Continuous Dynamical System Models of Steady-State Genetic Algorithms

Alden H. Wright
Computer Science Department
University of Montana
Missoula, MT 59812
USA
wright@cs.umt.edu

Jonathan E. Rowe *
School of Computer Science
University of Birmingham
Birmingham B15 2TT
Great Britain
J.E.Rowe@cs.bham.ac.uk

May 21, 2002

This paper appears in **Foundations of Genetic Algorithms 6**, edited by Worthy N. Martin and William M. Spears, Morgan Kaufmann, 2001, pp. 209–226.

Abstract

This paper constructs discrete-time and continuous-time dynamical system expected value and infinite population models for steady-state genetic and evolutionary search algorithms. Conditions are given under which the discrete-time expected value models converge to the continuous-time models as the population size goes to infinity. Existence and uniqueness theorems are proved for solutions of the continuous-time models. The fixed points of these models and their asymptotic stability are compared.

1 Introduction

There has been considerable development of expected value and infinite population models for genetic algorithms. To date, this work has concentrated on generational genetic algorithms. These models tend to be discrete-time dynamical systems, where each time step corresponds to one generation of the genetic algorithm.

Many practitioners (such as [Dav91]) advocate the use of steady-state genetic algorithms where a single individual is replaced at each step. This paper develops expected value and infinite population models for steady-state genetic algorithms. First, discrete-time expected value models are described, where each time step corresponds to the replacement of an individual. It is natural to consider these models in the limit when the population goes to infinity and the time step goes to zero. This paper shows how this limiting process leads in a natural way to a continuous-time dynamical system model. Conditions for the existence and uniqueness of solutions of this model are given.

The steady-state model that uses random deletion has a very close correspondence with the generational model that uses the same crossover, mutation, and selection. The fixed points of the two models are the same, and a fixed point where all of the eigenvalues of the differential of the generational model heuristic function have modulus less than one must be stable under the discrete-time and continuous-time steady-state models. However, a numerical example is given of a fixed point which is asymptotically stable under the continuous-time steady-state model but not asymptotically stable under the generational model.

* This work was completed while Jonathan E. Rowe was at De Montfort University.

Let Ω denote the search space for a search problem. We identify Ω with the integers in the range from 0 to $n-1$, where n is the cardinality of Ω. We assume a real-valued nonnegative fitness function f over Ω. We will denote $f(i)$ by f_i. Our objective is to model population-based search algorithms that search for elements of Ω with high fitness. Such algorithms can be *generational*, where a large proportion of the population is replaced at each time step (or generation). Or they can be *steady-state*, where only a single or small number of population members are replaced in a time step.

A population is a multiset (set with repeated elements) with elements drawn from Ω. We will represent populations over Ω by nonnegative vectors indexed over the integers in the interval $[0, n)$ whose sum is 1. If a population of size r is represented by a vector p, then rp_i is the number of copies of i in the population. For example, if $\Omega = \{0, 1, 2, 3\}$, and the population is the multiset $\{0, 0, 1, 2, 2\}$, then the population is represented by the vector $\langle \ 2/5 \quad 1/5 \quad 2/5 \quad 0 \ \rangle^T$.

Let $\Lambda = \{x \ : \ \sum_i x_i = 1 \text{ and } x_j \geq 0 \text{ for all } j\}$. Then all populations over Ω are elements of Λ. Λ can also be interpreted as the set of probability distributions over Ω. It is natural to think of elements of Λ as *infinite populations*. Geometrically, Λ is the unit simplex in $\Re^n$.

The ith unit vector in $\Re^n$ is denoted by e_i. The Euclidean norm on $\Re^n$ is denoted by $\| \cdot \| = \| \cdot \|_2$, the max norm by $\| \cdot \|_\infty$, and the sum norm by $\| \cdot \|_1$. The Euclidean norm is the default.

Brackets are used to denote an indicatior function. Thus,

$$[expression] = \begin{cases} 1 & \text{if } expression \text{ is true} \\ 0 & \text{if } expression \text{ is false} \end{cases}$$

Vose's random heuristic search algorithm describes a class of generational population-based search algorithms. The model is defined by a *heuristic function* $\mathcal{G} : \Lambda \to \Lambda$. If x is a population of size r, then the next generation population is obtained by taking r independent samples from the probability distribution $\mathcal{G}(x)$. When random heuristic search is used to model the simple genetic algorithm, $\mathcal{G}$ is the composition of a selection heuristic function $\mathcal{F} : \Lambda \to \Lambda$ and a mixing heuristic function $\mathcal{M} : \Lambda \to \Lambda$. The mixing function describes the properties of crossover and mutation. Properties of the $\mathcal{M}$ and $\mathcal{F}$ functions are explored in detail in [Vos99].

Given a population $x \in \Lambda$, it is not hard to show that the expected next generation population is $\mathcal{G}(x)$. As the population size goes to infinity, the next generation population converges in probability to its expectation, so it is natural to use $\mathcal{G}$ to define an infinite population model. Thus, $x \longrightarrow \mathcal{G}(x)$ defines a discrete-time dynamical system on Λ that we will call the *generational model*. Given an initial population x, the trajectory of this population is the sequence $x, \mathcal{G}(x), \mathcal{G}^2(x), \mathcal{G}^3(x), \ldots$.

Note that after the first step, the populations produced by this model do not necessarily correspond to populations of size r.

2 Steady-state evolutionary computation algorithms

Whitley's Genitor algorithm [Whi89] was the first "steady state" genetic algorithm. Genitor selects two parent individuals by ranking selection and applies mixing to them to produce one offspring, which replaces the worst element of the population. Syswerda ([Sys89] and [Sys91]) described variations of the steady-state genetic algorithm and empirically compared various deletion methods. Davis [Dav91] also empirically tested steady-state genetic algorithms and advocates them as being superior to generational GAs when combined with a feature that eliminates duplicate chromosomes.

In this section, we describe two versions of steady-state search algorithms. Both algorithms start with a population η of size r. In most applications, this population would be chosen randomly from the search space, but there is no requirement for a random initial population. At each step of both algorithms, an element j is removed from the population, and an element i of Ω is added to the population, The selection

of the element i is described by a heuristic function $\mathcal{G}$. (For a genetic algorithm, $\mathcal{G}$ will describe crossover, mutation, and usually selection.) The selection of element j is described by another heuristic function $\mathcal{D}_r$. (We include the population size r as a subscript since there may be a dependence on population size.)

