Annals of Mathematics and Artificial Intelligence 5 (1992) 79–88

Modeling genetic algorithms with Markov chains *

Allen E. Nix and Michael D. Vose

Computer Science Department, 107 Ayres Hall, The University of Tennessee, Knoxville, TN 37996-1301, USA

We model a simple genetic algorithm as a Markov chain. Our method is both complete (selection, mutation, and crossover are incorporated into an explicitly given transition matrix) and exact; no special assumptions are made which restrict populations or population trajectories. We also consider the asymptotics of the steady state distributions as population size increases.

1. Introduction

Designed to search irregular, poorly understood spaces, genetic algorithms (GAs) are general purpose algorithms developed by Holland [2]. Inspired by the example of population genetics, genetic search proceeds over a number of generations. The criteria of "survival of the fittest" provides evolutionary pressure for populations to develop increasingly fit individuals. Although there are many variants, the basic mechanism of a GA consists of:

(1) Evaluation of individual fitness and selection of a gene pool.
(2) Mutation and crossover.

Individuals resulting from these operations form the members of the next generation, and the process is iterated until the system ceases to improve.

Fixed length binary strings are typically the members of the population. They are selected (with replacement) for the gene pool with probability proportional to their relative fitness, which is determined by the objective function. There, they are recombined by mutation and crossover. Mutation corresponds to flipping the bits of an individual with some small probability (the mutation rate). The simplest implementation of crossover selects two "parents" from the pool and, after choosing the same random position within each string, exchanges their tails. Crossover is typically performed with some probability (the crossover rate), and parents are otherwise cloned. This recombination cycle repeats,

* This research was supported by the National Science Foundation (IRI-8917545).

contributing one of the resulting "offspring" each time until the next generation is full.

While this description may suffice for successful application of the genetic paradigm, it does not place GAs within a mathematical framework suitable for analysis.

Vose and Liepins [4] have previously presented a rigorous mathematical formalism for a simple GA. However, their approach models GAs by computing expected population trajectories through time based on an infinite population.

In contrast, we make no special assumptions. We neither assume populations are infinite, nor do we limit them to following expected evolutionary paths. Our populations are finite, and population trajectories are unconstrained. Our model is completely stochastic.

Independently, Davis has also modeled a simple GA as a Markov chain [1]. However, our work differs significantly from his. While he considers the asymptotics of steady state distributions as the mutation rate decreases, we investigate the asymptotics as population size increases. Moreover, our results concerning the matrix of transition probabilities are based on the model of Vose and Liepins, which simplifies representation and calculation.

2. States

Let Ω be the collection of length l binary strings, and let $r = |\Omega| = 2^l$ be the number of possible strings. Let P be a population of elements from Ω, let $n = |P|$ be the population size, and let N be the number of possible populations.

DEFINITION 1

Let Z be an $r \times N$ matrix whose columns represent the possible populations of size n. The ith column $\phi_i = \langle z_{0,i}, \ldots, z_{r-1,i} \rangle^{\mathrm{T}}$ of Z is the incidence vector for the ith population P_i. [1] That is,

$$z_{y,i} = \text{the number of occurrences of string } y \text{ in } P_i,$$

where integers y are identified with their binary representations, and indexing begins with 0.

As an example, if $l = 2$ and $n = 1$, then $\Omega = \{00, 01, 10, 11\}$ and the possible populations can be enumerated as

$$P_0 = \{11\}, \quad P_1 = \{10\}, \quad P_2 = \{01\}, \quad P_3 = \{00\}.$$

[1] T denotes transpose.

With this enumeration, we have

$$Z = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \phi_0 = \langle 0001 \rangle^{\mathrm{T}}, \quad z_{3,0} = 1.$$

This example points out that Z is well defined only with respect to some enumeration of the possible populations. Our convention will be to enumerate P_i before P_j if interpreting (the transpose of) incidence vectors as n-ary integers yields $\phi_i < \phi_j$.

