Techniques for Bounding the Convergence Rate of Genetic Algorithms

Yuri Rabinovich[1], Avi Wigderson[2]
[1]*Department of Computer Science, University of Toronto, Ontario, Canada M5S 1A4; e-mail: yuri@cs.haifa.ac.il*
[2]*Department of Computer Science, Hebrew University, Jerusalem, Israel 91904; e-mail: avi@cs.huji.ac.il*

ABSTRACT: The main purpose of the present paper is the study of computational aspects, and primarily the convergence rate, of genetic algorithms (GAs). Despite the fact that such algorithms are widely used in practice, little is known so far about their theoretical properties, and in particular about their long-term behavior. This situation is perhaps not too surprising, given the inherent hardness of analyzing nonlinear dynamical systems, and the complexity of the problems to which GAs are usually applied. In the present paper we concentrate on a number of very simple and natural systems of this sort, and show that at least for these systems the analysis can be properly carried out. Various properties and tight quantitative bounds on the long-term behavior of such systems are established. It is our hope that the techniques developed for analyzing these simple systems prove to be applicable to a wider range of genetic algorithms, and contribute to the development of the mathematical foundations of this promising optimization method. © 1999 John Wiley & Sons, Inc. Random Struct. Alg., 14, 111–138, 1999

1. INTRODUCTION

Genetic algorithms provide a heuristic paradigm based on a biological analogy for solving combinatorial optimization problems. They were formally introduced by Holland in [5] in 1975, and ever since were continuously gaining popularity as an

appealing and efficient tool for finding satisfactory solutions for hard practical optimization problems.

Suppose one is searching for an optimal (or, rather, a near optimal) solution among the set of all feasible solutions. The genetic algorithm (GA) approach is to design a dynamical system, in which the feasible solutions, viewed as "creatures," "evolve" in a space-homogeneous world under artificial selection. To this end, each feasible solution is prescribed a distinct genetic type. The *quality* or *fitness* of a solution is defined in such a way that "high quality" corresponds to "nearly optimal." The algorithm starts with an initial distribution (or, using a more lively term, population) p, consisting of simple-structures feasible solutions. The population then evolves in discrete time-steps under the action of two operators, the *mating* operator [denoted $M(p)$], and the *fitness* operator [denoted $W(p)$]. Sometimes a third operator, called *mutation*, is added; in this paper, for the sake of clarity and simplicity, we only consider the systems defined by the first two operators.

The mating operator is designed to increase variety in the population. It acts by forming random (according to the current distribution) pairs of solutions, and replacing them by their randomly obtained (according to the definition of the mating operator) children. Informally, all it does is create new legal genetic types from combinations of the old ones. In all the cases we discuss, the mating operator preserves the expected quality of the population. This important property is not a part of the standard definition GA.

The fitness operator W embodies the selection principle; it is responsible for making the average quality increase. It acts by reducing the proportion of the low quality solutions. A fine point about W is that if it is too "tolerant," the evolution may take too long, while if it is too "picky," a good portion of the population may become extinct before being able to distribute their genes. The danger in the latter case is that with a good portion of the genes lost, the really good solutions might become unreachable. All fitness functions discussed in this paper are linear.

In the overall genetic algorithm, the population quality constantly increases (due to W), while the possibility of appearance of near optimal solutions in the population is, hopefully, never lost (due to the reproductive power of M, which should be in a good balance with the extinctive power of W).

Despite the apparent success of this approach in many experimental applications (see, e.g., the annual proceedings of the International Conference on Genetic Algorithms, [11, 6] and many others), few rigorous theoretical results exist to explain and motivate the experimental results or to guide the design of such algorithms. This state of affairs is, perhaps, not too surprising for complex and poorly understood systems for which GAs are usually applied. There is no reason, however, why it should be so for the simpler systems. Motivated by a belief that a good understanding of the simpler special cases should eventually lead to a better understanding of the general case, we concentrate in the present paper on the study of three different simple but naturally arising GAs.

To the best of our knowledge, the first quantitative results on the long-term behavior of any concrete, nontrivial, and naturally arising genetic algorithm were obtained in [10] and [8]. While the systems discussed were quite simple to define, the mathematics involved in their analysis turned out to be both interesting and suggestive. They have inspired to some degree the latter, much more general, [9]

and [7]. The present paper is dedicated to the presentation of extended and improved results of [10] and [8].

Among more recent papers on analysis of genetic algorithms, one should mention [1]; it provides perhaps the first example in a literature of a genetic algorithm which solves a nontrivial (theoretical) problem, and whose running time compares favorably with that of existing nongenetic algorithms. For interesting qualitative results on the asymptotic behavior of genetic algorithms, see also the recent [3].

As in [10] and [8], the focus of the present paper is not so much on applications of GAs, or on the in-depth study of the particular systems discussed, but rather on developing mathematical tools and methods needed for their analysis. The goal is to acquire a better understanding of how genetic algorithms work in general. Three different genetic algorithms are analyzed; the analysis proceeds along similar lines in all three cases: first some properties of the mating operator M are established; then upper and lower bounds on average fitness of time-t population are obtained; finally the dynamical system defined by the mating operator M alone is discussed. The operator M usually gets much more attention than W; the reason for this is that in all cases considered, M is far more complex than W, and a good understanding of its properties is essential for the analysis of the entire system.

In the case of the most important GA discussed in this paper (Section 3), we also briefly discuss (in Section 3.5) the important issue of behavior of the more practical systems of a finite population size, versus that of the abstract GAs, which model infinite-size populations.

The most interesting system analyzed in this paper (Section 3) is a GA optimizing a linear function on the set of all binary strings of length n. Although it is common knowledge that GAs perform well on this problem (e.g., computer simulations, heuristic arguments about short and nonconflicting schemata, etc.; see [5, 12]), no rigorous results supporting this empirical knowledge exist. The analysis was first performed for a special case of $f(s) = \sum_i s_i$ in [10], then again in [1] for a different system, suited for a particular application. In the present paper we follow [10], and generalize its results in several directions.

Linear fitness functions arise whenever the quality of a solution depends on a number of factors, so that the contribution of each factor is independent from the others. Such fitness functions, although at first glance not too powerful, were successfully used in many complicated applications, e.g., in the famous classical program by Samuel for the game of checkers [13]. The power of such systems lies in a clever choice of factors determining the fitness functions, and in assigning them proper relative weights. Often the linear function is hidden, i.e., implemented by an adversary or by nature, and not known in advance to the designer of the genetic algorithm. Optimizing linear functions is a good starting point for the theoretical study of GAs for a number of reasons. They are naturally related to harder optimization problems, which often maximize linear functions on a complex domain. Being sufficiently simple and thus hopefully tractable, they are at the same time sufficiently hard to analyze. Finally, they display some fundamental properties shared by a much wider class of systems, and thus can be used for forming new conjectures and checking new proof techniques for systems in this class. Some of the general results of [9] and [7] were in fact obtained following this venue.

Two other systems investigated in this paper (in Sections 4 and 5) are abstract GAs defined on the domain of natural numbers. Despite their natural and simple definitions, there is no easy way to predict their long-term behavior, and no standard tools exist to this end. Despite the similarity in definitions of the two systems, they display a radically different behavior. The rate of growth of the average fitness is linear for one and exponential for another. The stationary distributions corresponding to the different mating operators are also quite different: Poissonian versus geometrical.

Although the paper is occasionally technical, it contains some general ideas and techniques as well. It is our hope that some general aspects of GAs are clarified by it, and that some of the methods used here prove useful in the future study of GAs.

2. SOME DEFINITIONS

In this section we introduce some notions which are used throughout the paper.