In the first algorithm, the heuristic functions $\mathcal{G}$ and $\mathcal{D}_r$ both depend on x, the current population. Thus, i is selected from the probability distribution $\mathcal{G}(x)$, and j is selected from the probability distribution $\mathcal{D}_r(x)$.

Steady-state random heuristic search algorithm 1:

1 Choose an initial population η of size r
2 $x \leftarrow \eta$
3 Select i from Ω using the probability distribution $\mathcal{G}(x)$.
4 Select j using the probability distribution $\mathcal{D}_r(x)$.
5 Replace x by $x - e_j/r + e_i/r$.
6 Go to step 3.

The second algorithm differs from the first by allowing for the possibility that the newly added element i might be deleted. Thus, j is selected from the probability distribution $\mathcal{D}_r(\frac{rx+e_i}{r+1})$. This algorithm is an $(r+1)$ algorithm in evolution strategy notation.

Steady-state random heuristic search algorithm 2:

1 Choose an initial population η of size r.
2 $x \leftarrow \eta$
3 Select i from Ω using the probability distribution $\mathcal{G}(x)$.
4′ Select j using the probability distribution $\mathcal{D}_r(\frac{rx+e_i}{r+1})$.
5 Replace x by $x - e_j/r + e_i/r$.
6 Go to step 3.

Some heuristics that have been suggested for for the $\mathcal{D}_r$ function include worst-element deletion, where a population element with the least fitness is chosen for deletion, reverse proportional selection, reverse ranking deletion, and random deletion, where the element to be deleted is chosen randomly from the population. Random deletion was suggested by Syswerda [Sys89]. He points out that random deletion is seldom used in practice. Because of this, one of the reviewers of this paper objected to the use of the term "steady-state genetic algorithm" for an algorithm that used random deletion. However, we feel that the term can be applied to any genetic algorithm that replaces only a few members of the population during a time step of the algorithm.

Random deletion can be modeled by choosing $\mathcal{D}_r(x) = x$.

If the fitness function is injective (the fitnesses of elements of Ω are distinct), then reverse ranking and worst-element deletion can be modeled using the framework developed for ranking selection in [Vos99].

$$\mathcal{D}_r(x)_i = \int_{\sum[f_j<f_i]x_j}^{\sum[f_j\le f_i]x_j} \rho(s)ds$$

The probability density function $\rho(s)$ can be chosen to be $2s$ to model standard ranking selection, and $2-2s$ to model reverse ranking deletion. To model worst-element deletion, we define $\rho(s)$ as follows:

$$\rho(s) = \begin{cases} r & \text{if } 0 \le s \le 1/r, \\ 0 & \text{otherwise.} \end{cases}$$

As an example, let $n = 3$, $x = \langle\ \frac{1}{3}\quad \frac{1}{6}\quad \frac{1}{2}\ \rangle^T$, $f = \langle\ 2\quad 1\quad 3\ \rangle^T$, and $r = 4$. Then $\rho(s) = 4$ if $0 \le s \le 1/4$ and $\rho(s) = 0$ if $1/4 < s \le 1$. (The population x does not correspond to a real finite population of size 4.

However, this choice leads to a more illustrative example. Also, if $\mathcal{D}_r$ is iterated, after the first iteration the populations produced will not necessarily correspond to finite populations of size r.) Then

$$\mathcal{D}_r(x)_1 = \int_0^{x_1} \rho(s)ds = \int_0^{1/6} 4ds = 2/3,$$

$$\mathcal{D}_r(x)_0 = \int_{x_1}^{x_1+x_0} \rho(s)ds = \int_{1/6}^{1/2} \rho(s)ds = \int_{1/6}^{1/4} 4ds = 1/3.$$

and

$$\mathcal{D}_r(x)_2 = \int_{x_1+x_0}^{x_1+x_0+x_2} \rho(s)ds = \int_{1/2}^{1} \rho(s)ds = 0.$$

For random deletion and reverse ranking deletion, $\mathcal{D}_r(x)$ does not depend on the population size and can be shown to be differentiable as a function of x.

For worst-element deletion, $\mathcal{D}_r(x)$ does depend on the population size, and is continuous but not differentiable.

Lemma 2.1 *If $\mathcal{D}_r$ is defined as above for worst-element deletion, then $\mathcal{D}_r$ satisfies a Lipschitz condition. In other words, there is a constant L_r so that $\|\mathcal{D}_r(x) - \mathcal{D}_r(y)\| \leq L_r\|x - y\|$ for all $x, y \in \Lambda$.*

Proof. Let $x, y \in \Lambda$. Then

$$\begin{aligned}
|\mathcal{D}_r(x)_i - \mathcal{D}_r(y)_i| &= \left| \int_{\sum[f_j<f_i]x_j}^{\sum[f_j\leq f_i]x_j} \rho(s)ds - \int_{\sum[f_j<f_i]y_j}^{\sum[f_j\leq f_i]y_j} \rho(s)ds \right| \\
&= \left| \int_{\sum[f_j<f_i]y_j}^{\sum[f_j<f_i]y_j} \rho(s)ds - \int_{\sum[f_j\leq f_i]x_j}^{\sum[f_j\leq f_i]y_j} \rho(s)ds \right| \\
&\leq r\sum[f_j < f_i]|y_j - x_j| + r\sum[f_j \leq f_i]|y_j - x_j| \\
&\leq 2r\sum|y_j - x_j| \\
&= 2r\|x - y\|_1.
\end{aligned}$$

Thus, $\|\mathcal{D}_r(x)-\mathcal{D}_r(y)\|_\infty \leq 2r\|x-y\|_1$. Since all norms are equivalent up to a constant, $\|\mathcal{D}_r(x)-\mathcal{D}_r(y)\|_2 \leq 2rK\|x - y\|_2$ for some constant K. □

The function

$$\mathcal{H}_r(x) = x + \frac{1}{r}\mathcal{G}(x) - \frac{1}{r}\mathcal{D}_r(x) \tag{1}$$

gives the expected population for Algorithm 1 at the next time step, and the function

$$\mathcal{K}_r(x) = x + \frac{1}{r}\mathcal{G}(x) - \frac{1}{r}\mathcal{D}_{r+1}\left(\frac{rx + \mathcal{G}(x)}{r+1}\right) \tag{2}$$

gives the expected population for Algorithm 2 at the next time step.