LEMMA 1

The number of possible populations is

$$N = \binom{n+r-1}{r-1}.$$

Proof

The incidence vector ϕ_i can be represented graphically by using dots and slashes. Each dot represents one string, therefore a population P_i of n strings is represented with n dots. To represent $z_{y,i}$ instances of string y, a slash is put after the $(z_{0,i} + \cdots + z_{y,i})$th dot. As an example, if $r = 4$ and $n = 5$ then

$$\cdot\cdot//\cdot\cdot/\cdot/$$

represents the incidence vector $\langle 2, 0, 2, 1 \rangle$ which indicates two strings each of 0 and 2, one string of 3, and no strings of 1. Note that only $r-1$ slashes are required to represent a distribution of r possible strings in a population, since the last slash is not necessary. If $r-1$ dots are added to the n dots, then any population of size n could be represented simply by correctly choosing the $r-1$ dots through which to put slashes. Using this scheme, this previous population would be represented as

$$\cdot\cdot//\cdot\cdot/\cdot$$

Since it is possible to represent all populations using this method (and each population is represented only once) the number of possible populations N is the number of ways of choosing $r-1$ dots from a total of $n+r-1$ dots. □

We model a genetic algorithm with a Markow chain having the N possible populations as states. Therefore, the columns of Z describe the states of our model.

3. Transition probabilities

DEFINITION 2

Assuming a simple genetic algorithm operates on each generation to produce the next, let Q be the $N \times N$ transition matrix where Q_{ij} is the probability that the kth generation will be P_j given that the $(k-1)$st generation is P_i.

The transition probabilities may be calculated by considering how the population incidence vector ϕ_j describes the composition of the next generation.

Let $p_i(y)$ be the probability of producing string y in the next generation given that the current population is P_i. If the next generation is P_j, then there must be $z_{y,j}$ occurrences of string y produced. The probability of this is almost given by $\{p_i(y)\}^{z_{y,j}}$, but we must take into account that exactly which $z_{y,j}$ members of the next population are y is not important; we only care that the correct number of strings occur.

The number of ways of choosing $z_{0,j}$ occurrences of string 0 for a population of size n is given by the binomial coefficient

$$\binom{n}{z_{0,j}}.$$

This leaves $n - z_{0,j}$ positions remaining in the next generation to fill. Moving on to string 1, we see that the number of ways of choosing its $z_{1,j}$ occurrences is given by

$$\binom{n - z_{0,j}}{z_{1,j}}.$$

Continuing in this fashion, the total combination for all strings is given by

$$\binom{n}{z_{0,j}}\binom{n - z_{0,j}}{z_{1,j}} \cdots \binom{n - z_{0,j} - z_{1,j} - \cdots - z_{r-2,j}}{z_{r-1,j}} = \frac{n!}{z_{0,j}! z_{1,j}! \cdots z_{r-1,j}!}.$$

Therefore, the probability of producing population P_j from population P_i can be written as a multinomial distribution with parameters n, $p_i(0), \ldots, p_i(r-1)$:

$$Q_{i,j} = \frac{n!}{z_{0,j}! z_{1,j}! \cdots z_{r-1,j}!} \prod_{y=0}^{r-1} \{p_i(y)\}^{z_{y,j}} = n! \prod_{y=0}^{r-1} \frac{\{p_i(y)\}^{z_{y,j}}}{z_{y,j}!}.$$

What remains is to calculate the conditional probability function $p_i(y)$ which gives the probability of producing string y from population P_i. The model of Vose and Liepins [4] contains the ingredients for this calculation. We summarize relevant material from their paper and refer to [4] for details.

Given a vector x, let $|x|$ denote the sum of its coordinates. Let $\oplus$ be *exclusive-or* on integers and let $\otimes$ be *logical-and*.

Let $m_{i,j}(k)$ be the probability that k results from the recombination process based on parents i and j. If recombination is 1-point crossover and mutation

with crossover rate χ and mutation rate μ, then

$$m_{i,j}(0) = \frac{(1-\mu)^l}{2}\left\{\eta^{|i|}\left(1-\chi+\frac{\chi}{l-1}\sum_{k=1}^{l-1}\eta^{-\Delta_{i,j,k}}\right)\right.$$
$$\left.+\eta^{|j|}\left(1-\chi+\frac{\chi}{l-1}\sum_{k=1}^{l-1}\eta^{\Delta_{i,j,k}}\right)\right\},$$

where $\eta = \mu/(1-\mu)$, integers are to be regarded as bit vectors when occurring in $|\cdot|$, where division by zero at $\mu = 0$ and $\mu = 1$ is to be removed by continuity, and where

$$\Delta_{i,j,k} = |(2^k - 1) \otimes i| - |(2^k - 1) \otimes j|.$$

Define permutations σ_j on $\mathbb{R}^r$ by

$$\sigma_j\langle x_0, \ldots, x_{r-1}\rangle^{\mathrm{T}} = \langle x_{j\oplus 0}, \ldots, x_{j\oplus(r-1)}\rangle^{\mathrm{T}}.$$

Let M be the matrix having (i, j)th entry $m_{i,j}(0)$, and define the operator $\mathscr{M}$ by

$$\mathscr{M}(x) = \langle(\sigma_0 x)^{\mathrm{T}} M\sigma_0 x, \ldots, (\sigma_{r-1}x)^{\mathrm{T}} M\sigma_{r-1}x\rangle^{\mathrm{T}}.$$

Let f be the positive objective function, and define F to be the linear operator having diagonal matrix with (i, i)th entry $f(i)$.