Let $\mathcal{N}$ be a set of some structures, to be called *types*. A probability distribution p on $\mathcal{N}$ shall sometimes be referred to as a *population*; p_l will denote the proportion of the type $l \in \mathcal{N}$ in the population. A *fitness function f* on $\mathcal{N}$ attaches to each $l \in \mathcal{N}$ a nonnegative value, reflecting the quality or the *fitness* of l. The *average* or *expected* fitness of a distribution p of $\mathcal{N}$ is

$$\text{Av}(p) = E_p(f) = \sum_{l \in \mathcal{N}} f(l) p_l.$$

The *fitness operator W*, acting on distributions on $\mathcal{N}$, is completely defined by the fitness function,

$$W(p)_l = \frac{f(l) p_l}{\text{Av}(p)}.$$

Note that it increases the probability of each string proportionally to its fitness. Thus, the proportion of the above-average strings in the population grows, and so does the average fitness. More formally,

$$\text{Av}(W(p)) = \frac{\sum_l f(l)^2 p_l}{\sum_l f(l) p_l} = \frac{E_p(f^2)}{E_p(f)} = E_p(f) + \frac{\text{Var}_p(f)}{E_p(f)} \geq E_p(f) = \text{Av}(p). \quad (1)$$

The *mating operator M* is somewhat more complicated. For each quadruple $(i, j, k, l) \in \mathcal{N}^4$, let β_{ijkl} denote the probability that, in a mating between the two "parent" types i and j, the two types k and l are produced as "offspring." Since any mating must have a definite outcome, it always holds $\sum_{k,l \in \mathcal{N}} \beta_{ijkl} = 1$ for all $i, j \in \mathcal{N}$. We adopt the view that the order of parents and offspring is not significant, so that β is symmetric in i and j, and in k and l, i.e., $\beta_{ijkl} = \beta_{jikl} = \beta_{ijlk}$. If, furthermore, $\beta_{ijkl} = \beta_{klij}$ (meaning that the mating operation is locally time reversible), β is called *symmetric*.

Given two populations p and q, we define their product $\times$ as

$$(p \times q)_l = \sum_{i,j,k \in \mathcal{N}} p_i q_j \beta_{ijkl} \qquad \forall l \in \mathcal{N}. \tag{2}$$

The mating operator M is defined as $M(p) = p \times p$. It should be clear that the definition of M describes a mating process in which two parents are selected independently at random from the population p and mated to form random offspring according to the distribution specified by β. The operator M (or, equivalently, β) defines a (deterministic) quadratic mapping from the set of all of probability distributions over $\mathcal{N}$ to itself. In all the cases discussed in this paper, M always preserves the average fitness.

Given such $\mathcal{N}, \beta, f$ and the initial population $p(0)$, a corresponding genetic algorithm (GA) is completely defined. In each (discrete) moment of time $t = 1, 2, \ldots$ it produces a new population $p(t)$ from the old one,

$$p(t) = WMp(t-1). \tag{3}$$

Sometimes M will be replaced by a "supermating" M^k, obtained by applying M k times.

In order to estimate how close two probability distributions are, we use the standard notion of *variation distance*. For distributions p, q on $\mathcal{N}$,

$$\Delta(p, q) = \frac{1}{2}\|p - q\|_1 = \frac{1}{2}\sum_{i \in \mathcal{N}}|p_i - q_i| = \max_{A \subseteq \mathcal{N}}|p(A) - q(A)|.$$

3. OPTIMIZING A LINEAR FUNCTION ON BINARY STRINGS

Systems studied in this section involve binary strings of a fixed length n, and linear fitness functions $f(s_1, \ldots, s_n) = A + \sum_i a_i s_i$. We assume that f is nonnegative on $\{0,1\}^n$; the condition of nonnegativity is essential for GA, and it can always be met by adding an appropriate positive constant A.

For the most part of the section the *bitwise independent* mating operator is used. The simplest way to describe it is as follows: given a pair of strings $\{s, r\}$, the ith bit of the first offspring is obtained by choosing randomly and independently the ith bit of s or the ith bit of r. The ith bit of the second offspring is the one rejected by the first offspring. The corresponding description in terms of explicitly defined β_{ijkl}s is: β_{ijkl} is $2^{-d(i,j)}$ if $k \cup l = i \cup j$ (viewed as multisets corresponding to strings), and is 0 otherwise. The power $d(i, j)$ stands for the Hamming distance between i and j.

In the concluding remarks of this section, we shall also (briefly) discuss other mating operators.

3.1. Basic Facts about the Bitwise Independent Mating Operator

In this subsection we concentrate on the mating operator, and establish some of its basic properties for future use.

Let s be a random binary string of length n chosen according to the underlying distribution p, and let s_i denote the ith bit of s. Let also $e_i = E_p[s_i]$ be the expected value of s_i. The following claim describes the effect of $\times$ on e_is.

Claim 3.1. *Let p and q be two distributions over binary strings of length n. Then, for every i between 1 and n,*

$$E_{p \times q}(s_i) = \tfrac{1}{2}\big(E_p(s_i) + E_q(s_i)\big).$$

In particular, $M(p)$ and p have the same bit expectations, and thus the same average fitness $Av(p)$.

Proof. Clearly, it suffices to verify that for any four strings s, t, x, y such that the pair $\{s, t\}$ may produce $\{x, y\}$, it holds $s_i + t_i = x_i + y_i$. However, this follows immediately from the definition of $\times$. ∎

Next, we introduce the important notion of the *basic* distribution. A distribution is called *basic* if the corresponding random variables (bits) $\{s_i\}_{i=1}^n$ are entirely independent. We shall use notation $N(\bar{e}) = N(e_1, \ldots, e_n)$ to denote the (unique) basic distribution with bit expectations e_i.

Basic distributions play in our system a role analogous to that of eigenvectors in linear systems; they will be used to produce an analogue of the spectral decomposition. The following proposition establishes the key property of such distributions:

Proposition 3.2. *We have*

$$N(e_1, \ldots, e_n) \times N(d_1, \ldots, d_n) = N\left(\frac{e_1 + d_1}{2}, \ldots, \frac{e_n + d_n}{2}\right).$$

In particular, $n(e_1, \ldots, e_n) \times N(e_1, \ldots, e_n) = N(e_1, \ldots, e_n)$.

Proof. Operator $\times$ acts independently on each bit, and thus preserves the bitwise independence. By the previous claim, the bit expectations in $p \times q$ are the average of those in p and q. ∎

To make the full use of Proposition 3.2, we introduce the notion of *association* between distributions on binary strings of length n, and distributions on $\mathbb{R}^n$ supported on $[0, 1]^n$.

Let q be a distribution on binary strings, given in the form of a convex combination of basic distributions,

$$q = \sum_{v \in [0,1]^n} w_v N(v). \tag{4}$$

The *associated* random variable Y on the n-dimensional unit cube is defined by

$$\Pr[Y = v] \equiv w_v.$$

On the converse, given a discrete random variable Y on $[0,1]^n$, the associated distribution on strings is defined by

$$q = \sum_{v \in [0,1]^n} \Pr[Y = v] \cdot N(v).$$

Notice that a distribution q on strings can always be represented as a convex sum of (degenerate) basic distributions,

$$q = \sum_{r \in \{0,1\}^n} q_r N(r),$$

where q_r is the probability of the string r under q. Thus, any q has an associated Y. In general, however, such Y is not unique, as the representation (4) is not unique. The other direction works better: any Y has a unique associated q.

