there is a unique minimum cost tributary of $\Im$. That such a tributary exists will also be assumed.

A later section will show how C_r captures the asymptotic behavior of random heuristic search as the population size $r \rightarrow \infty$. A consequence will be that for large populations, RHS will with large probability be asymptotically near that fixed point $\omega_{k'}$ of $\mathcal{G}$ corresponding to the minimum cost tributary of $\Im$.

5. Unstable Fixed Points

The existence of unstable fixed points presents added complexity to the proof of the main result. The purpose of this section is to develop preliminary results that will allow them to be successfully dealt with. This section assumes familiarity with the stable manifold theorem and related results (see Akin, 1993, Chapter 10).

The next theorem partially justifies the focus in Section 4 on stable fixed points. The heuristic $\mathcal{G}$ of random heuristic search is called *normal* if

1. $\mathcal{G}$ is hyperbolic.
2. $\mathcal{G}$ has a Lyapunov function: continuous ϕ such that $x \neq \mathcal{G}(x) \Longrightarrow \phi(x) < \phi(\mathcal{G}(x))$.

Assuming a normal, ergodic heuristic—an assumption made throughout the rest of this paper—the proportion of time spent by random heuristic search near unstable fixed points vanishes as $r \to \infty$. That is the main result of this section. Before proceeding, a completion operator will be defined and a few of its properties will be noted.

Let $\mathcal{S}$ be the basin of attraction for a fixed point of $\mathcal{G}$. Define $\mathcal{S}_0 = \mathcal{S}$ and let $\mathcal{S}_{i+1} = \bigcup \mathcal{S}'$ where the union is over all basins of attraction $\mathcal{S}'$ corresponding to fixed points in the closure of $\mathcal{S}_i$. Since it is assumed that there are finitely many fixed points (a consequence of hyperbolicity and the compactness of Λ), the sequence $\mathcal{S}_0, \mathcal{S}_1, \ldots$ must become stationary. The *completion* of $\mathcal{S}$ is defined to be the limit of this sequence and is denoted by $\overline{\mathcal{S}}$. For $A \subset \mathcal{S}, \overline{A}$ is defined as $\overline{\mathcal{S}}$.

LEMMA 5.1:

- *$\overline{\mathcal{S}}$ is closed.*
- *If $\mathcal{S}$ corresponds to an unstable fixed point, $\overline{\mathcal{S}}$ contains no stable fixed point.*
- *If $x \notin \overline{\mathcal{S}}$, $\{\mathcal{G}^k(x)\colon\ k \geq 0\}$ is a positive distance away from $\overline{\mathcal{S}}$.*

SKETCH OF PROOF: If $\overline{\mathcal{S}}$ were not closed, let $\overline{\mathcal{S}} = \mathcal{S}_i$ and let $x_j \in \overline{\mathcal{S}}$ be a sequence converging to $x \notin \overline{\mathcal{S}}$. Since $\mathcal{G}$ is focused, x is contained in the basin of attraction of a fixed point ω' in the complement of the closure of $\mathcal{S}_i$ (otherwise $x \in \mathcal{S}_{i+1} = \mathcal{S}_i$). Let U be any open neighborhood of ω', and let V be an open neighborhood of x that is mapped into U by $\mathcal{G}^k$ for some k (the orbit of x converges to ω' and $\mathcal{G}$ is continuous). Since $\mathcal{S}_i$ is invariant under $\mathcal{G}$ (it is a union of basins of attraction), it follows that $\mathcal{S}_i \cap U \neq \emptyset$ (it contains $\mathcal{G}^k(x_j)$ for some j and k). Hence ω' is in the closure of $\mathcal{S}_i$, a contradiction.

Suppose that ω is a stable fixed point and that $\mathcal{S}$ corresponds to an unstable fixed point. If $\mathcal{S}_i$ is separated from ω by an open set (true for $i = 0$), the basin of attraction of ω will not be included in the union defining $\mathcal{S}_{i+1}$. Hence $\mathcal{S}_i$ is separated from ω by an open set for all i.

Since $\mathcal{G}$ is focused, the set $\{\mathcal{G}^k(x)\colon\ k \geq 0\}$ has unique limit point ω. Thus, $\{\omega\} \cup \{\mathcal{G}^k(x)\colon\ k \geq 0\}$ is a closed set disjoint from $\overline{\mathcal{S}}$ (otherwise $x \in \overline{\mathcal{S}}$). Hence there is a minimum distance between these disjoint compact sets. □

The following lemma is needed in the proof of Lemma 5.3. The open ball of radius ε about the element or set x is denoted by $\mathcal{B}_\varepsilon(x)$.