The operators F and $\mathscr{M}$ form the basis of the model of Vose and Liepins. The utility of these operators follows from:

- If v is a vector with yth component equal to the proportion of y in the population, then Fv is a vector pointing in the same direction as that vector having yth component equal to the probability that y will be selected for recombination.
- If v is a vector with yth component equal to the probability that y will be selected for recombination, then $\mathscr{M}(v)$ is a vector with yth component equal to the expected proportion of y in the next generation.

Note that the yth component of $\phi_i/|\phi_i|$ is the proportion of y in population P_i. Since F is a linear operator, it follows from the first point above that

$$\frac{F\phi_i}{|F\phi_i|}$$

has yth component equal to the probability that y will be selected for recombination. Therefore by the second point above, the yth component of

$$\mathscr{M}\left(\frac{F\phi_i}{|F\phi_i|}\right)$$

is the expected proportion of y in the next generation. Since the expected proportion is the probability of occurrence, we have established the following:

THEOREM 1

The transition matrix is given by

$$Q_{i,j} = n! \prod_{y=0}^{r-1} \frac{\left\{ \mathcal{M}\left(\frac{F\phi_i}{|F\phi_i|} \right)_y \right\}^{z_{y,j}}}{z_{y,j}!}.$$

We have constructed an exact model of a simple GA. The states of the Markov chain are given by the columns of the matrix Z, and the transition probabilities are given by the matrix Q.

4. Asymptotics

Let π^k be the probability row vector having jth component equal to the probability that the kth generation is P_j. It follows from the definition of the transition matrix Q that

$$\pi^k = \pi^0 Q^k,$$

where π^0 is the initial distribution. The steady state distribution π is given by

$$\lim_{k\to\infty} \pi^k = \lim_{k\to\infty} \pi^0 Q^k = \text{the solution to the equation } "\pi = \pi Q"$$

and has jth component which may be interpreted as the relative proportion of the time that the GA has a population corresponding to P_j. The steady state distribution is independent of the initial population because we assume that mutation is nonzero and hence the Markov chain is ergodic.

If the GA were to converge to some population P_j, then it would asymptotically (as the number of generations went to infinity) spend all its time in state P_j. Hence π would be the basis vector e_j which has 1 in the jth component and 0 elsewhere.

This cannot happen since there are no absorbing states, and a GA will tend to wander due to the stochastic nature of selection, crossover, and mutation. However, if we take the limit of steady state distributions as population size goes to infinity

$$\pi^* = \lim_{n\to\infty} \pi,$$

then the distribution π^* may very well correspond to convergence to some fixed population since fluctuations are averaged over increasingly large populations. This section is devoted to making this precise, and culminates in a characterization of the distribution π^*. We will use some basic probability theory, and refer the reader to [3] for background.

Note that the states of our Markov chain can be regarded as points on the simplex

$$\Lambda = \{x \in \mathbb{R}^r : x \text{ is nonnegative and } |x| = 1\}$$

through the correspondence $P_j \leftrightarrow \phi_j/n$, since a population incidence vector has coordinates which sum to n. As $n \to \infty$, the states, which we will from here on identify with points of the simplex, become dense in Λ.

Since π is for each n a probability measure over the compact set Λ, a theorem of Prokhorov (see [3]) implies that every infinite sequence has an infinite subsequence which converges weakly. The main result of this section is that an accumulation point π^* of $\{\pi\}_{n>0}$ can give positive probability only to the fixed points of the operator

$$\mathcal{G}(x) = \mathcal{M}\left(\frac{Fx}{|Fx|}\right),$$

provided that the iterates of $\mathcal{G}$ converge.