Thus, $x \longrightarrow \mathcal{H}_r(x)$ and $x \longrightarrow \mathcal{K}_r(x)$ define discrete-time expected-value models of the above steady-state algorithms. We will call these the *discrete-time steady-state models*.

The following is straightforward.

Lemma 2.2 *If the deletion heuristic $\mathcal{D}_r$ of the discrete-time steady-state models satisfy*

$$\mathcal{D}_r(y) \leq rx + \mathcal{G}(x), \tag{3}$$

where $y = x$ for (1) and $y = \frac{rx+\mathcal{G}(x)}{r+1}$ for (2), then the trajectories of the systems defined by $\mathcal{H}_r$ and $\mathcal{K}_r$ remain in the simplex Λ.

The models for random deletion, reverse ranking deletion, and worst-element deletion all satisfy the hypotheses of lemma 2.2.

3 Convergence of the $\mathcal{K}_r$ heuristic.

In this section we assume that the deletion heuristic is defined by worst element deletion. We give conditions on the fitness function and on $\mathcal{G}$ that assure that $\lim_{t\to\infty} \mathcal{K}_r^t(x)$ exists and is the uniform population consisting of copies of the global optimum.

In evolution strategy terminology, this is an $(r+1)$-ES algorithm which uses an elitist selection method. Rudolph [Rud98] has shown that for this class of algorithms, if there is mutation rate that is greater than zero and less than one, then the finite population algorithm converges completely and in mean. These are statements about the best element in the population rather than the whole population, so these results do not imply our result.

We assume that the fitness function is injective. In other words, we assume that if $i \neq j$, then $f_i \neq f_j$. Since we will not be concerned with the internal structure of Ω, without loss of generality we can assume that $f_0 < f_1 < \ldots < f_{n-1}$. This assumption will simplify notation.

Under this assumption, we can give a simplified definition for the worst-element deletion heuristic $\mathcal{D}_{r+1}$ that is used in the definition of $\mathcal{K}_r$.

$$\mathcal{D}_{r+1}(y)_i = \begin{cases} (r+1)y_i & \text{if } \sum_{j\le i} y_j \le \frac{1}{r+1} \\ 1-(r+1)\sum_{j<i} y_j & \text{if } \sum_{j<i} y_j \le \frac{1}{r+1} < \sum_{j\le i} y_j \\ 0 & \text{if } \frac{1}{r+1} < \sum_{j<i} y_j \end{cases}$$

Now define $m(x) = \min\{i : x_i > 0\}$, and define $M(x) = 2m(x) + 1 - x_{m(x)}$.

Theorem 3.1 *If $f_0 < f_1 < \ldots < f_{n-1}$ and if there is a $\delta > 0$ such that for all $x \in \Lambda$ such that $x \neq e_{n-1}$,*

$$\sum_{j>m(x)} \mathcal{G}_j(x) \ge \delta,$$

then for any $x \in \Lambda$ there is a $T > 0$ such that $\mathcal{K}_r^t(x) = e_{n-1}$ for all $t \ge T$.

This condition says that $\mathcal{G}(x)$ has a combined weight of at least δ on those points of Ω whose fitness is higher than the worst-fitness element of x. (By "element of x", we mean any $i \in \Omega$ such that $x_i > 0$.) This condition would be satisfied by any $\mathcal{G}$ heuristic that allowed for a positive probability of mutation between any elements of Ω.

To prove this theorem, we need the following results.

Lemma 3.2 *For any $x \in \Lambda$, if $j < m(x)$, then $\mathcal{K}_r(x)_j = 0$.*

Proof. To simplify notation, let m denote $m(x)$.

Let $y = \frac{rx+\mathcal{G}(x)}{r+1}$. Then $\sum_{j<m} \mathcal{G}(x)_j \le \frac{1}{r+1}$ since $\sum_{j<m} x_j = 0$ and $\sum_{j<m} \mathcal{G}_j \le 1$.
Thus, for $j < m$, $\mathcal{D}_{r+1}(y)_j = y_j$, and $\mathcal{K}_r(x)_j = y_j - \mathcal{D}_{r+1}(y)_j = 0$. □

Lemma 3.3 *For any $x \in \Lambda$, if there is a $\delta > 0$ such that $\sum_{j>m(x)} \mathcal{G}(x)_j \ge \delta$, then*

$$M(\mathcal{K}_r(x)) \ge M(x) + \frac{\delta}{r}.$$

Proof. To simplify notation, again let m denote $m(x)$.

Let $y = \frac{rx+\mathcal{G}(x)}{r+1}$.

Case 1: $\sum_{j\le m} y_j \le \frac{1}{r+1}$.

Then

$$\mathcal{D}_{r+1}(y)_m = (r+1)y_m = rx_m + \mathcal{G}(x)_m,$$

and

$$\mathcal{K}_r(x)_m = x_m + \frac{1}{r}\mathcal{G}(x)_m - \frac{1}{r}(rx_m + \mathcal{G}(x)_m) = 0.$$

Thus,

$$M(\mathcal{K}_r(x)) \ge 2(m+1) + 1 - x_{m+1} \ge 2m + 2 \ge M(x) + 1.$$

Case 2: $\sum_{j<m} y_j \le \frac{1}{r+1} < \sum_{j\le m} y_j$.