LEMMA 5.2: *Let $\mathcal{G}$ be normal. Given $\eta > 0$, there exists $\xi > 0$ such that*

$$x \in \Lambda \setminus \bigcup \mathcal{B}_\eta(\omega_j) \Longrightarrow \phi(\mathcal{G}(x)) > \phi(x) + \xi$$

SKETCH OF PROOF: First it will be proved by way of contradiction that the function

$$g(x) = \phi(\mathcal{G}(x)) - \phi(x)$$

is positive on $\Lambda \setminus \bigcup \mathcal{B}_\eta(\omega_j)$. Let $z_j \in \Lambda \setminus \bigcup \mathcal{B}_\eta(\omega_j)$ such that $g(z_j) \leq 0$. By compactness (and passing to a subsequence if necessary), assume that $z_j \to z \in \Lambda \setminus \bigcup \mathcal{B}_\eta(\omega_j)$. Now, by continuity,

$$\phi(\mathcal{G}(z)) \leq \phi(z)$$

which contradicts the claim that ϕ is a Lyapunov function. Hence the continuous function $g(z)$ is positive on the compact set

$$\Lambda \setminus \bigcup \mathcal{B}_\eta(\omega_j)$$

Choosing positive ξ less than the minimum of g over this set finishes the proof. □

LEMMA 5.3: *Let $\mathcal{G}$ be normal and let $\mathcal{T}$ be the stable manifold of the fixed point ω. There exists an open neighborhood U of ω such that $\mathcal{T} \cap U = \overline{\mathcal{T}} \cap U$.*

SKETCH OF PROOF: If not, let $\mathcal{T}$ correspond to the fixed point ω with basin $\mathcal{S}$, let $\overline{\mathcal{T}} = \mathcal{S}_i$, and suppose $x \in (\overline{\mathcal{T}} \cap U) \setminus \mathcal{T}$. Let $y_i = x$ and note that for some $\varepsilon > 0$,

$$\phi(y_i) < \phi(\lim_{k \to \infty} \mathcal{G}^k(y_i)) - \varepsilon$$

since ϕ is a Lyapunov function for $\mathcal{G}$. Since $\mathcal{G}^{\infty}(y_i)$ is in the closure of $\mathcal{S}_{i-1}$, there exists $y_{i-1} \in \mathcal{S}_{i-1}$ such that

$$\phi(y_i) < \phi(y_{i-1}) - \varepsilon$$

Repeating this construction yields

$$\phi(y_{i-1}) < \cdots < \phi(y_0)$$

It follows that $\phi(y_i) < \phi(\omega) - \varepsilon$, which leads to the contradiction $\phi(\omega) < \phi(\omega)$ as U contracts to ω, provided that ε can be bounded from below. A positive lower bound follows from Lemma 5.2 (since $x \notin \mathcal{T}$, the stable manifold theorem indicates that the orbit of x leaves a neighborhood of ω). □

THEOREM 5.4: *Suppose $\mathcal{G}$ is ergodic and normal with unstable fixed point ω, and let π be the steady-state distribution of random heuristic search. Then there exists a neighborhood U of ω such that $\pi(U) \to 0$ as $r \to \infty$.*

SKETCH OF PROOF: Let $\mathcal{H}_s$ be the subspace corresponding to those eigenvalues of $d\mathcal{G}_\omega$ (abbreviated $d\mathcal{G}$) within the unit disk, let $\mathcal{H}_u$ be the subspace corresponding to those eigenvalues outside the unit disk, and let ρ_s and ρ_u be the corresponding projections. Let the respective unit balls of these spaces be $\mathcal{B}_s$ and $\mathcal{B}_u$. By assumption, $d\mathcal{G}$ has no eigenvalues of modulus 1, and is hence a contraction on $\mathcal{H}_s$ with Lipschitz constant $\alpha < 1$ and an expansion on $\mathcal{H}_u$ with constant $\beta > 1$ with respect to suitable Euclidean norms $\|\cdot\|_s$ and $\|\cdot\|_u$ respectively. Given a vector v, the norm that will be used is $\|v\| = (\|\rho_s v\|_s^2 + \|\rho_u v\|_u^2)^{1/2}$. Note that the distance between the disjoint closed sets $\mathcal{H}_s$ and the boundary of $\mathcal{B}_u$ is 1. Let $\varepsilon > 0$ be sufficiently small that the set $\mathcal{L} + \omega$ where

$$\mathcal{L} = \mathcal{L}_u \times \mathcal{L}_s = \varepsilon\, \mathcal{B}_u \times \frac{\varepsilon}{1-\alpha} \mathcal{B}_s$$

is well within the "linear region" where the stable manifold $\mathcal{T}$ at ω is well approximated by $\omega + \mathcal{H}_s$ and $\mathcal{G}(\omega + z)$ is well approximated by $\omega + d\mathcal{G}z$ (for notational convenience and to simplify exposition, they will be treated as equal). Since $\mathcal{G}$ is normal, take the linear region to be so small that $\mathcal{T} = \overline{\mathcal{T}}$ within it (Lemma 5.3). Define, for $\varepsilon > 0$,

$$\begin{aligned} S_\varepsilon &= \{x\colon \|x - \omega\| < \varepsilon\} \\ \mathcal{T}_\varepsilon &= \{x\colon \|x - \mathcal{T}\| < \varepsilon\} \\ \overline{\mathcal{T}}_\varepsilon &= \{x\colon \|x - \overline{\mathcal{T}}\| < \varepsilon\} \end{aligned}$$

The proof is based on the idea that if the Markov chain is at state $x \in S_\varepsilon$, it will quickly move to the complement of $\overline{\mathcal{T}}$, whereupon it will spend almost all its time (as $r \to \infty$) at some distance η away from ω. Hence $\pi(S_{\min\{\eta,\varepsilon\}}) \to 0$ as $r \to \infty$, and so $\pi(\omega) = 0$.