The convenience of using the notion of association stems from the following fundamental fact, implied by Proposition 3.2:

Proposition 3.3. *Let p, q be distributions on binary n-strings, and let X, Y be corresponding associated random variables. Then $p \times q$ is associated with $(X + Y)/2$.* ∎

The last notion to be introduced in this subsection is that of the *generalized B-moment e_B*. Let p be a distribution on binary strings of length n, and let $B \subseteq \{1, 2, \ldots, n\}$ be a set of indices. Then $e_B = e_B(p)$ is defined as

$$e_B = \Pr_p\left(\bigwedge_{i \in B} (s_i = 1)\right) = E_p\left(\prod_{i \in B} s_i\right),$$

where s_i is the ith bit of a random string s. By convention, $e_\varnothing$ is 1. An important observation is that if X is a random variable associated with p, then

$$e_B = E\left(\prod_{i \in B} X_i\right), \tag{5}$$

where X_i is the value of the ith coordinate of X. We obtain the following formula for the value of $e_B(p \times q)$:

Proposition 3.4. *It holds*

$$e_B(p \times q) = 2^{-|B|} \sum_{K \subseteq B} e_K(p) e_{B-K}(q).$$

Proof. Let X and Y be random variables on $[0,1]^n$ associated with p and q, respectively. By Proposition 3.3 and Eq. (5),

$$e_B(p \times q) = \left(\prod_{i \in B} \left(\frac{X_i + Y_i}{2}\right)\right) = 2^{-|B|} \sum_{K \subseteq B} e_K(p) e_{B-K}(q). \qquad ∎$$

Observe that $\{e_B(p)\}_{B \subseteq \{1,...,n\}}$ can be obtained from $\{p_l\}_{l \in \mathscr{N}}$ by a linear transformation defined by n alone, and that, furthermore, this linear transformation is reversible. Thus, $\{e_B(p)\}_{B \subseteq \{1,...,n\}}$ provides an alternative basis for the space of distributions over the binary strings of length n. In view of Proposition 3.4, this basis is particularly convenient for working with $\times$. As we shall see later, the action of W in this basis can also be described in convenient terms.

We conclude this subsection with the following nice and useful corollary about the effect of M on the covariances. The covariance $\mathrm{Cov}(s_i^+, s_j^+)$ is

$$\mathrm{Cov}(s_i, s_j) = e_{\{i,j\}} - e_i e_j.$$

It is easy to check that in the case of $\{0,1\}$-variables (that is, in our case), a pair of variables is independent if and only if their covariance is 0. Thus, the following corollary implies that every application of M makes s_i and s_j "half" as dependent:

Corollary 3.5. *Let p be a distribution on n-strings, s_i, s_j are random variables corresponding to the ith and the jth bits, and let s_i^+, s_j^+ be the corresponding bits in $p^+ = M(p) = p \times p$. Then*

$$\mathrm{Cov}(s_i^+, s_j^+) = \tfrac{1}{2}\mathrm{Cov}(s_i, s_j).$$

Proof. By Proposition 3.4, $\mathrm{Cov}(s_i^+, s_j^+) = e_{\{i,j\}}^+ - e_i^+ e_j^+ = \tfrac{1}{2}(e_{\{i,j\}} + e_i e_j) - e_i e_j = \tfrac{1}{2}\mathrm{Cov}(s_i, s_j).$ ∎

Remark. Corollary 3.5 is in fact quite simple, and can be proved without any e_Bs by a direct argument (see, e.g., the proof of Lemma 3.14). The e_Bs, however, will be repeatedly used later, and in order to give the reader the opportunity to get more used to them, we prefer the proof given above. ∎

3.2. The Rate of Convergence: Symmetric Case

Having gained some understanding of how the mating operator works, we return to our original system involving both fitness and mating. We start our investigation with a special case (to be called "symmetric"), when the fitness function is just $f(s_1, \ldots, s_n) = \sum s_i$, and the initial distribution $p(0)$ is symmetric, i.e., is invariant under any permutation $\pi \in S_n$ of positions. The symmetry is obviously preserved under both the mating and the fitness operators, and thus all $p(t)$s are symmetric. Clearly, under the action of M and W, the time-t distribution $p(t)$ will eventually converge to the one supported on the all-ones string.

Recall that by (1), it holds that

$$\mathrm{Av}(W(p)) = E(f(s)) + \frac{\mathrm{Var}(f(s))}{E(f(s))}.$$

For our fitness function $f(p) = \sum s_i$, this implies

$$\mathrm{Av}(W(p)) = \mathrm{Av}(p) + \frac{1}{\mathrm{Av}(p)}\left(\sum_i \mathrm{Var}(s_i) + \sum_{i \neq j} \mathrm{Cov}(s_i, s_j)\right). \tag{6}$$

In order to compute the average fitness of $WM(p)$, recall that by Corollary 3.5, M acting on p preserves all e_is [and hence all $\text{Var}(s_i)$], while reducing all the covariances $\text{Cov}(s_i, s_j)$ by a factor of 2. Combining this with (6) we arrive at

$$\text{Av}(WM(p)) = \text{Av}(p) + \frac{1}{\text{Av}(p)}\left(\sum_i \text{Var}(s_i) + \frac{1}{2}\sum_{i \neq j} \text{Cov}(s_i, s_j) \right)$$

$$= \text{Av}(p) + \frac{1}{2}\frac{\sum_i \text{Var}(s_i)}{\text{Av}(p)} + \frac{1}{2}\frac{\text{Var}(\sum_i s_i)}{\text{Av}(p)}$$

$$\geq \text{Av}(p) + \frac{1}{2}\frac{\sum_i \text{Var}(s_i)}{\text{Av}(p)}. \tag{7}$$

By symmetry, all $E_p(s_i) = e_i$ are equal, say to e; so, $\text{Av}(p) = ne$ and $\sum_i \text{Var}(s_i) = ne(1-e)$. Substituting these values in (6), we conclude that

$$\text{Av}(WM(p)) \geq \text{Av}(p) + \frac{1}{2}\left(1 - \frac{\text{Av}(p)}{n}\right).$$

Without risk of confusion, let $\text{Av}(t)$ be shorthand for $\text{Av}(p(t))$. Assuming $p(0)$ does not have all its weight on the all-zeroes string, all $\text{Av}(t)$s are positive. The following theorem provides an upper bound on the time t it takes $\text{Av}(t)$ to get ϵ-close to its limit value n:

Theorem 3.6. $\text{Av}(t) \geq n - \epsilon$, provided that $t \geq 2n \log(1/\epsilon) + 2n \ln(n - \text{Av}(0))$.

Proof. By (7), one has

$$\text{Av}(t+1) \geq \text{Av}(t) + \frac{1}{2}\left(1 - \frac{\text{Av}(t)}{n}\right).$$

Equivalently,

$$n - \text{Av}(t+1) \leq \left(1 - \frac{1}{2n}\right)(n - \text{Av}(t)).$$

Whence, $n - \text{Av}(t) \leq (1 - 1/2n)^t (n - \text{Av}(0))$, and the statement follows.   ∎

How sharp is the upper bound of Theorem 3.6? In order to answer this question, let us first examine the generalized moments e_B of the previous subsection more closely. Observe that due to the symmetry, only the size $|B| = k$ matters now, and all e_Bs corresponding to Bs of size k are equal. Define E_k, to be called the *generalized kth moment* of symmetric p, as

$$E_k = e_B \quad \text{where } |B| = k, \quad \text{or} \quad \text{just } E_k = \text{Pr}_p\left(\bigwedge_{i=1}^{k}(s_i = 1) \right).$$

The inequality (7) becomes loose when the covariances $\text{Cov}(s_i, s_j)$ are positive. Note that $\text{Cov}(s_i, s_j) = E_2 - E_1^2$. How will this quantity behave in the course of

iterative applications of M and W? It turns out that if the initial distribution has a certain nice property, these covariances are never positive; this will yield a lower bound on the rate of convergence of such distributions.

Let $\mathscr{E}(p) = \{E_0 = 1, E_1, E_2, \ldots, E_n\}$ be the sequence of moments of distribution p. Notice that it is monotone nonincreasing.