Then

$$\mathcal{D}_{r+1}(y)_m = 1 - (r+1)\sum_{j<m} y_j = 1 - \sum_{j<m} \mathcal{G}(x)_j.$$

Thus,

$$\begin{aligned}
\mathcal{K}_r(x)_m &= x_m + \frac{1}{r}\mathcal{G}(x)_m - \frac{1}{r}\mathcal{D}_{r+1}(y)_m \\
&= x_m - \frac{1}{r}\left(1 - \sum_{j\le m} \mathcal{G}(x)_j\right) \\
&= x_m - \frac{1}{r}\sum_{j>m} \mathcal{G}(x)_j \\
&\le x_m - \frac{\delta}{r}.
\end{aligned}$$

Also note that

$$\begin{aligned}
\frac{1}{r+1} < \sum_{j\le m} y_j &\implies 1 < \sum_{j\le m} (rx_j + \mathcal{G}(x)_j) \\
&\implies x_m - \frac{1}{r}\left(1 - \sum_{j\le m} \mathcal{G}(x)_j\right) > 0 \\
&\implies \mathcal{K}_r(x)_m > 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
M(\mathcal{K}_r(x)) &= 2m + 1 - x_m + \frac{1}{r}\sum_{j>m} \mathcal{G}(x)_j \\
&= M(x) + \frac{1}{r}\sum_{j>m} \mathcal{G}(x) \\
&\ge M(x) + \frac{\delta}{r}.
\end{aligned}$$

Case 3: $\frac{1}{r+1} < \sum_{j<m} y_j$.

In this case, $1 < \sum_{j<m} (rx_j + \mathcal{G}(x)_j)$, which implies $1 < \sum_{j<m} \mathcal{G}(x)_j$. This is clearly impossible, so this case never happens. □

Proof of theorem 3.1: Let $x \in \Lambda$. Each term of the sequence $M(x), M(\mathcal{K}_r(x)), M(\mathcal{K}_r^2(x)), \ldots$ increases by δ/r from the previous term unless the previous term is $M(e_{n-1})$. Since the terms of the sequence are bounded above by $M(e_{n-1}) = 2(n-1)$, there is a $T > 0$ such that for all $t \geq T$, $M(\mathcal{K}_r^t(x)) = 2(n-1)$ and thus $\mathcal{K}_r^t(x) = e_{n-1}$. □

4 Continuous-time dynamical system models

Our objective in this section is to move from the expected value models of the previous section to an infinite population model. The incremental step in the simplex from one population to the next in the expected value models is either $\frac{1}{r}\mathcal{G}(x) - \frac{1}{r}\mathcal{D}_r(x)$ or $\frac{1}{r}\mathcal{G}(x) - \frac{1}{r}\mathcal{D}_r\left(\frac{rx+\mathcal{G}(x)}{r+1}\right)$. If the population size r is doubled then the size of the incremental step is halved in the first case and is approximately halved in the second case. Thus, in order to make the same progress in moving through the simplex, we need to take twice as many incremental steps of the expected value model. We can think of this as halving the time between incremental steps of the expected value model. We show below that this process corresponds to the well known limiting process of going from the Euler approximation of a differential equation to the differential equation itself.

We define a continuous-time dynamical system model which can be interpreted as the limit of the systems (1) and (2) as the population size goes to infinity and as the time step simultaneously goes to zero. Thus, we are interested in the limits of the functions $\mathcal{D}_r(x)$ for (1) and of $\mathcal{D}_r\left(\frac{rx+\mathcal{G}(x)}{r+1}\right)$ for (2). If this limit defines a continuous function $\mathcal{D}(x)$ that satisfies a Lipschitz condition, then we will show that the continuous-time system defined by the initial value problem

$$y' = \mathcal{E}(y) \qquad y(\tau) = \eta, \qquad \eta \in \Lambda, \tag{4}$$

where $\mathcal{E}(y) = \mathcal{G}(y) - \mathcal{D}(y)$, has a unique solution that exists for all $t \geq \tau$ and lies in the simplex. Further, it can be interpreted as the limit of the solutions of the systems (1) and (2) as the population size goes to infinity and the time step goes to zero.

It is easier to define what we mean by the convergence of the solutions to a family of discrete-time systems if we extend the discrete-time solutions to continuous-time solutions. An obvious way to do this is to connect successive points of the discrete-time trajectory by straight lines. The following makes this more precise.

Define $\mathcal{E}_r(x) = \mathcal{G}(x) - \mathcal{D}_r(x)$ to model the system (1) and define $\mathcal{E}_r(x) = \mathcal{G}(x) - \mathcal{D}_r\left(\frac{rx+\mathcal{G}(x)}{r+1}\right)$ to model the system (2).

Define

$$\begin{aligned} e_r(\tau) &= \eta \\ e_r(t) &= e_r(\tau + k/r) + \mathcal{E}_r(e_r(\tau + k/r))(t - (\tau + k/r)) \quad \text{for } \tau + k/r < t \leq \tau + (k+1)/r \end{aligned}$$

The following Lemma shows the $e_r(t)$ functions interpolate the solutions to the discrete-time systems (1) and (2). The proof is a straightforward induction.

Lemma 4.1 *For $k = 0, 1, \ldots,$*

$$e_r(\tau + k/r) = \mathcal{H}_r^k(\tau) = \mathcal{H}_r(\mathcal{H}_r(\ldots \mathcal{H}_r(\tau)\ldots))$$

or

$$e_r(\tau + k/r) = \mathcal{K}_r^k(\tau) = \mathcal{K}_r(\mathcal{K}_r(\ldots \mathcal{K}_r(\tau)\ldots)).$$

Note that if the solutions to (1) and (2) are in the simplex, then the convexity of the simplex implies that $e_r(t)$ is in the simplex for all $t \geq \tau$.

4.1 Extending the functions $\mathcal{E}$ and $\mathcal{E}_r$ to all of $\Re^n$

The standard existence and uniqueness theorems from the theory of differential equations are stated for a system $y' = F(t, y)$ where y ranges over $\Re^n$. (For example, see theorems 4.3 and 4.5 below.) In many cases, the $\mathcal{E}$ and $\mathcal{E}_r$ functions have natural extensions to all of $\Re^n$. In this case, these theorems can be directly applied. However, we would rather not make this assumption. Thus, to prove existence of solutions, we would like to extend the function $\mathcal{E} : \Lambda \to \Lambda$ to a continuous function defined over all of $\Re^n$. (The same technique can be applied to the $\mathcal{E}_r$ functions.)