The transition from $z + \omega$ to $z' + \omega$ can be viewed as a Bernoulli trial where *success* means:

1. $\|z' - d\mathcal{G}z\| < \gamma$.
2. $z' - d\mathcal{G}z$ has a component in the direction of $\rho_u d\mathcal{G}z$.

Transitions satisfying the first condition are called *semisuccessful*. Note that the probability of a semisuccessful transition is at least $1 - \exp\{-\gamma^2 r \theta_1\}$ for some $\theta_1 > 0$ (Theorem 3.3).

If $z + \omega \in S_\varepsilon$ and a series of semisuccessful transitions takes place, the resulting state $z^* + \omega$ satisfies $\rho_s z^* \in \mathcal{L}_s$ provided the transitions are within the linear region and $\gamma < \varepsilon$. This follows from the observation $\rho_s(S_\varepsilon - \omega) \subset \mathcal{L}_s$ and the transition inequality

$$\|\rho_s z'\| \leq \|\rho_s d\mathcal{G}z\| + \gamma < \alpha\|\rho_s z\| + \varepsilon \leq \frac{\varepsilon}{1-\alpha}$$

which is valid for any z satisfying $\rho_s z \in \mathcal{L}_s$. The probability of $e^{\gamma r \theta_2}$ consecutive semisuccessful transitions is $1 - o(1)$ as $r \to \infty$ since

$$e^{\gamma r \theta_2} \ln(1 - e^{-\gamma^2 r \theta_1}) \geq \frac{e^{\gamma r(\theta_2 - \gamma\theta_1)}}{e^{-\gamma^2 r \theta_1} - 1} = o(1)$$

provided that $\theta_2 < \gamma\theta_1$.

The second condition defining a successful transition says that the asymptotically normally distributed random vector (Theorem 3.6) whose addition to $d\mathcal{G}z$ produces z' should lie in a particular half space. Since the normal density is invariant under central inversion, the probability of satisfying the second condition is $1/2 + o(1)$. The probability of a subseries of k consecutive successful transitions out of a series of e^k semisuccessful transitions is therefore at least

$$1 - (1 - (2 + o(1))^{-k})^{e^k/k} \geq 1 - \exp\left\{-e^k(2 + o(1))^{-k}/k\right\} = 1 - o(1)$$

provided that $k \to \infty$. Note that if a series of k successful transitions takes place, the resulting state $z^* + \omega$ satisfies $\|\rho_u z^*\| \geq \beta^k\|\rho_u z\|$, since the second condition defining a successful transition guarantees that the expected motion away from ω (the factor of β provided by $d\mathcal{G}$) takes place at each transition.

Let $V = \{x\colon \|\rho_u x\| \leq 1 + \|d\mathcal{G}\|\}$. Using the normal approximation (Theorem 3.6) yields

$$\lim_{r \to \infty} prob\{\tau(\xi) \in \mathcal{G}(\xi) + r^{-1}V\} = 0$$

If $\xi \in S_\varepsilon \cap \mathcal{T}_{1/r}$, then, since ω is a fixed point of $\mathcal{G}$, choosing ε sufficiently small guarantees that $\mathcal{G}(\xi)$ is well within the "linear region," and hence, by using the normal approximation once more, $\tau(\xi)$ is well within the linear region with probability $1 - o(1)$ as $r \to \infty$. Thus it is permissible to regard $\mathcal{T}_{1/r} - \omega$ as $(\mathcal{H}_s)_{1/r}$ in the calculation below (the difference between the sets is only relevant outside the linear region and the impact on the probability in question is $o(1)$).

$$\begin{aligned} prob\{\tau(\xi) \in \mathcal{T}_{1/r}\} &= prob\{\tau(\xi) \in \mathcal{G}(\xi) + (\mathcal{T}_{1/r} - \omega) - d\mathcal{G}(\xi - \omega)\} \\ &\leq prob\{\tau(\xi) \in \mathcal{G}(\xi) + \{x\colon \|\rho_u x\| \leq (1 + \|d\mathcal{G}\|)/r\}\} \\ &= prob\{\tau(\xi) \in \mathcal{G}(\xi) + r^{-1}V\} \end{aligned}$$

Hence if $z + \omega \in S_\varepsilon$ then with probability $1 - o(1)$ either $\|\rho_u z\| \geq 1/r$ or $\|\rho_u z'\| \geq 1/r$. Therefore a series of at most $k = \theta_3 \ln r$ successful transitions beginning from a point

$z + \omega \in S_\varepsilon$ will with probability $1 - o(1)$ produce a state $z^* + \omega$ satisfying

$$\|\rho_u z^*\| \geq \frac{\beta^{\theta_3 \ln r}}{r} = \exp\{\theta_3 \ln r \ln \beta - \ln r\} > \varepsilon$$

for some $\theta_3 > 0$. In other words, z^* is at least ε away from $\mathcal{H}_s$. Taking $z^* + \omega$ to be the first such state encountered puts the transitions within the linear region and so $\rho_s z^* \in \mathcal{L}_s$. Hence $z^* + \omega \notin \mathcal{T}_\varepsilon$.