LEMMA 2

Let the Markov chain be in state x at time t and in state x' at time $t+1$. Given $\epsilon > 0$ and $\gamma < 1$, there exists N independent of x such that with probability at least γ

$$n > N \Rightarrow \| x' - \mathcal{G}(x) \| < \epsilon.$$

Proof

We have seen that the distribution of the vector ϕ_j describing the population produced from P_i is multinomial with parameters $n, p_i(0), \ldots, p_i(r-1)$. It follows that the expectation of ϕ_j is

$$\langle np_i(0), \ldots, np_i(r-1) \rangle^{\mathrm{T}} = n\mathcal{G}(\phi_i).$$

Since all norms are equivalent on $\mathbb{R}^r$ we may use $\| \cdot \|_\infty$. Hence the probability in question is the probability that the coordinate x_j' is not within ϵ of $p_i(j)$. Since the coordinates are binomially distributed, Chebyshev's inequality gives

$$Prob\{|x_j' - p_i(j)| > \epsilon\} < \frac{p_i(j)(1-p_i(j))}{n\epsilon^2} = O\left(\frac{1}{n\epsilon^2}\right). \qquad \square$$

Since the fitness function f is positive, the operator $F \cdot / |F \cdot|$ is continuous on Λ. It follows that $\mathcal{G}$ is continuous since it has coordinate functions which are the composition of this with quadratic maps. Moreover, since Λ is compact, $\mathcal{G}$ is uniformly continuous.

Let the Markov chain be in state x_t at time t, and let $\mathcal{G}^t$ denote the tth iterate of $\mathcal{G}$. The next theorem shows that with probability arbitrarily close to 1, population trajectories converge to iterates of $\mathcal{G}$ as $n \to \infty$.

THEOREM 2

Given $k > 0$, $\epsilon > 0$ and $\gamma < 1$, there exists N such that with probability at least γ and for all $0 \leqslant t \leqslant k$

$$n > N \Rightarrow \| x_t - \mathcal{G}^t(x_0) \| < \epsilon.$$

Proof

We induct on k. The base case $k = 1$ is lemma 2. Since $\mathscr{G}$ is uniformly continuous, choose δ such that

$$\|x_{k-1} - \mathscr{G}^{k-1}(x_0)\| < \delta \Rightarrow \|\mathscr{G}(x_{k-1}) - \mathscr{G}^{k}(x_0)\| < \epsilon/2.$$

By the inductive hypothesis, if $n > N_0$ then with probability at least $1 - (1 - \gamma)/2$ we have

$$\|x_{k-1} - \mathscr{G}^{k-1}(x_0)\| < \delta.$$

By lemma 2, let N_1 be such that with probability at least $1 - (1 - \gamma)/2$

$$n > N_1 \Rightarrow \|x_k - \mathscr{G}(x_{k-1})\| < \epsilon/2.$$

It follows that if $n > N = \max\{N_0, N_1\}$, then with probability at least γ

$$\|x_k - \mathscr{G}^{k}(x_0)\| \leqslant \|x_k - \mathscr{G}(x_{k-1})\| + \|\mathscr{G}(x_{k-1}) - \mathscr{G}^{k}(x_0)\| < \epsilon/2 + \epsilon/2.$$

□

Let $\mathscr{F}$ be the fixed points of $\mathscr{G}$, and let π be the probability measure corresponding to the steady state distribution of the Markov chain for population size n. The next theorem shows that under certain conditions, the probability of being away from $\mathscr{F}$ vanishes as $n \to \infty$.

THEOREM 3

Suppose that $\lim_{t\to\infty} \mathscr{G}^t(x) \in \mathscr{F}$ for every $x \in \Lambda$. If π^* is a weak accumulation point of $\{\pi\}_{n>0}$, then

$$\pi^*(\mathscr{F}) = 1.$$

Proof

It suffices to show $\pi^*(\mathscr{U}) = 1$ for every open set $\mathscr{U}$ containing $\mathscr{F}$. Since $\mathscr{G}$ is continuous, its fixed points form a compact set. Note that $\mathscr{C} = \Lambda \setminus \mathscr{U}$ is also compact, and let ρ be the distance between $\mathscr{C}$ and $\mathscr{F}$. Let $\mathscr{F}_{\rho/4}$ be the set of points within a distance of $\rho/4$ from $\mathscr{F}$.