Let H denote the hyperplane $\{x : \sum x_i = 1\}$ of $\Re^n$, and let $\mathbf{1}$ denote the vector of all ones. We first define a function R which retracts H onto the simplex Λ. Let $R(x)_i = \max(0, x_i)$. Clearly R is continuous, and

$$\|R(x) - R(y)\|_\infty \leq \|x - y\|_\infty \tag{5}$$

for all x, y.

Then we define a orthogonal projection p from $\Re^n$ onto H. Define p by $p(x) = x + (1 - \sum x_i)\mathbf{1}$. Clearly, p is continuous, and

$$\|p(x) - p(y)\|_\infty \leq \|x - y\|_\infty \tag{6}$$

for all x, y.

If $\mathcal{E} : \Lambda \to \Lambda$ is continuous, then $\mathcal{E}$ can be extended to a continuous function $\widetilde{\mathcal{E}} : \Re^n \to \Lambda$ by defining $\widetilde{\mathcal{E}}(x) = \mathcal{E}(R(p(x)))$. Clearly $\widetilde{\mathcal{E}}$ is bounded.

Lemma 4.2 *If $\mathcal{E}$ satisfies a Lipschitz condition, then so does $\widetilde{\mathcal{E}}$.*

Proof. Let $x, y \in \Re^n$. Then

$$\|\widetilde{\mathcal{E}}(x) - \widetilde{\mathcal{E}}(y)\|_\infty \leq L\|R(p(x)) - R(p(y))\|_\infty \leq L\|R(x) - R(y)\|_\infty \leq L\|x - y\|_\infty.$$

□

4.2 Existence of solutions in the simplex

The following theorem gives conditions under which a family of approximate solutions to an initial value problem converges to a solution to the problem. We will apply the theorem to show that a subsequence of the $e_r(t)$ functions converge to a solution to (4). If we know that the $e_r(t)$ solutions are contained in the simplex, then this will give us the existence of solutions in the simplex.

Theorem 4.3 *(Theorem 3.1 of [Rei71].) Suppose that the vector function $F(t, y)$ is continuous on the infinite strip $\Delta = [a, b] \times \Re^n$, and (τ, η) is a point of Δ. If ϵ_m, $(m = 1, 2, \ldots)$, is a sequence of positive constants converging to zero, and $y^{(m)}$ is a corresponding sequence of approximate solutions satisfying*

$$\|y^{(m)}(t) - \eta - \int_\tau^t F(s, y^{(m)}(s))ds\| < \epsilon_m \tag{7}$$

and for which there are corresponding constants κ, κ_1 such that

$$\|F(t, y^{(m)}(t))\| < \kappa\|y^{(m)}(t)\| + \kappa_1 \quad \text{on} \quad [a, b], \quad (m = 1, 2, \ldots) \tag{8}$$

then there is a subsequence $\{y^{(m_k)}\}$, $(m_1 < m_2 < \ldots)$, which converges uniformly on $[a, b]$ to a solution $y(t)$ of $y' = F(t, y)$ satisfying $y(\tau) = \eta$.

The following theorem gives conditions under which (4) has solutions in Λ that are limits of the the functions e_r. The uniqueness of these solutions is considered later.

Theorem 4.4 *Suppose that the functions $\mathcal{E}_r$ satisfy the condition of Lemma 2.2 for all r, and that $\lim_{r\to\infty}\mathcal{E}_r = \mathcal{E}$, where $\mathcal{E}$ is continuous on Λ. Let $[a,b]$ be an interval containing τ, and let $\eta \in \Lambda$. Then $y' = \mathcal{E}(y)$, $y(\tau) = \eta$ has a solution $y(t)$ defined on $[a,b]$ which is the limit of a subsequence of the functions $e_r(t)$.*

Proof. Our objective is to apply Theorem 4.3 with $\mathcal{E}(y)$ in place of $F(t,y)$. The hypothesis requires that $\mathcal{E}$ be defined on all of $\Re^n$. The previous subsection shows how to extend $\mathcal{E}$ to $\widetilde{\mathcal{E}}$ which is defined on all of $\Re^n$. To simplify notation, in this proof we use $\mathcal{E}$ in place of $\widetilde{\mathcal{E}}$.

Since any continuous function on a compact set is bounded, $\mathcal{E}$ is bounded by a positive number κ_1 on Λ. This shows that (8) holds for $\mathcal{E}$.

We also need to show that (7) holds. In other words, given $\epsilon > 0$, we need to show that there is an r_1 so that for $r \geq r_1$,

$$\left\| e_r(t) - \eta - \int_\tau^t \mathcal{E}(e_r(s))ds \right\| < \epsilon.$$

Since Λ is compact, the sequence of functions $\mathcal{E}_r$ converges uniformly to $\mathcal{E}$. Thus, there is an r_2 so that if $r \geq r_2$,

$$\|\mathcal{E}_r(x) - \mathcal{E}(x)\| < \frac{\epsilon}{6(b-a)} \quad \text{for all } x \in \Lambda. \tag{9}$$

We claim that given $\epsilon > 0$, there exists an $\delta > 0$ so that for all x,y with $\|x-y\| < \delta$ and $r \geq r_2$,

$$\|\mathcal{E}_r(x) - \mathcal{E}_r(y)\| < \frac{\epsilon}{2(b-a)} \tag{10}$$

To show this, choose δ sufficiently small so that $\|x-y\| < \delta$ implies $\|\mathcal{E}(x) - \mathcal{E}(y)\| < \frac{\epsilon}{6(b-a)}$. Then for $r \geq r_2$ and $\|x-y\| < \delta$,

$$\|\mathcal{E}_r(x) - \mathcal{E}_r(y)\| \leq \|\mathcal{E}_r(x) - \mathcal{E}(x)\| + \|\mathcal{E}(x) - \mathcal{E}(y)\| + \|\mathcal{E}(y) - \mathcal{E}_r(y)\| \leq \frac{\epsilon}{2(b-a)}.$$

Let the step function $g_r(t)$ be defined by

$$g_r(t) \quad = \quad \mathcal{E}_r\left(e_r\left(\tau + k/r\right)\right) \qquad \text{for } \ \tau + k/r \leq t < \tau + (k+1)/r.$$

Then $e_r(t)$ has the integral representation

$$e_r(t) = \eta + \int_\tau^t g_r(s)ds \quad \text{for } t \in [a,b]$$

We claim that that there is an r_1 so that $\|g_r(t) - \mathcal{E}_r(e_r(t))\| \leq \frac{\epsilon}{2(b-a)}$ for all $t \in [a,b]$ and $r \geq r_1$.