Definition 1. *A distribution p is called* balanced *if the sequence $\mathscr{E}(p)$ is log-concave, i.e., $E_k^2 \geq E_{k-1}E_{k+1}$, for all $1 \leq k \leq n - 1$. Equivalently,*

$$\frac{E_{k+1}}{E_k} = \Pr_p\left(s_n = 1 \,\Big|\, \bigwedge_{i=1}^{k} (s_i = 1)\right),$$

is monotone nonincreasing in k.

The simplest example of a balanced distribution is a symmetric distribution on singletons, where $E_1 = 1/n$, and all the subsequent moments are 0. The following lemma explains the importance of being balanced:

Lemma 3.7. *The property of being balanced is preserved under both W and M.*

Proof. Let p_i denote the total weight of strings with exactly i 1s in p. Then,

$$E_k = \sum_{i=k}^{n} \frac{\binom{n-k}{i-k}}{\binom{n}{i}} P_i.$$

Using this representation and keeping in mind that $\mathrm{Av}(p) = nE_1$, it is not hard to verify that the kth moment E_k' of $p' = W(p)$ can be expressed in terms of moments of p as follows,

$$E_k' = \sum_{i=k}^{n} \frac{\binom{n-k}{i-k}}{\binom{n}{i}} P_i' = \frac{1}{\mathrm{Av}(W(p))} \sum_{i=k}^{n} \frac{\binom{n-k}{i-k}}{\binom{n}{i}} iP_i = \frac{n-k}{n}\frac{E_{k+1}}{E_1} + \frac{k}{n}\frac{E_k}{E_1}.$$

Also, by Proposition 3.4, the kth moment E_k^+ of $p^+ = M(p)$ is

$$E_k^+ = 2^{-k} \sum_{i=0}^{k} \binom{k}{i} E_i E_{k-i}.$$

The proof of the claim is relatively simple for W. Expressing the moments of $W(p)$ in terms of moments of p, one needs to show that

$$\{kE_k + (n-k)E_{k+1}\}^2$$
$$\geq \{(k-1)E_{k-1} + (n-k+1)E_k\} \cdot \{(k+1)E_{k+1} + (n-k-1)E_{k+2}\}.$$

Expanding the above expressions, and using the log-concavity of $\mathscr{E}(p)$, the claim follows. For M the lemma turns out to be surprisingly difficult. Fortunately, it has

been proven before (for a completely different need and in a slightly more general setting) by Walkup in [14].   ■

Now, for a balanced p all the covariances $\text{Cov}(s_i, s_j) = E_2 - E_1^2$ are nonpositive. Therefore, by preservation of balancedness, if $p(0)$ is balanced, the covariances $\text{Cov}(s_i(t), s_j(t))$ are nonpositive for all $t > 0$. Thus, inequality (7) has the following counterpart for balanced distributions,

$$\text{Av}(t+1) \le \text{Av}(t) + \frac{\sum \text{Var}(s_i)}{\text{Av}(t)} = \text{Av}(t) + \left(1 - \frac{\text{Av}(t)}{n}\right). \tag{8}$$

As an immediate consequence of (8) we obtain a lower bound essentially matching the upper bound of Theorem 3.8:

Theorem 3.8. *If $p(0)$ is balanced, then $\text{Av}(t) \le n - \epsilon$ for all $t \le (n-1)\log(1/\epsilon) + (n-1)\log(n - \text{Av}(0))$.*

3.3. The Rate of Convergence: An Outline of the General Case

We now address the general case, when the fitness function is an arbitrary linear function. The analysis can be performed for this case as well, at the expense of making a small modification in the definition of the system. We give a detailed outline of the argument here, referring the interested reader to [8] for the omitted (simple) technical details. The only two properties of M used in the proof are the preservation of bit expectations and the reduction of covariances by a factor of 2.

Let the fitness function f be $f(s) = A + \sum_i a_i s_i$ and assume that f is nonnegative on $\{0, 1\}^n$. The condition of nonnegativity is essential for GA; it can always be satisfied by choosing an appropriate A. Also define $L = A + \sum |a_i|$. Without loss of generality, one may assume that all a_i are nonnegative: otherwise, replace (for the needs of the analysis alone, and without altering the actual system) s_i by $1 - s_i$, a_i by $-a_i$, and A by $A - a_i$.

In what follows we consider a modified system, where $p(t+1) = WM^r(p(t))$, i.e., $p(t+1)$ is obtained by r successive applications of M (rather than a single one), followed by a single W.

Recall that for any fitness function, the expectation e_i' of the ith bit in $W(p)$ is

$$e_i' = \frac{E(f(s) \cdot s_i)}{E(f(s))}.$$

Keeping in mind that $\text{Var}(s_i) = e_i(1 - e_i)$, the last identity becomes in our case,

$$e_i' = e_i + \frac{a_i}{\text{Av}(p)}(1 - e_i)e_i + \frac{1}{\text{Av}(p)}\text{Cov}\left(\sum_{j \ne i} a_j s_j, s_i\right). \tag{9}$$

Consider the modified system with $p(t+1) = WM^r(p(t))$. Combining Corollary 3.5 with (9) gives for this system,

$$e_i(t+1) = e_i(t) + \frac{a_i}{\text{Av}(t)}(1 - e_i(t))e_i(t) + 2^{-r}R_i(t), \tag{10}$$

where

$$R_i = \text{Cov}\left(\frac{\sum_{j \neq i} a_j s_j(t)}{\text{Av}(t)}, s_i(t)\right) \geq -e_i(t).$$

Theorem 3.9. *Consider the system with $r \geq 4 + \log_2 n/\epsilon$, and assume that in the initial distribution $p(0)$ all e_is are between δ and $1 - \delta$. Then $\text{Av}(t) \geq L(1 - \epsilon)$ for all $t \geq c \cdot (n/\epsilon \log 1/\delta\epsilon)$, where c is some universal constant bounded by 70.*

Proof. Let i be a position such that $a_i \geq \epsilon L/2n$. By (10),

$$e_i(t+1) \geq e_i(t) + \frac{\epsilon}{2n}(1 - e_i(t))e_i(t) - \frac{\epsilon}{16n}e_i(t) \geq e_i(t)\left(1 + \frac{7\epsilon}{16n}\right) - \frac{\epsilon}{2n}e_i^2(t).$$
(11)

It is easily verified (straightforward technical details are omitted) that the number of steps it takes $e_i(t)$ to get from δ to $\frac{1}{2}$ is $O(n/\epsilon \log 1/\delta)$. Observe also that e_i strictly increases until it passes $\frac{1}{2}$, and once it passes $\frac{1}{2}$, it can never again go below this value. Thus we may assume that after such a number of steps, all such e_i are $\geq \frac{1}{2}$. This concludes the first part of the analysis.