To summarize: If $z + \omega \in S_\varepsilon$, with probability $1 - o(1)$ the next $e^{\gamma r \theta_2}$ transitions are semisuccessful, and a state $z^* + \omega$ in the complement of $\overline{\mathcal{T}}_\varepsilon$ is encountered in fewer than r^{θ_3} transitions (provided θ_2 is sufficiently small and θ_3 is sufficiently large). Next, γ will be chosen making it impossible for semisuccessful transitions from $z^* + \omega$ to enter $\overline{\mathcal{T}}_\eta$ (for some $\eta > 0$). This implies that with probability $1 - o(1)$ the proportion of time the Markov chain can spend in $S_{\min\{\eta,\varepsilon\}}$ is $o(1)$.

Since the set of fixed points in the complement of $\overline{\mathcal{T}}$ is closed, let U be a closed neighborhood that separates it from $\overline{\mathcal{T}}$. For $k \geq 0$, define the sets

$$E^k = \{x\colon\ k \geq 0 \text{ is minimal such that } j \geq k \Rightarrow \mathcal{G}^j(x) \in U\}$$

It is easily verified that $\mathcal{G}(E^{k+1}) \subset E^k$ and that the sets

$$C^k = \bigcup_{j \leq k} E^j$$

are compact and satisfy $\mathcal{G}(C^{k+1}) \subset C^k$. Define sets F^k inductively as follows: For $x \in C^0$, let O_x^0 be a closed neighborhood separating x from $\overline{\mathcal{T}}$. Since C^0 is compact, let F_0 be the union of a finite subcover of the O_x^0 (their interiors form an open cover of C^0). For $x \in C^j$, let O_x^j be a closed neighborhood separating x from $\overline{\mathcal{T}}$ such that $\mathcal{G}(O_x^j)$ is contained in the interior of some O_y^{j-1} belonging to the finite subcover of C^{j-1}. Since C^j is compact, let F_j be the union of a finite subcover (of the O_x^j). By construction,

- The F^k are closed sets in the complement of $\overline{\mathcal{T}}$.
- $\mathcal{G}(F^{k+1})$ is contained in the interior of F^k.
- The interiors of the F^k cover the complement of $\overline{\mathcal{T}}$.

Because $\mathcal{G}$ is continuous, $\mathcal{G}(F^0)$ is compact and disjoint from $\overline{\mathcal{T}}$ (recall that orbits of points not in $\overline{\mathcal{T}}$ are disjoint from $\overline{\mathcal{T}}$). Therefore let ε be sufficiently small that $\mathcal{G}(F^0)$ is contained in the complement of $\overline{\mathcal{T}}_{2\varepsilon}$. Since the complement of $\overline{\mathcal{T}}_\varepsilon$ is compact, let the interiors of $F^0, F^1, \ldots, F^K$ be a finite cover. Let η be the distance from the closure of this cover to $\overline{\mathcal{T}}$. Note that the complement of the interior of F^k is disjoint from $\mathcal{G}(F^{k+1})$, hence let δ_k be the distance between them. Choose positive γ less than the minimum of $\varepsilon, \delta_0, \delta_1, \ldots \delta_{K-1}$. To summarize:

- If x is in the complement of $\overline{\mathcal{T}}_\varepsilon$, then $x \in F^k$ for some $k \leq K$.
- If $x \in F^k$ for some $k \leq K$, then x is in the complement of $\overline{\mathcal{T}}_\eta$.

- $\mathcal{G}(F^0)$ is contained in the cover $F^0, F^1, \ldots, F^K$ at a distance of at least γ from its complement.
- $\mathcal{G}(F^{k+1})$ is contained in F^k at a distance of at least γ from its complement.

It follows that a series of semisuccessful transitions from within the cover $F^0, F^1, \ldots, F^K$ cannot escape it, and must therefore remain a distance of η from $\overline{\mathcal{T}}$. □

The following corollary puts the result of Theorem 5.4 in a form which will be useful. Recall that the ω_j are the stable fixed points of $\mathcal{G}$.

COROLLARY 5.5: *Under the same hypotheses as Theorem 5.4, there exists a function* $\varepsilon = o(1)$ *such that*

$$\pi\,(\bigcup_k \mathcal{B}_\varepsilon(\omega_k)) = 1 - o(1)$$

as $r \to \infty$. *Moreover,* ε *may be chosen to converge arbitrarily slowly.*

SKETCH OF PROOF: Let $\lambda > 0$. Because the time spent away from stable fixed points vanishes (Theorem 5.4), there exists n_λ such that if $r \geq n_\lambda$ then

$$\pi\left(\bigcup_k \mathcal{B}_\lambda(\omega_k)\right) > 1 - \lambda$$

Now let λ decrease to 0 and take ε to decrease no faster than would the step function defined by $\varepsilon(n_\lambda) = \lambda$. □

6. Asymptotic Approximation

The main result of Nix and Vose (1991) is that as $r \to \infty$, the steady-state distribution of the finite-population model can give nonvanishing probability only to fixed points of $\mathcal{G}$. Since the time spent by random heuristic search away from a fixed point is negligible, steady-state behavior for large populations is therefore captured by a Markov chain—it will turn out to be C_r—having fixed points as states.