To show this, consider

$$\begin{aligned}
&\|g_r(t) - \mathcal{E}_r(e_r(t))\| \\
&= \|\mathcal{E}_r(e_r(\tau + k/r)) - \mathcal{E}_r(e_r(t))\| \text{ where } k \text{ is chosen so that } |t - (\tau + k/r)| \leq 1/r \\
&= \|\mathcal{E}_r(e_r(\tau + k/r)) - \mathcal{E}_r(e_r(\tau + k/r) + \mathcal{E}_r(e_r(\tau + k/r))(t - (\tau + k/r))\| \\
&= \|\mathcal{E}_r(x) - \mathcal{E}_r(y)\|
\end{aligned}$$

where $\|x-y\| = \|\mathcal{E}_r(e_r(\tau + k/r))(t - (\tau + k/r))\| \leq \kappa_1/r$. Choose r_1 so that $\kappa_1/r_1 < \delta$ and $r_1 \geq r_2$. Then (10) holds, and $\|g_r(t) - \mathcal{E}_r(e_r(t))\| \leq \frac{\epsilon}{2(b-a)}$ for all $t \in [a,b]$ and $r \geq r_1$.

Thus, for all $r \geq r_1$,

$$\begin{aligned}\left\| e_r(t) - \eta - \int_\tau^t \mathcal{E}(e_r(s))ds \right\| &= \left\| \int_\tau^t \left(g_r(s) - \mathcal{E}(e_r(s))\right) ds \right\| \\ &\leq \left\| \int_\tau^t \left(g_r(s) - \mathcal{E}_r(e_r(s))\right) ds \right\| + \left\| \int_\tau^t \mathcal{E}_r(e_r(s)) - \mathcal{E}(e_r(s))ds \right\| \\ &\leq \frac{\epsilon}{2(b-a)}(b-a) + \frac{\epsilon}{2(b-a)}(b-a) \\ &\leq \epsilon.\end{aligned}$$

□

4.3 Uniqueness of solutions

The following is a standard result on uniqueness from the theory of differential equations.

Theorem 4.5 *(Theorem 5.1 of [Rei71].) If on the infinite strip $\Delta = [a,b] \times \Re^n$ the vector function $F(t,y)$ is continuous and satisfies a Lipschitz condition*

$$\|F(t,z) - F(t,y)\| \leq \kappa \|z - y\|,$$

then for each point (τ, η) of Δ there is a unique solution $y(t)$ of $y' = F(t,y)$ on $[a,b]$ satisfying $y(\tau) = \eta$.

Corollary 4.6 *If the $\mathcal{E}_r$ satisfy the hypothesis of Lemma 2.2 and if $\mathcal{E} = \lim_{r\to\infty} \mathcal{E}_r$ satisfies a Lipschitz condition, then the system (4) has a unique solution which is defined for all $t \geq \tau$, and which lies in the simplex Λ.*

Proof. Given any interval $[a,b]$ with $\tau \in [a,b]$ and given $\eta \in \Lambda$, theorem 4.4 shows that (4) has a solution defined on $[a,b]$ which is contained in the simplex. The Lipschitz hypothesis on $\mathcal{E}$ shows that this solution is unique. Since the interval $[a,b]$ is arbitrary, this solution can be defined for all t. □

Let us summarize what we have shown. When the deletion heuristic is independent of population size, as it is for random deletion and inverse ranking deletion, then theorems 4.4 and 4.6 show that the trajectories of the discrete-time systems (1) and (2) approach the solution to the continuous time system (4) as the population size goes to infinity and the time step goes to zero. Thus, (4) is a natural infinite-population model for these discrete-time systems.

Theorems 4.4 and 4.6 do not apply to the case of worst-element deletion since the limit of the $\mathcal{D}_r$ functions as r goes to infinity is not continuous. (However, these theorems can be applied in the interior of the simplex and in the interior of every face of the simplex.) If the fitness is injective, then the function $\mathcal{D} = \lim_{r\to\infty} \mathcal{D}_r$ (where $\mathcal{D}_r$ denotes worst-element deletion) can be defined as follows. Let $k = k(x)$ have the property that $x_k > 0$ and $f_k \leq f_j$ for all j such that $x_j > 0$. Then $\mathcal{D}(x)_k = 1$ and $\mathcal{D}(x)_j = 0$ for all $j \neq k$. Figure 1 shows a trajectory of the system $y' = y - \mathcal{D}(y)$ where $\mathcal{D}$ has this definition. In this figure, e_0, e_1, e_2 are the unit vectors in R^3, and the fitnesses are ordered by $f_2 < f_1 < f_0$. The trajectory starts near e_2, and goes in a straight line with constant velocity to the (e_1, e_0) face. In the (e_1, e_0) face, the trajectory goes to e_0 with constant velocity.

5 Fixed points for random deletion

Theorem 5.1 *Under random deletion ($\mathcal{D}(x) = x$), all of the following systems:*

$$y' = \mathcal{G}(y) - y, \tag{11}$$

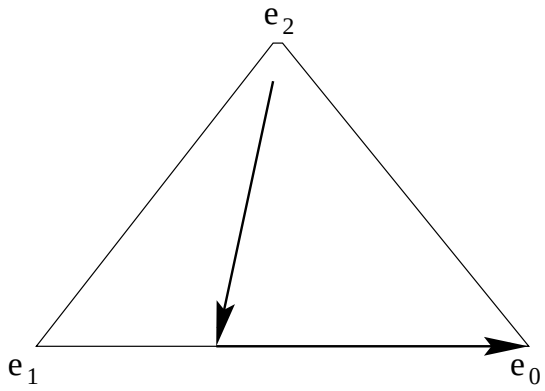

Figure 1: Trajectory of Worst-Element Deletion Continuous-time Model

$$x \to x + \frac{1}{r}(\mathcal{G}(x) - x) = \frac{r-1}{r}x + \frac{1}{r}\mathcal{G}(x), \tag{12}$$

$$x \to x + \frac{1}{r}\left(\mathcal{G}(x) - \frac{rx + \mathcal{G}(x)}{r+1}\right) = \frac{r}{r+1}x + \frac{1}{r+1}\mathcal{G}(x) \tag{13}$$

$$x \to \mathcal{G}(x) \tag{14}$$

have the same set of fixed points.