The second part of the analysis, i.e., from the moment when i with such a_is holds $e_i(t) \geq \frac{1}{2}$, until the moment it holds $\text{Av}(t) \geq L(1 - \epsilon)$, resembles that of the symmetric case. In fact, the value of r does no longer matter, as long as $r \geq 1$. Arguing along the same lines as in (6) and (7), one gets

$$\text{Av}(t+1) \geq \text{Av}(t) + (1 - 2^{-r})\frac{\sum_i \text{Var}(a_i s_i(t))}{\text{Av}(t)}$$

$$\geq \text{Av}(t) + \frac{1}{2}\frac{\sum a_i^2(1 - e_i(t))e_i(t)}{\text{Av}(t)}.$$

Observe that by our assumption, for all a_i as above it holds $a_i e_i \geq \epsilon L/4n$. Also, the sum $\sum_j a_j$ where j runs over the remaining positions, is bounded by $n \epsilon L/2n = \epsilon L/2$. Therefore, the last inequality implies

$$\text{Av}(t+1) \geq \text{Av}(t) + \frac{1}{2}\frac{\epsilon}{4n}\sum a_i(1 - e_i(t)) - \frac{1}{2}\frac{\epsilon}{4n}\frac{\epsilon L}{2}$$

$$\geq \text{Av}(t) + \frac{\epsilon}{8n}(L - \text{Av}(t)) - \frac{\epsilon^2 L}{16n}.$$

Equivalently,

$$L - \text{Av}(t+1) \leq (L - \text{Av}(t))\left(1 - \frac{\epsilon}{8n}\right) + \frac{\epsilon^2 L}{16n}.$$

Once more, it is a simple matter to verify that the number of steps it takes $\text{Av}(t)$ to get from $L(\frac{1}{2} - \epsilon/2)$ [a lower bound on $\text{Av}(t)$ after the conclusion of the first stage] to $L(1 - \epsilon)$, is $O(n/\epsilon \log 1/\epsilon)$. The straightforward technical details are omitted. Combining the terms obtained in the two stages, we arrive at the desired conclusion. ∎

3.4. More on the Mating Operator

This subsection is devoted to the study of systems defined by the mating operator alone. Our aim is to understand the structure of the stationary distributions and establish the rate of convergence for these systems. While similar results were recently obtained in [7] for more general crossover systems, the analysis given here uses entirely different ideas, and thus, given the rarity and difficulty of rapid mixing results for nonlinear systems, is of an independent interest.

Let $\Delta[*;*]$ denote the variation distance between two distributions. Given a distribution p over binary n-strings, $M^k p$ is obtained by k successive applications of M. As before, e_i is used for the expectation of the ith bit in p, and define $\bar{e} = (e_1, e_2, \ldots, e_n)$. The main result of this subsection is:

Theorem 3.10. *Given an initial distribution p, $M^k p$ converges to $N(\bar{e})$, where $\bar{e}$ is the vector of bit expectations in p. Moreover, the distance between $M^k p$ and $N(\bar{e})$ is bounded by*

$$\Delta\left[M^k p; N(\bar{e})\right] \leq n\sqrt{k}\, 2^{-(k/2)}.$$

Equivalently, $\Delta[M^k p; N(\bar{e})] \leq \epsilon$, provided that

$$k \geq 2\log_2 \frac{n}{\epsilon} + (1 + o(1))\log_2 \log_2 \frac{n}{\epsilon}.$$

Proof. Let us but mention that since the bitwise independent mating increases the independence of the bits without altering their expectations, the fact of convergence to $N(\bar{e})$ should not be surprising.

We start with the observation that the variation distance between two basic distributions can be bounded as follows:

Claim 3.11. *For $u = (u_1, \ldots, u_n)$, $v = (v_1, \ldots, v_n) \in [0,1]^n$, it holds $\Delta[N(u); N(v)] \leq \|u - v\|_1$.*

Proof. By induction on n, notice that

$$
\begin{aligned}
\Delta\big[N(u_1, &\ldots, u_{n-1}, u_n); N(v_1, \ldots, v_{n-1}, v_n)\big] \\
&\leq \Delta\big[N(u_1, \ldots, u_{n-1}, u_n); N(u_1, \ldots, u_{n-1}, v_n)\big] \\
&\quad + \Delta\big[N(u_1, \ldots, u_{n-1}, v_n); N(v_1, \ldots, v_{n-1}, v_n)\big] \\
&= |u_n - v_n| + \Delta\big[N(u_1, \ldots, u_{n-1}); N(v_1, \ldots, v_{n-1})\big].
\end{aligned}
$$

Repeatedly applying the same argument, the claim follows. ∎

Let q be a distribution on strings, and let $\bar{e}$ be the vector of bit expectations of q. The next step is to relate the variation distance between q and the basic distribution $N(\bar{e})$ to the expectation of the l_1 distance between the associated random variables. Notice that $N(\bar{e})$ is associated with the random variable having all its weight on $\bar{e}$.

Claim 3.12. *Let Y be a random variable on $[0,1]^n$ associated with q, and let $\bar{e} \in [0,1]^n$ be a vector. Then*

$$\Delta[q; N(\bar{e})] \leq E(\|Y - \bar{e}\|_1).$$

Proof. By Claim 3.11,

$$E(\|Y - \bar{e}\|_1) = \sum_{v \in [0,1]^n} \Pr[Y = v] \cdot \|v - \bar{e}\|_1$$

$$\geq \sum_{v \in [0,1]^n} \Pr[Y = v] \cdot \Delta[N(v); N(\bar{e})].$$

Recall that $\sum_{v \in [0,1]^n} \Pr(Y = v) N(v)$ is q by definition. Keeping in mind that variation distance between two distributions is half the l_1-norm of their difference (viewed as functions on the underlying space), the triangle inequality for the l_1-norms implies

$$E(\|Y - \bar{e}\|_1) \geq \Delta\left[\sum_{v \in [0,1]^n} \Pr[Y = v] \cdot N(v); N(\bar{e})\right] = \Delta[q; N(\bar{e})]. \qquad \blacksquare$$

We proceed presently with the proof of the theorem. Let X be the random variable associated with the initial distribution p. Let S_{2^k} denote the sum of 2^k independent identical random variables X. According to Proposition 3.3, $2^{-k} S_{2^k}$ is associated with $M^k p$. Call this random variable Y. By Claim 3.12,

$$\Delta[M^k p; N(\bar{e})] \leq E(\|Y - \bar{e}\|_1) = \sum_{i=1}^n E(|Y_i - e_i|),$$

and it suffices to bound the sum $\sum_{i=1}^n E(|Y_i - p_i|)$. Notice that Y_i, the ith coordinate of Y, is the average of the sum of 2^k independent X_is (as Y is of Xes), where each X_i is supported on interval $[0,1]$, and has mean e_i. Applying the Hoeffding inequality (see, e.g., [4]), we conclude that for each i,

$$\Pr(|Y_i - e_i| \geq t) \leq 2e^{-2 \cdot 2^k t^2}.$$

Therefore, for any parameter $t \in (0,1)$,

$$E(|Y_i - e_i|) \leq t + \Pr(|Y_i - e_i| \geq t) \leq t + 2e^{-2^{k+1} t^2}.$$

Choosing $t = 2^{-k/2}\sqrt{k}(1 - \epsilon)$ for a suitably chosen ϵ, we obtain $E(|Y_i - e_i|) \leq 2^{-k/2}\sqrt{k}$. Finally,

$$\Delta[M^k p; N(\bar{e})] \leq E(\|Y - \bar{e}\|_1) = \sum_{i=1}^n E(|Y_i - e_i|) \leq n\sqrt{k} \, 2^{-(k/2)}. \qquad \blacksquare$$

3.5. Two Concluding Remarks

The main object of investigation in Section 3 were the dynamic systems defined on distributions of binary strings. Such systems are an idealization of the systems occurring in practical implementations of genetic algorithms, where one is dealing

with evolution of finite-size multisets of binary strings. To what extent are the methods developed for idealized systems applicable to the practical situation? We claim that at least in the case of linear fitness function and bitwise independent mating operator, these methods can be extended to cover the new situation as well.

We start by reviewing the definitions. The basic object is now a multiset containing N binary strings of length n. The population size N will be maintained under the new operators. The new mating operator (denoted as $\tilde{M}$) is defined as follows: given a multiset of n strings, choose randomly and independently (with replacement) N items (we assume N is even). Pair the first item with the second, the third with the fourth, etc. Each pair produces two offsprings. Consider one such pair: for each position $1 \leq i \leq n$, randomly and independently from others, the ith bit of the first offspring is taken from one random parent, while the ith bit of the second offspring is taken from the other.