Although the results described in the preceding paragraph were proved in the context of a simple genetic algorithm, there was nothing crucial about the specific details of the heuristic function $\mathcal{G}$. The same conclusions hold (by the same proofs) for random heuristic search when hyperbolic and ergodic.

This section presents the main result of the paper. Let F denote the set of fixed points of $\mathcal{G}$. The following lemmas will be useful.

LEMMA 6.1: *Let* ρ *be a length-k path that is constrained to lie a distance of* $\varepsilon > 0$ *away from F. If* $\mathcal{G}$ *is focused, then there exist positive constants* η *and* δ *(depending on* ε*) such that for all* $k > 0$

$$|\rho| \geq \left\lfloor \frac{k}{\eta} \right\rfloor \delta$$

SKETCH OF PROOF: For $\xi \in \Lambda$ let n_ξ be such that $\mathcal{G}^{n_\xi}(\xi)$ is within ε of a fixed point. By continuity, let N_ξ be a neighborhood of ξ such that $\mathcal{G}^{n_\xi}(N_\xi)$ is also within ε of the fixed point. For parameter $\eta > 0$, define the sequence of sets

$$\begin{aligned} V_0 &= \eta(N_\xi - \xi) + \xi \\ V_{j+1} &= \mathcal{B}_\eta(\mathcal{G}(V_j)) \end{aligned}$$

Now choose $\eta = \eta_\xi$ sufficiently small such that for all $0 \le i \le n_\xi$,

$$V_i \subset \mathcal{G}^i(N_\xi)$$

It follows that if a path ρ of length n_ξ beginning from V_0 does not have a step x_{i-1} to x_i for which $\|x_i - \mathcal{G}(x_{i-1})\| > \eta_\xi$, ρ cannot stay ε away from F. On the other hand, if ρ does contain such a step, $|\rho| > \theta\eta_\xi^2$ for some $\theta > 0$ (Lemma 3.2).

The existence of a finite subcover of the V_0 (by compactness of Λ) implies the existence of η (the maximum over n_ξ) and δ (the minimum over $\theta\eta_\xi^2$) such that every path of length η maintaining a distance of ε away from F has a cost of at least δ. □

LEMMA 6.2: *Let ω and ω' be distinct fixed points of $\mathcal{G}$ (not necessarily stable) and suppose positive ε is $o(1)$ as $r \to \infty$. Then there exists a path ρ' of length K from ω to ω' such that*

1. *$K \to \infty$ as $r \to \infty$ (ρ' depends on r).*
2. *The interior points of ρ' come from X_n^r and are at least ε away from F.*
3. *$|\rho'| = \rho_{\omega,\omega'} + o(1)$.*

Moreover, K may be chosen to diverge arbitrarily slowly.

SKETCH OF PROOF: Let ρ_j be a sequence of paths from ω to ω' such that

$$\lim_{j \to \infty} |\rho_j| = \rho_{\omega,\omega'}$$

Without loss of generality, the length of ρ_j diverges since it can be made longer at an arbitrarily small cost: To increase the length of $\omega, x_1, \ldots, \omega'$ modify it to $\omega, z, x_1, \ldots, \omega'$. The cost of the modified path is

$$\alpha_{\omega,z} + \alpha_{z,x_1} + \alpha_{x_1,x_2} + \cdots \to 0 + \alpha_{\omega,x_1} + \alpha_{x_1,x_2} + \cdots$$

as $z \to \omega$. In particular, it may be assumed that ρ_j has length j. Also, by perturbing the interior points of ρ_j as necessary, it may be assumed that no interior point is a fixed point.

As $r \to \infty$, the set X_n^r becomes dense in Λ and the interior points of ρ_j become at least ε away from F (since $\varepsilon = o(1)$). Thus if r is sufficiently large, ρ_j can be approximated arbitrarily closely by a path ρ' of the required type such that

$$|\rho'| < |\rho_j| + 1/j$$

Moreover, r may be arbitrarily large with respect to j. □

THEOREM 6.3: *Suppose $\mathcal{G}$ is ergodic and normal. As $r \to \infty$, the probability measure corresponding to the steady-state distribution of random heuristic search converges to point mass at the fixed point corresponding to the minimum cost tributary of the fixed-point graph (provided it exists).*

SKETCH OF PROOF: Consider the auxiliary Markov chain having as "states" the neighborhoods $\mathcal{B}_\varepsilon(\omega_k)$ together with $\mathcal{B}_\varepsilon(\omega'_l)$ where the ω'_l are unstable fixed points. Since ε will eventually be chosen according to Corollary 5.5, a state is an open ball of radius $o(1)$ as $r \to \infty$. Let S_i be the general name for the ith such state, and to streamline notation, let ω_j denote a general fixed point (possibly unstable). According to Corollary 5.5, the time spent away from these states vanishes as $r \to \infty$, and hence the steady-state behavior of RHS converges to the steady-state behavior of this auxiliary chain, provided that its transition probabilities can be appropriately defined. The auxiliary chain is obtained sequentially in two steps:

1. Aggregate the points (states) of RHS belonging to the sets S_i while leaving all other points alone.
2. Restrict to the subchain over the states S_i.