Proof. A necessary and sufficient condition for $\overline{x}$ to be a fixed point of all of these systems is $\mathcal{G}(\overline{x}) = \overline{x}$. □

The results of section 3 and the above results can be used to give conditions under which the fixed points of the steady-state $\mathcal{K}_r$ heuristic of equation (2) using worst-element deletion cannot be the same as the fixed points of the simple GA (or of steady-state with random deletion). We assume injective fitness and positive mutation for both algorithms. (By "positive mutation", we mean a nonzero probability of mutation from any string in the search space to any other.) The results of section 3 show that the only fixed point of the steady-state heuristic of equation (2) is the uniform population consisting of the optimum element in the search space. Any fixed point of the simple GA with positive mutation must be in the interior of the simplex.

6 Stability of fixed points

A fixed point $\overline{x}$ is said to be *stable* if for any $\epsilon > 0$, there is a $\delta > 0$ such that for any solution $y = y(t)$ satisfying $\|\overline{x} - y(\tau)\| < \delta$, then $\|\overline{x} - y(t)\| < \epsilon$ for all $t > \tau$. (For a discrete system, we can take $\tau = 0$, and interpret $t > \tau$ as meaning $t = 1, 2, 3, \ldots$)

A fixed point $\overline{x}$ is said to be *asymptotically stable* if $\overline{x}$ is stable and if there is an $\epsilon > 0$ so that if $\|y - \overline{x}\| < \epsilon$, then $\lim_{t\to\infty} y(t) = \overline{x}$.

The first-order Taylor approximation around $\overline{x}$ of (11) is given by

$$y' = \mathcal{G}(\overline{x}) - \overline{x} + (d\mathcal{G}_{\overline{x}} - I)(y - \overline{x}) + o(\|y - \overline{x}\|^2).$$

It is not hard to show (see Theorem 1.1.1 of [Wig90] for example) that if all of the eigenvalues of $d\mathcal{G}_{\overline{x}} - I$ have negative real parts, then the fixed point $\overline{x}$ is asymptotically stable.

The first-order Taylor approximation around $\overline{x}$ of (14) is given by

$$\mathcal{G}(y) = \mathcal{G}(\overline{x}) + d\mathcal{G}_{\overline{x}}(y - \overline{x}) + o(\|y - \overline{x}\|^2).$$

It is not hard to show (see Theorem 1.1.1 of [Wig90] for example) that if all of the eigenvalues of $d\mathcal{G}_{\overline{x}}$ have modulus less than 1 (has spectral radius less than 1), then the fixed point $\overline{x}$ is asymptotically stable.

The following lemma is straightforward.

Lemma 6.1 *Let $a \neq 0$ and b be scalars. Then λ is a multiplicity m eigenvalue of an $n \times n$ matrix A if and only if $a\lambda + b$ is a multiplicity m eigenvalue of the matrix $aA + bI$, where I is the $n \times n$ identity matrix.*

Theorem 6.2 *Let $\overline{x}$ be a fixed point of the system (14) where the modulus of all eigenvalues of $d\mathcal{G}_{\overline{x}}$ is less than 1. Then $\overline{x}$ is an asymptotically stable fixed point of (11), (12) and (13).*

Proof. Let λ be an eigenvalue of $d\mathcal{G}_{\overline{x}}$. By assumption $|\lambda| < 1$. Then $\lambda - 1$ is the corresponding eigenvalue for the system (11), and the real part of $\lambda - 1$ is negative.

The corresponding eigenvalue for (12) is $\frac{r-1}{r} + \frac{1}{r}\lambda$, and

$$\left|\frac{r-1}{r} + \frac{1}{r}\lambda\right| \leq \frac{r-1}{r} + \frac{1}{r}|\lambda| < 1$$

The argument for (13) is similar. □

If $d\mathcal{G}_{\overline{x}}$ has all eigenvalues with real parts less than 1 and some eigenvalue whose modulus is greater than 1, then $\overline{x}$ would be a stable fixed point of the continuous system (11) but an unstable fixed point of the generational discrete system (14). For the steady-state discrete system (12), the differential of the linear approximation is $\frac{r-1}{r}I + \frac{1}{r}d\mathcal{G}_{\overline{x}}$. As r goes to infinity, at some point the modulus of all eigenvalues of this differential will become less than 1, and the fixed point will become asymptotically stable.

We give a numerical example that demonstrates that this can happen. (See [WB97] for more details of the methodology used to find this example.) Assume a binary string representation with a string length of 3. The probability distribution over the mutation masks is

$$\langle\ 0.0 \quad 0.0 \quad 0.0 \quad 0.87873415 \quad 0.0 \quad 0.0 \quad 0.12126585 \quad 0.0\ \rangle^T.$$

The probability distribution over the crossover masks is

$$\langle\ 0.26654992 \quad 0.0 \quad 0.73345008 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0\ \rangle^T.$$

The fitness vector (proportional selection) is

$$\begin{array}{llll} \langle 0.03767273 & 0.40882046 & 3.34011500 & 3.57501693 \\ 0.00000004 & 3.89672742 & 0.21183468 & 15.55715272 \rangle^T. \end{array}$$

The fixed point is

$$\begin{array}{llll} \langle 0.20101565 & 0.21467902 & 0.07547095 & 0.06249578 \\ 0.26848520 & 0.04502642 & 0.11812778 & 0.01469920 \rangle^T. \end{array}$$

This gives a set of eigenvalues:

$$\begin{array}{l} \{\ -1.027821882 + 0.01639853054i, \quad -1.027821882 - 0.01639853054i, \quad 0.5097754068, \\ \quad -0.3498815639, \quad 0.1348641055, \quad -0.01080298133, \\ \quad 0.2146271583 \times 10^{-5}, \quad 0.6960358287 \times 10^{-9}\} \end{array}$$

7 An illustrative experiment

It is a remarkable result that a steady-state genetic algorithm with random deletion has the same fixed-points as a generational genetic algorithm with common heuristic function $\mathcal{G}$. We can illustrate this result

experimentally as follows. Firstly, we choose some selection, crossover and mutation scheme from the wide variety available. It doesn't matter which are chosen as long as the same choice is used for the steady-state and generational GAs. In our experiments we have used binary tournament selection, uniform crossover and bitwise mutation with a rate of 0.01. Together, these constitute our choice of heuristic function $\mathcal{G}$. Secondly, we pick a simple fitness function, for example, the one-max function on 100 bits. Thirdly, we choose two different initial populations, one for each GA. These should be chosen to be far apart; for example, at different vertices of the simplex. In our experiments, the steady-state GA starts with a population of strings containing all ones, whereas the generational GA has an initial population of strings containing only zeros. A population size of 1000 was used.