Since $\lim_{t\to\infty} \mathscr{G}^t(x) \in \mathscr{F}$ for every $x \in \Lambda$, the set $\mathscr{C}$ is the disjoint union of the sets

$$E^j = \left\{x \in \mathscr{C} : \mathscr{G}^j(x) \notin \mathscr{F}_{\rho/4} \wedge t > j \Rightarrow \mathscr{G}^t(x) \in \mathscr{F}_{\rho/4}\right\}.$$

Therefore

$$\pi^*(\mathscr{U}) = 1 - \pi^*(\mathscr{C}) = 1 - \sum \pi^*(E^j).$$

The proof is completed by showing $\pi^*(E^j) = 0$. Let $h : \Lambda \to [0,1]$ be a continuous function which is 1 on E^j. Then

$$\pi^*(E^j) \leqslant \int h \, \mathrm{d}\pi^* \leqslant \limsup_{n\to\infty} \int h \, \mathrm{d}\pi.$$

Hence we are done if, given $\epsilon > 0$, there exist N and suitable h such that

$$n > N \Rightarrow \int h \, d\pi < \epsilon.$$

Let E_δ^j be the set of points within a distance of δ from E^j, and given $y_0 \in E_\delta^j$, let $x_0 \in E^j$ be such that $\| y_0 - x_0 \| < \delta$. Let $\epsilon > 0$ and let $k > 2j\epsilon^{-1}$. Since $\mathscr{G}$ is uniformly continuous, let δ be smaller than $\rho/4$ and such that for all $0 \leqslant t \leqslant k$

$$\| \mathscr{G}^t(y_0) - \mathscr{G}^t(x_0) \| < \rho/4.$$

By theorem 2, let N be such that with probability $1 - \epsilon/2$ and for all $0 \leqslant t \leqslant k$

$$n > N \Rightarrow \| y_t - \mathscr{G}^t(y_0) \| < \rho/4.$$

Let $d(\cdot, \mathscr{F})$ be the function measuring distance from $\mathscr{F}$. If follows that if $n > N$ and $j < t \leqslant k$, then with probability $1 - \epsilon/2$, and for all $y_0 \in E_\delta^j$,

$$\begin{aligned} d(y_t, \mathscr{F}) &\leqslant \| y_t - \mathscr{G}^t(y_0) \| + \| \mathscr{G}^t(y_0) - \mathscr{G}^t(x_0) \| + d(\mathscr{G}^t(x_0), \mathscr{F}) \\ &< 3\rho/4. \end{aligned}$$

This means that if the Markov chain was at a state in E_δ^j, then with probability $1 - \epsilon/2$ it will spend at least the fraction $(k-j)/k > 1 - \epsilon/2$ of the time outside of E_δ^j provided that $n > N$. Hence choosing h to be zero outside of E_δ^j gives

$$\int h \, d\pi \leqslant \pi(E_\delta^j) < \epsilon,$$

provided that $n > N$. □

5. Conclusion

We model a simple genetic algorithm as a Markov chain. Our model is both complete (selection, mutation, and crossover are incorporated into an explicitly given transition matrix) and exact; no special assumptions are made which restrict populations or population trajectories.

Our Markov model is described in detail; the states and the transition probability matrix are given in terms of the parameters of the GA (population size, mutation rate, and crossover rate).

We investigate the asymptotic behavior of the steady state distributions as population size increases. The model of Vose and Liepins is embedded in the machinery of the exact Markov chain representation for genetic search, and the asymptotic behavior of the steady state distributions is characterized in terms of the fixed points of their model.

In particular, if their model has a unique (attracting) fixed point, then theorem 3 implies that as population size grows, the GA spends asymptotically

all its time at the population corresponding to that fixed point. In other words: if the finite population is sufficiently large, *we can accurately predict the convergence behaviour of a real GA.*

If there is more than a single fixed point, then theorem 3 only characterizes where the limit points of the steady state distributions can give positive probability. In this general case, exactly how mass is distributed over the fixed points has not yet been established. We believe it is determined by the relative sizes of the basins of attraction of the fixed points of $\mathscr{G}$.

We conjecture that the fixed point of $\mathscr{G}$ with largest basin is where (as population size grows) the GA spends asymptotically all its time. If this is true, then the model of Vose and Liepins would determine the convergence behaviour of real GAs (with large populations) in every case.

It should be remarked that the results of the previous section were proved uniformly in the parameters of the GA. In particular, if a Markov chain were non-stationary in that the mutation parameter μ were gradually lowered to zero (as suggested by Davis [1]), then every weak accumulation point π^* distributes mass only at the fixed points of the operator $\mathscr{G}$ corresponding to $\mu = 0$.

References

[1] T. Davis, Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm, dissertation presented to the University of Florada (1991).

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

[3] R.G. Laha and V.K. Rohatgi, *Probability Theory* (Wiley, 1979).

[4] M.D. Vose and G.E. Liepins, Punctuated equilibria in genetic search, Complex Systems 5 (1991) 31–44.