To streamline notation, the transition probability from state S_i to state S_j is denoted by $A_{i,j}$ and ρ_{ω_i,ω_j} is abbreviated by $\rho_{i,j}$. In accordance with the aggregation and restriction theorems for Markov chains (see Appendix), the transition probability $A_{i,j}$ of the auxiliary chain is defined as

$$
\begin{aligned}
A_{i,j} \quad = \quad & \frac{1}{\pi(S_i)} \sum_{x \in S_i} \pi_x \sum_{y \in S_j} Q_{x,y} \\
& + \sum_{z_1}{}^{*} \frac{1}{\pi(S_i)} \sum_{x \in S_i} \pi_x \, Q_{x,z_1} \sum_{y \in S_j} Q_{z_1,y} \\
& + \sum_{z_1,z_2}{}^{*} \frac{1}{\pi(S_i)} \sum_{x \in S_i} \pi_x \, Q_{x,z_1} \, Q_{z_1,z_2} \sum_{y \in S_j} Q_{z_2,y} \\
& + \ldots
\end{aligned}
$$

where $\sum^*$ denotes summation over the complement of the union of the S_i, and π is the steady-state distribution of RHS. By Lemma 3.1, the first term is

$$
\begin{aligned}
\sum_{y \in S_j} \frac{1}{\pi(S_i)} \sum_{x \in S_i} \pi_x \, e^{-r(\alpha_{\omega_i,\omega_j} + \alpha_{x,y} - \alpha_{\omega_i,\omega_j} + O(r^{-1} \ln r))} \quad & = \quad \sum_{y \in S_j} e^{-r\,\alpha_{\omega_i,\omega_j}} \frac{1}{\pi(S_i)} \sum_{x \in S_i} \pi_x \, e^{-r\,o(1)} \\
& = \quad \sum_{y \in S_j} e^{-r(\alpha_{\omega_i,\omega_j} + o(1))}
\end{aligned}
$$

since $S_i \to \{\omega_i\}$ and $S_j \to \{\omega_j\}$ and hence $\alpha_{x,y} \to \alpha_{\omega_i,\omega_j}$. Similarly, the kth term ($k > 1$) is

$$
\sum_{y \in S_j} t_k \qquad \text{where} \qquad t_k = \sum_{z_1,\ldots,z_{k-1}}{}^{*} \; e^{-r(\alpha_{\omega_i,z_1} + \cdots + \alpha_{z_{k-1},\omega_j} + o(1)) + O(k \ln r)}
$$

The expression above is also valid for the first term ($k = 1$) if t_1 is interpreted as $e^{-r(\alpha_{\omega_i,\omega_j} + o(1))}$. Note that each t_k corresponds to a sum over length-k paths from ω_i to ω_j. The paths of t_k are constrained to follow along states of RHS—there are r^θ states[7]—and to avoid states of the auxiliary chain (i. e., their interior points must be at least ε from F). According to Lemma 6.2, a minimal cost path ρ_K of t_K has cost $\rho_{i,j} + o(1)$ as $K, r \to \infty$. Next, define c by

$$
\sum_{k > K} t_k = c \sum_{k \le K} t_k \qquad \text{so that} \qquad \sum_{k} t_k = (1 + c) \sum_{k \le K} t_k
$$

7 Here θ is used to represent a positive function of r for which both θ and θ^{-1} are bounded.

Since there are at most $r^{k\theta}$ paths of length k, it follows that

$$\begin{aligned} e^{-r(\rho_{i,j}+o(1))+O(K\ln r)} &\leq \sum_k t_k \\ &\leq (1+c)\sum_{k\leq K} r^{k\theta} e^{-r(\rho_{i,j}+o(1))+O(k\ln r)} \\ &\leq e^{-r(\rho_{i,j}+o(1))+\ln(1+c)+O(K\ln r)} \end{aligned}$$

Therefore,

$$\sum_k t_k = e^{-r(\rho_{i,j}+o(1))}$$

provided that $\ln(1+c) = o(r)$ and $K \to \infty$ at a rate $o(r/\ln r)$. This yields

$$\begin{aligned} A_{i,j} &= \sum_k \sum_{y\in S_j} t_k \\ &= \sum_{y\in S_j} e^{-r(\rho_{i,j}+o(1))} \\ &= e^{-r(\rho_{i,j}+o(1))} \end{aligned}$$

since the number of states in RHS is $r^{\theta} = e^{o(r)}$. Next, $\ln(1+c) = o(r)$ will be established. Let ρ_k be a minimal cost path of t_k. By choosing $K' > K = O(\ln r)$, which is justified since K diverges arbitrarily slowly,

$$\begin{aligned} c &= \sum_{k>K} t_k \Big/ \sum_{k\leq K} t_k \\ &\leq e^{r(\rho_{i,j}+o(1))} \sum_{k>K} r^{k\theta} e^{-r(|\rho_k|+o(1))+O(k\ln r)} \\ &\leq e^{r(\rho_{i,j}+o(1))} \left(\sum_{k\leq K'} e^{-r(\rho_{i,j}+o(1))+O(k\ln r)} + \sum_{k>K'} e^{-r(|\rho_k|+o(1))+O(k\ln r)} \right) \end{aligned}$$