The two GAs were run with these initial populations. To give a rough idea of what is happening, the average population fitness for each was recorded for each "generation". For the steady-state GA this means every time 1000 offspring have been generated (that is, equivalent to the population size). This was repeated ten times. The average results are plotted in the first graph of figure 2.

To show that the two genetic algorithms are tending towards exactly the same population, the (Euclidean) distance was calculated between the corresponding population vectors at each generation. By "population vector" is here meant a vector whose components give the proportions of the population within each *unitation class*. The results for a typical run are shown in the second graph of figure 2. It can be seen that after around 70 generations, the two GAs have very similar populations. Figure 3 shows the average (over 20 runs) distance between the algorithms where both algorithms are started with the population consisting entirely of the all-zeros string. The error bars are one standard deviation. These figures show that the two algorithms follow very different trajectories, but with the same fixed points.

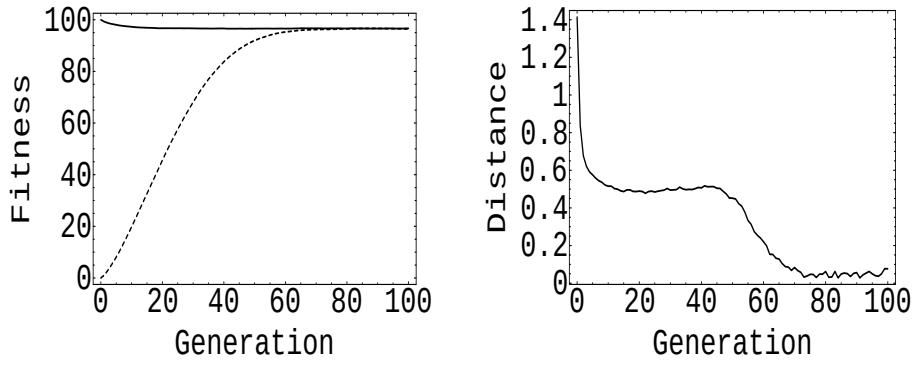

Figure 2: a) average population fitness of steady-state GA (solid line) and generational GA (dashed line), averaged over ten runs. b) Distance between steady-state GA and generational GA for a typical run.

8 Conclusion and further work

We have given discrete-time expected-value and continuous-time infinite-population dynamical system models of steady-state genetic algorithms. For one of these models and worst-element deletion, we have given conditions under which convergence to the uniform population consisting of copies of the optimum element is guaranteed.

We have shown the existence of solutions to the continuous-time model by giving conditions under which the discrete-time models converge to the solution of the continuous-time model. And we have given conditions for uniqueness of solutions to the continuous-time model.

We have investigated the fixed points and stability of these fixed points for these models in the case of

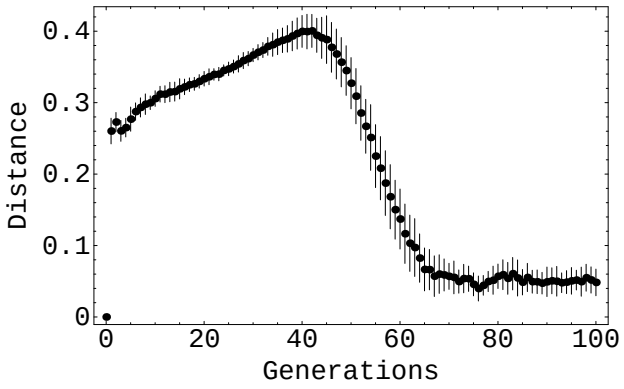

Figure 3: The distance between the steady-state GA and the generational GA averaged over 20 runs. The error bars represent one standard deviation.

worst-element and random deletion. Further work is needed to investigate the properties of fixed points for these and other deletion methods.

The relationship of these models to the Markov chain models of steady-state algorithms given in [WZ99] could also be investigated.

Acknowledgments

The first author thanks Alex Agapie for discussions regarding section 3.

References

[Dav91] Lawrence Davis. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, 1991.

[Rei71] William T. Reid. *Ordinary Differential Equations*. John Wiley & Sons, New York, 1971.

[Rud98] Günter Rudolph. Finite markov chain results in evolutionary computation: A tour d'horizon. *Fundamenta Informaticae*, 35:67–89, 1998.

[Sys89] Gilbert Syswerda. Uniform crossover in genetic algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 2–9. Morgan Kaufman, 1989.

[Sys91] Gilbert Syswerda. A study of reproduction in generational and steady state genetic algorithms. In Gregory J. E. Rawlings, editor, *Foundations of genetic algorithms*, pages 94–101, San Mateo, 1991. Morgan Kaufmann.

[Vos99] M. D. Vose. *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge, MA, 1999.

[WB97] A. H. Wright and G. L. Bidwell. A search for counterexamples to two conjectures on the simple genetic algorithm. In *Foundations of genetic algorithms 4*, pages 73–84, San Mateo, 1997. Morgan Kaufmann.

[Whi89] Darrell Whitley. The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive trials is best. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 116–123. Morgan Kaufman, 1989.

[Wig90] S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer-Verlag, New York, 1990.

[WZ99] A. H. Wright and Y. Zhao. Markov chain models of genetic algorithms. In *Proceedings of the Genetic and Evolutionary Computation (GECCO) conference*, pages 734–742, San Francisco, CA., 1999. Morgan Kaufmann Publishers.