The new $\tilde{W}$ acts by forming a new multiset, obtained by choosing N strings randomly and independently, such that the probability of a string to be chosen (at any time) is proportional to its relative frequency times its fitness in the old multiset.

It is convenient to associate with a multiset of N binary strings a distribution $\tilde{p}$ on binary strings, induced by the relative frequencies of strings in the multiset. $\tilde{W}$ and $\tilde{M}$ can be viewed as probabilistic operators defined on $\tilde{p}$.

An important observation relating the old and the new definitions is:

Lemma 3.13. *Let $\tilde{p}$ be the distribution induced by some multiset of N strings. Let z and $\tilde{z}$ stand, respectively, for one of the following four quantities: the expectations of the ith bit or the covariances of the ith and jth bits, in either $W(\tilde{p})$ and $\tilde{W}(\tilde{p})$, or in $M(\tilde{p})$ and $\tilde{M}(\tilde{p})$. Then, comparing the values of z and $\tilde{z}$, one always gets*

$$\Pr(|z - \tilde{z}| \geq \epsilon) \leq 2 e^{-N\epsilon^2}.$$

Proof. It suffices to notice that in fact $\tilde{M}(\tilde{p})$ and $\tilde{W}(\tilde{p})$ are obtained by taking an N-sample from $M(\tilde{p})$ and $W(\tilde{p})$, respectively. Using the Hoeffding inequality (see [4]) to bound the deviation of the sample from its mean (which is 0), we arrive at the desired result. ∎

The meaning of the above lemma is that despite the fact that the system is nonlinear, the errors in e_i and $\text{Cov}(s_i, s_j)$ accumulate but *linearly*. As a consequence of this, the methods similar to those developed previously still apply, only now they should be carried out in a presence of random noise. For N big enough, this noise becomes arbitrarily small. For instance, the finite analogue of (11) is

$$\tilde{e}_i(t+1) \geq \tilde{e}_i(t) + \frac{\epsilon}{2n}(1 - \tilde{e}_i(t))\tilde{e}_i(t) - \frac{\epsilon}{16n}\tilde{e}_i(t) - O(r\delta) \text{ with prob.}$$

$$\geq \Omega\left((1 - 2e^{-N\delta^2})^r\right),$$

where δ is an arbitrary positive value, $\tilde{p}(t+1) = \tilde{W}\tilde{M}^r(\tilde{p}(t))$, and $r = 4 + \log_2 n/\epsilon$. Clearly, for any sufficiently large N a result analogous to Theorem 3.9 can be obtained; we omit here the exact statement and its proof. For further details the reader is referred to [8] and [7]. It is worth a comment here that the asymptotic

behavior of the finite-size system is radically different from the corresponding dynamical system, eventually all the strings in the multiset will be the same. Thus, if one wishes to design a genetic algorithm whose behavior will be similar for a sufficiently long time to that of the idealized GA, N should be chosen sufficiently large.

The second remark is about using different mating operators. Indeed, the most popular mating operator in GA is the *crossover*, defined as follows: given a pair of strings $\{s, r\}$, they are cut at the same random location $0 \le i \le n$, then the head of s is glued together with the tail of r, and vice versa, producing two offspring. Another interesting and practical mating operator was used in [1]; the bits become totally independent in one step, while their expectations are preserved. Do our methods apply for such operators?

The answer is positive. Notice that the only two properties of the mating operator used in the analysis of the system were the preservation of e_i, and the fact that the covariance $\mathrm{Cov}(s_i, s_j)$ was reduced each time by a factor c. However, these properties are quite universal, and shared in particular by both the above-mentioned operators. Indeed, both preserve e_is for obvious reasons. As for the covariances, the latter operator brings them to 0 in one step, while for the former it holds.

Lemma 3.14. *Let p be a distribution on strings, and let $p^+ = M(p)$ where M is the crossover operator. Then*

$$\mathrm{Cov}(s_i^+, s_j^+) = (1 - c_{ij})\mathrm{Cov}(s_i, s_j),$$

where $c_{ij} = |i - j|/n \ge 1/n$.

Proof. Roughly, the argument can be described as follows. Consider a random string s from p^+ (viewed as a random variable), and its two parents $\{r, t\}$ from p. Notice that c_{ij} is exactly the probability that the crossover that produced s from r, t, occurred between positions i and j. Now, if it did occur there, the ith and jth bits of s are uncorrelated, they came from different random strings. Therefore, this case contributes 0 to $\mathrm{Cov}(s_i^+, s_j^+)$. If the crossover did not occur there, the covariance of the ith and jth bits of s is the same as in p. Since the second possibility happens with probability $(1 - c_{ij})$, the lemma follows. For a more detailed proof, the reader is referred to [8]. ∎

The conclusion is that an analogue of Theorem 3.9 holds in this setting as well.

4. THE BINOMIAL SYSTEM ON INTEGERS

The system is defined on the domain of all nonnegative integers, i.e., the types N are $\{0\} \cup \mathbb{N}$. The transition probabilities β_{ijkl} of M are

$$\beta_{ijkl} = \begin{cases} \binom{i+j}{k} 2^{-(i+j)}, & i+j = k+l, \\ 0, & \text{otherwise}. \end{cases}$$

Equivalently, when types i and j meet, the type of the first offspring is distributed binomially on the interval $[0, i+j]$, with parameter $\frac{1}{2}$. Notice that the above transition probabilities β_{ijkl} (probability that the parents $\{i, j\}$ will have offspring $\{k, l\}$) are 0 whenever $i + j \neq k + l$.

The fitness of type "i" will be simply i. Clearly, above M is fitness-preserving.

The operator M is not time reversible, and thus the general theory of [9] does not apply to it. Its analysis is thus of a special interest. On the other hand, it can be viewed as a limit of time-reversible systems. Indeed, consider the bitwise indepen-dent mating operator on the binary strings of length n of the previous section. For a symmetric p (i.e., the probability of a string depends solely on the number of 1s it has), let $p_{[i]}$ be the total weight of all strings with i 1s. Then for a large n, the behavior of $M(p)_i$ and $W(p)_i$ in the present system is very close to that of properly normalized $M(p)_{[i]}$ and $W(p)_{[i]}$ in the binomial system. This analogy provides a possibility to transfer the methods of the previous section (the symmetric case) to the new setting, and is very fruitful. In many cases the reader will find a close resemblance between the results of this and the previous sections. The possibility of using the properties of the better understood time-reversible system for the analysis of a nonreversible one is surprising.

4.1. Some Preliminaries

The infinity of our systems raises a problem we did not have before: it may happen that the expected fitness of a distribution is ∞, in which case all the questions we ask are meaningless. In what follows, we assume that the initial distribution $p(0)$ has a finite support (e.g., only a finite number of types have positive probability). Since this property is preserved under $\times$ and W, all $p(k)$ will have a finite support.

In what follows, we shall often use generating functions, obtained by associating the distribution p with the polynomial $P(x) = \sum_{i \geq 0} p_i x^i$. Operators on polynomi-als corresponding to $\mathrm{Av}(p)$ and $W(p)$ are defined as follows:

Claim 4.1. *Let p be a distribution, and $P(x)$ the associated polynomial. Define*

$$\mathrm{Av}(P(x)) = \frac{P'(1)}{P(1)}; \qquad W(P(x)) = \frac{xP'(x)}{P'(1)}.$$

Then $\mathrm{Av}(P(x)) = \mathrm{Av}(p)$, and the polynomial associated with $W(p)$ is $W(P(x))$. ∎

Observe that for two polynomials P and Q, $\mathrm{Av}(PQ) = \mathrm{Av}(P) + \mathrm{Av}(Q)$, and for a constant c, $\mathrm{Av}(cP) = \mathrm{Av}(P)$.