Estimating the first sum in the second factor as before and applying Lemma 6.1 to the second sum produces

$$\begin{aligned} &\leq e^{r(\rho_{i,j}+o(1))} \left(e^{-r(\rho_{i,j}+o(1))+O(K'\ln r)} + \sum_{k>K'} e^{-r(k\eta^{-1}-1)\delta+o(r)+O(k\ln r)} \right) \\ &\leq e^{o(r)+O(K'\ln r)} + e^{r(\rho_{i,j}+o(1))} \sum_{k>K'} e^{-r(k\eta^{-1}-1)\delta+k\theta\ln r} \end{aligned}$$

for some θ (perhaps differing from previous use of θ). Choosing $K' = o(r/\ln r)$ and bounding the second sum with an integral yields

$$\begin{aligned} &\leq e^{o(r)} + e^{r(\rho_{i,j}+\delta+o(1))} \int_{K'}^{\infty} e^{k(\theta\ln r - r\eta^{-1}\delta)}\, dk \\ &\leq e^{o(r)} + e^{r(\rho_{i,j}+\delta+o(1))+K'(\theta\ln r - r\eta^{-1}\delta)} / (r\eta^{-1}\delta - \theta\ln r) \end{aligned}$$

provided $r\eta^{-1}\delta > \theta \ln r$. Since δ and η depend on ε, which is varying arbitrarily slowly with respect to r, without loss of generality $r\eta^{-1}\delta > r/\ln r$. Thus, for large r, the following bound is obtained:

$$c \leq e^{o(r)} \left(1 + e^{r\,(\rho_{i,j}+\delta-K'/\ln r)}\right)$$

Choosing $K' = \sqrt{r}$ gives $c = e^{o(r)}$ as required.

According to the restriction and aggregation theorems for Markov chains (see Appendix), the steady-state distribution π' of the auxiliary chain satisfies

$$\pi'(S_j) = \frac{\pi(S_j)}{\pi(\bigcup_j S_j)}$$

Applying Corollary 5.5 shows $\pi(S_j) = \pi'(S_j) + o(1)$ as $r \to \infty$. Therefore, the auxiliary chain correctly captures the asymptotic behavior of random heuristic search as $r \to \infty$. Moreover, it has been established that the auxiliary chain has transition matrix satisfying

$$A_{i,j} = e^{-r\,(\rho_{i,j}+o(1))}$$

It follows that the auxiliary chain is identical to the Markov chain C_r, with the exception that it contains states corresponding to unstable fixed points whereas C_r does not. The final step in the proof is to show that the steady-state distribution of the auxiliary chain converges to the steady-state distribution of C_r. This is accomplished via Theorem 4.1 by showing that the inclusion of states corresponding to unstable fixed points cannot affect the location where the minimal cost tributary is rooted.

Let t be a minimal cost tributary of the auxiliary chain, and to simplify exposition, nodes (of t) corresponding to stable/unstable fixed points will be referred to as stable/unstable nodes. Also, node S_i will be called "adjacent" to S_j if ω_i is in the closure of the basin of attraction of ω_j.

First consider the case where an unstable interior node x of t does not contain in its subtree every stable node to which it is adjacent. Let the children of x be $y_0, \ldots, y_k$, and let z be a node, to which x is adjacent, that is not in the subtree rooted at x. Modifying t by removing the subtrees rooted at y_i and reattaching them as children of z affects the cost of the tree by at most $o(r)$. Removing x and reattaching it as a child of z has similar cost. The resulting tree has one less unstable interior node; x is now a leaf and is adjacent to its parent.

In the case where x contains in its subtree every stable node to which it is adjacent, let z and y_i be as before, except that z is located in the subtree rooted at y_0. Modifying t by removing the subtrees rooted at y_i, for $i > 0$, and reattaching them as children of z affects the cost of the tree by at most $o(r)$. Next, remove the subtree rooted at y_0 and reattach it as a child of the parent of x, then remove x and reattach it as a child of z. The resulting tree has one less unstable interior node; x is now a leaf and is adjacent to its parent. As before, this is accomplished at a cost of $o(r)$.

The two operations described above may be repeated to transform t into a tree t' that, ignoring $o(r)$ terms, has identical cost. Moreover, t' has all unstable nodes as leaves adjacent to their stable parent nodes. Hence by removing all unstable leaves from t', a tree t'' is obtained that corresponds to C_r and that, neglecting $o(r)$ terms, has cost identical to t.

Conversely, if t'' is a minimal cost tributary corresponding to C_r, then by adjoining unstable nodes as leaves adjacent to their parents, a tree t corresponding to the auxiliary chain is obtained that, neglecting $o(r)$ terms, has cost identical to t''.

Since, as $r \to \infty$, the $o(r)$ terms have no influence, it follows that the states corresponding to unstable fixed points (i.e., the difference between the auxiliary chain and C_r) cannot affect the location where the minimal cost tributary is rooted. $\square$

7. Conclusion

A general class of search algorithms, *random heuristic search*, is reviewed, and it is indicated how a variety of generational style GAs are special cases of it. Abstractly, the message of this paper is summarized by this statement: Let Υ be a normal, ergodic, large population instance of random heuristic search. If the minimum cost tributary of $\Im$ is unique,

- Short-term behavior of Υ is determined by the local basin in which the initial population finds itself.
- Long-term behavior is determined by that local fixed point having the largest basin (i.e., minimum spanning intree).

To put this more concretely, let $\mathcal{G}$ be a heuristic that instantiates random heuristic search to a genetic algorithm. The above conclusions would apply based on the assumption of nonzero mutation and the conjectures that

1. $\mathcal{G}$ has a Lyapunov function.[8]
2. $\mathcal{G}$ is hyperbolic.[9]
3. The minimum cost tributary of $\Im$ is unique.

While these conjectures are interesting in themselves, the fact that this work has woven them into finite-population GA behavior will, it is hoped, encourage their consideration and the further development of the interrelationships between the finite- and infinite-population models.

Readers interested in further details concerning how random heuristic search instantiates to the genetic algorithm are referred to Vose (in press).

Acknowledgments

This research was supported by the National Science Foundation: IRI-8917545, IRI-9224917, and CDA-9115428.

References

Akin, E. (1993). *The general topology of dynamical systems.* Providence, RI: American Mathematical Society.

Davis, T. (1991). *Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm.* Ph. D. Dissertation, University of Florida, Gainesville.

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal, deceptive problem. In L. Davis (Ed.), *Genetic algorithms and simulated annealing* (pp. 74–88). London: Pitman.

Nix, A., & Vose, M. D. (1991). Modeling genetic algorithms with Markov chains. *Annals of Mathematics and Artificial Intelligence*, *5* (1992), 79–88.

Renyi, A. (1970). *Probability theory.* Amsterdam: Elsevier North-Holland.

8 This conjecture has been proved—assuming proportional selection—for fitness functions with low epistasis (Vose & Wright, 1994a).

9 This conjecture has recently been proved—assuming proportional selection—by Mary Eberlein and Michael D. Vose.

Vose, M. D. (1990). Formalizing genetic algorithms. *Proceedings of the IEEE Workshop on Genetic Algorithms, Neural Nets, and Simulated Annealing Applied to Problems in Signal and Image Processing*, May 1990. Glasgow, U. K.

Vose, M. D. (1993a). Modeling simple genetic algorithms. In D. Whitley (Ed.), *Foundations of genetic algorithms II*. San Mateo, CA: Morgan Kaufmann.

Vose, M. D. (1993b). *Computing steady state distributions without cancellation*. Technical Report (CS-93-189), Department of Computer Science, University of Tennessee, Knoxville.

Vose, M. D. (in press). *The simple genetic algorithm: Foundations and theory*.

Vose, M. D., & Wright, A. H. (1994a). Simple genetic algorithms with linear fitness. *Evolutionary Computation*, *2*(4), 347–368.

Wright, A. H., & Vose, M. D. (1994b). Finiteness of the fixed point set for the simple genetic algorithm. Manuscript submitted for publication.

Appendix

This section describes the results concerning aggregation and restriction of Markov chains that were used in this article.

Let Q be the positive transition matrix for a Markov chain over the states given by elements of the set S, and let the nonempty sets $P_1, \ldots, P_l$ partition S. Let π be the steady-state distribution corresponding to Q. Consider a Markov chain having as states the sets P_j and transition matrix Q'.

THEOREM (*Aggregation*) *If the transition matrices Q and Q' are related by*

$$Q'_{i,j} = \frac{1}{\pi(P_i)} \sum_{x \in P_i} \pi_x \sum_{y \in P_j} Q_{x,y}$$

then $\pi'_j = \pi(P_j)$.
PROOF: *Check that π' satisfies the steady-state equation $(Q')^T \pi' = \pi'$.* □

Next, consider a Markov chain having as states the elements of some nonempty subset S' of S and transition matrix Q'.

THEOREM (*Restriction*) *If the transition matrices Q and Q' are related by*

$$\begin{aligned} Q'_{i,j} &= Q_{i,j} \\ &+ \sum_{z_1}{}^{*} Q_{i,z_1} Q_{z_1,j} \\ &+ \sum_{z_1,z_2}{}^{*} Q_{i,z_1} Q_{z_1,z_2} Q_{z_2,j} \\ &+ \ldots \end{aligned}$$

where $\sum^*$ denotes summation over the complement of S', then $\pi'_j = \pi_j / \pi(S')$.
PROOF: *Induct on the cardinality of $S \setminus S'$.* □