The following simple inequality will repeatedly prove useful:

Claim 4.2. *Let α, a_i, b_i, c_i be nonnegative reals $(i = 1, 2, \ldots)$ and suppose that for each i,*

$$\frac{a_i}{b_i} \geq \alpha \frac{b_i}{c_i}.$$

Then

$$\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \geq \alpha \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n c_i}.$$

Proof. Define $x_i = \sqrt{a_i/\alpha}$, $y_i = \sqrt{c_i/\alpha}$. By the Cauchy–Schwartz inequality, $(\sum x_i y_i)^2 \leq (\sum x_i^2)(\sum y_i^2)$. Substituting the expressions for x_i, y_i, we deduce the claim. ∎

4.2. The e-Sequence of a Distribution

The following claim shows that the employment of generating functions in the case of binomial systems is indeed appropriate:

Claim 4.3. *Let p, q be distributions, and let $P(x), Q(x)$ be the associated polynomials. Then the polynomial associated with $p \times q$ is $(P \cdot Q)((x + 1)/2)$.*

Proof. Since the $\times$ operator is bilinear, it suffices to check the statement for distributions concentrated on a single point. Let p be concentrated on some type i, and q on some type j. By definition of $\times$, in this case $(p \times q)_k = \binom{i+j}{k} 2^{-(i+j)}$. However, this is exactly the coefficient of x^k in $((x + 1)/2)^j((x + 1)/2)^i$. ∎

Definition 2. *Given a distribution p, define its e-sequence $(e_0, e_1, e_2, \ldots)$ by the generating relation,*

$$P(x) = \sum_{i=0}^{\infty} p_i x^i = \sum_{k=0}^{\infty} \frac{e_k}{k!}(x - 1)^k.$$

In particular, $e_0 = 1$, $e_1 = \mathrm{Av}(p)$. Observe that e_is are nonnegative, and $e_i = 0$ implies that so are all e_k with $k \geq i$. This can be deduced from the nonnegativity of p_is, and the fact that e_k is the value of the kth derivative of $P(x)$ at 1.

Claim 4.4. *Let p, q be two distributions with corresponding e-sequences $\{c_i\}_{i=0}^{\infty}$ and $\{d_i\}_{i=0}^{\infty}$ respectfully. Then*

- *The kth moment of $p \times q$ is*

$$\sum_{i=0}^k \frac{\binom{k}{i}}{2^k} c_i d_{k-i}.$$

- *The kth moment of $W(p)$ is*

$$\frac{e_{k+1}}{e_1} + k\frac{e_k}{e_1}.$$

In particular, e_1', the first moment of $W(p)$, equals $e_2/e_1 + 1$.

Proof. Immediate from Claims 4.3 and 4.1 and the definition of the e-sequence. ∎

4.3. A Lower Bound on $\text{Av}(p(k))$

Proposition 4.5. *The generalized moments of p (denoted as e_i), and those of $W(p)$ (denoted as e'_i) are related as follows,*

$$\frac{e'_k}{e'_{k-1}} \geq \frac{e_k}{e_{k-1}} \quad \text{for all } k \text{ with } e_k > 0.$$

Proof. Expanding the e'_ks in terms of e_ks (by Claim 4.4), we obtain an equivalent statement,

$$\frac{e_{k+1} + ke_k}{e_k + (k-1)e_{k-1}} \geq \frac{e_k}{e_{k-1}},$$

or

$$\frac{e_{k+1}}{e_k} + 1 \geq \frac{e_k}{e_{k-1}}. \tag{12}$$

Expressing the e_ks in the terms of the p_is [recall that by Definition 2, $e_k = P^{(k)}(1)$], we obtain for the left side of (12),

$$\frac{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+1)(i-k)p_i}{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+1)p_i} + 1$$

$$= \frac{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+1)(i-k+1)p_i}{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+1)p_i}.$$

The right side of (12) is

$$\frac{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+2)(i-k+1)p_i}{\sum_{i=0}^{\infty} i(i-1) \cdots (i-k+2)p_i}.$$

Now the left and the right sides can be represented (respectively) in the form,

$$\frac{\sum_{i=k-1}^{\infty} a_i}{\sum_{i=k-1}^{\infty} b_i} \quad \text{and} \quad \frac{\sum_{i=k-1}^{\infty} b_i}{\sum_{i=k-1}^{\infty} c_i},$$

where $a_i/b_i = b_i/c_i = i - k + 1$ for all $i \geq k - 1$. The assumptions of Claim 4.2 hold, and therefore the left side is greater than or equal to the right side, as claimed. ∎

The parameter e_2/e_1 is important, as the average fitness of the next stage distribution depends solely on it [recall that $e'_1 = (e_2/e_1) + 1$]. The following lemma plays the key role in obtaining a lower bound on $\text{Av}(p(n))$:

Lemma 4.6. *Let p be a distribution with $p_0 < 1$. Let $\{e_i\}$, $\{e'_i\}$, and $\{e^*_i\}$ denote the e-sequences of p, $W(p)$, and $MW(p)$, respectively. Then*

$$\frac{e^*_2}{e^*_1} \geq \frac{e_2}{e_1} + \frac{1}{2}.$$

Proof.   It follows that

$$\frac{e_2^*}{e_1^*} \overset{(1)}{=} \frac{e_2' + e_1'^2}{2e_1'} = \frac{1}{2}e_1' + \frac{1}{2}\frac{e_2'}{e_1'} \overset{(2)}{\geq} \frac{1}{2}e_1' + \frac{1}{2}\frac{e_2}{e_1} \overset{(3)}{=} \frac{1}{2}\left(\frac{e_2}{e_1} + 1\right) + \frac{1}{2}\frac{e_2}{e_1} = \frac{e_2}{e_1} + \frac{1}{2}.$$

(1) the first two moments of $MW(p)$ expressed in terms of those of $W(p)$;
(2) by Proposition 4.5;
(3) e_1' of $W(p)$ expressed in terms of e_1, e_2 of p.   ■

Lemma 4.7.   *Let p be some distribution with $p_0 = 0$. Then*

$$\mathrm{Av}(WM(p)) \geq \mathrm{Av}(p) + \tfrac{1}{2}.$$

Proof.   Since $\mathrm{Av}(p)$ equals e_1, it is sufficient to prove the corresponding inequality for e_1s. For every distribution p with $p_0 = 0$, there exists a distribution q such that $W(q) = p$. Let $\{e_i\}$, $\{e_i'\}\{e_i^*\}$, and $\{e_i^+\}$ denote the e-sequences of q, $W(q) = p$, $MW(q)$, $WMW(q) = WM(p)$, respectively. Then

$$e_1^+ \overset{(1)}{=} \frac{e_2^*}{e_1^*} + 1 \overset{(2)}{\geq} \frac{e_2}{e_1} + \frac{1}{2} + 1 \overset{(3)}{=} e_1' + \frac{1}{2}.$$

Where

(1) e_1^+ of $WMW(q)$ expressed in terms of e_1^*, e_2^* of $MW(q)$;
(2) by the previous lemma;
(3) e_1' of $W(q)$ expressed in terms of e_1, e_2 of q.   ■

We can now establish a lower bound:

Theorem 4.8.   *We have*

$$\mathrm{Av}(p(n+1)) \geq \mathrm{Av}(p(1)) + \frac{n}{2}.$$

Proof.   The proof follows by induction from Lemma 4.7. Indeed, since all $p(k)$, $k > 0$, assign zero probability to the type "0" [as $p(k) = WM(p(k-1))$],

$$\mathrm{Av}(p(n+1)) \geq \mathrm{Av}(p(n)) + \frac{1}{2} \geq \cdots \geq \mathrm{Av}(p(1)) + \frac{n}{2}.$$ ■

4.4. An Upper Bound

Now we would like to show that the obtained lower bound is tight, at least for certain kinds of distributions.

Definition 3.   *A distribution p will be called* balanced *if its e-sequence is log-concave, e.g., $e_k^2 \geq e_{k-1}e_{k+1}$ for all natural k.*

We have already encountered balanced distributions in Section 3. Here they are used in a very similar fashion. The simplest example of a balanced distribution is that with $p_1 = 1$. The corresponding e-sequence is $(1, 1, 0, 0, \dots)$.

Balanced distributions have (as before) the following important property:

Lemma 4.9. *If distributions p, q are balanced, so are $W(p)$ and $p \times q$*

The proof (and the statement) are essentially the same as in Lemma 3.7; the second part is proved by Walkup in [14].

Lemmas 4.6 and 4.7 and Theorem 4.8 all have reversed counterparts for balanced distributions:

Lemma 4.10. *Let p be a balanced distribution. Using the notation of Lemma 4.6, we claim that*

$$\frac{e_2^*}{e_1^*} \leq \frac{e_2}{e_1} + 1.$$

Proof. It follows that

$$\frac{e_2^*}{e_1^*} \overset{(1)}{=} \frac{e_2' + e_1'^2}{2e_1'} = \frac{1}{2}e_1' + \frac{1}{2}\frac{e_2'}{e_1'} \overset{(2)}{\leq} \frac{1}{2}e_1' + \frac{1}{2}e_1' = e_1' \overset{(3)}{=} \frac{e_2}{e_1} + 1.$$

(**1**) the first two moments of $MW(p)$ expressed in terms of those of $W(p)$;
(**2**) by preservation of balancedness under W;
(**3**) e_1' of $W(p)$ expressed in terms of e_1, e_2 of p. ∎

Lemma 4.11. *Let $p(0)$ be some nonzero balanced distribution. Then all $p(i)$ and $W(p(i))$ are balanced, and for all natural k it holds*

$$\mathrm{Av}(p(k+1)) \leq \mathrm{Av}(p(k)) + 1.$$

Proof. Let $q = M(p(k-1))$. Let $\{e_i\}$, $\{e_i'\}\{e_i^*\}$, and $\{e_i^+\}$ denote the e-sequences of q, $W(q) = p(k)$, $MW(q)$, and $WMW(q) = p(k+1)$, respectively. Then

$$e_1^+ \overset{(1)}{=} \frac{e_2^*}{e_1^*} + 1 \overset{(2)}{\leq} \frac{e_2}{e_1} + 1 + 1 \overset{(3)}{=} e_1' + 1.$$

Where

(**1**) e_1^+ of $WMW(q)$ expressed in terms of e_1^*, e_2^* of $MW(q)$;
(**2**) by the preservation of balancedness and the previous lemma.
(**3**) e_1' of $W(q)$ expressed in terms of e_1, e_2 of q. ∎

Thus, the bound of Theorem 4.8 is tight:

Theorem 4.12. *For a balanced initial distribution $p(0)$, $\mathrm{Av}(p(n+1)) \leq \mathrm{Av}(p(1)) + n$.*

4.5. Stationary Distributions

As we have mentioned, the present system is *not* time reversible. Therefore the structure of its stationary distributions, and even the fact of their existence, cannot be inferred from the general theory [9]. However, keeping in mind the analogy with the symmetric systems of the previous section, a natural guess would be that the stationary distributions are Poissonian. This is indeed the case.

Definition 4.   *Let $N(\alpha)$ denote the distribution with* $\Pr(i) = (\alpha^i/i!)e^{-\alpha}$ *for* $i = 0, 1, 2, \dots$.

The generating function of $N(\alpha)$ is $e^{\alpha(x-1)}$, and its e-sequence is $(1, \alpha, \alpha^2, \dots)$. Observe that $\mathrm{Av}(N(\alpha)) = e_1 = \alpha$.

Theorem 4.13.   *For any $\alpha \geq 0$ there exists a unique stationary distribution with expected fitness α, and that is $N(\alpha)$. Moreover, if we start with an arbitrary p with $e_1 = \alpha$, and apply M repeatedly, the distribution $M^n p$ converges to $N(\alpha)$ in the variation norm.*

Proof (Sketch).   Using generating functions, we see that $N(\alpha) \times N(\beta) = N((\alpha + \beta)/2)$. Thus, our special distributions are stationary, and behave in a nice way. It would be tempting to represent p as a convex sum of $N(\alpha)$s with $\alpha \geq 0$, which is so convenient for working with M. Unfortunately, such representation is not always possible. Since introducing $N(\alpha)$s with $\alpha < 0$ leads to new technical complications, we prefer to give a more direct (and less precise) argument.

Using the e-sequences, it is not hard to infer from Claim 4.4 that $N(\alpha)$s are the *only* stationary distributions. Therefore, since $e_1 = \alpha$ is preserved under $\times$, the only possible limit point of the set $\{M^n p\}_{n=0}^{\infty}$ is $N(\alpha)$.

The fact of convergence and an estimation of its rate (it may depend on α) can be obtained using the following observation: If the polynomial associated with p is $P(x)$, then, by Claim 4.3, the polynomial corresponding to $M^n(p)$ is

$$\left[P\left(1 + \frac{x-1}{2^n}\right) \right]^{2^n}.$$

A tedious but straightforward elementary calculus argument, combined with induction on i, shows that the ith derivative of the above polynomial at $x = 0$ (that is, p_i) tends to $\alpha^i e^{-\alpha}$, where $\alpha = P'(1) = \mathrm{Av}(p)$. To get an estimate on the rate of convergence to $N(\alpha)$, one needs to consider only a sufficiently long (depending on α) "head" of the distribution. Since all the coefficients are nonnegative and sum up to 1, the contribution of the corresponding "tail" to the variation distance can be effectively bounded provided the head is sufficiently close to convergence.   ∎

5. THE UNIFORM SYSTEM

The last system we are going to investigate differs from the previous one only in definition of the mating operator. The transition probabilities β_{ijkl} are this time,

$$\beta_{ijkl} = \begin{cases} \dfrac{1}{i+j+1}, & i+j = k+l, \\ 0, & \text{otherwise.} \end{cases}$$

That is, when types i and j meet, the type of the first offspring is uniformly distributed on the interval $[0, i+j]$. Notice that the average fitness is preserved under $\times$.

The uniform system is time-reversible, and this fact is reflected in the structure of its stationary distributions.

Despite the superficial resemblance in the definitions of the present and the previous systems, they display a significantly different behavior; the rates of growth of $Av(p(t))$ are exponential versus linear, respectively.

5.1. Some Facts Related to the Mating Operator

We continue to use generating functions in the analysis of the present system as well; their usage, however, poses more difficulties than before.

Claim 5.1. *Let $P(x), Q(x)$ be polynomials associated with p, q. Then the polynomial associated with $p \times q$ is*

$$\frac{\int_1^x P(t)Q(t)\, dt}{x - 1}.$$

Proof. Immediate from the definition of $\times$. ∎

Recall that, as before, the polynomial $W(P(x))$ associated with $W(p)$ is $xP'(x)/P'(1)$.

The following two simple facts will prove useful:

Claim 5.2. *Let $F(x)$ be some function analytic in the neighborhood of 1, and $f(x) = (\int_1^x F(t)\, dt)/(x - 1)$. Then $f^{(n)}(1) = F^{(n)}(1)/(n + 1)$ for every natural n.*

Proof. It suffices to notice that for $F(x) = \sum_{i=0}^{\infty} a_i(x - 1)^i$, $f(x)$ has expansion $f(x) = \sum_{i=0}^{\infty} (a_i/(i + 1))(x - 1)^i$. ∎

Claim 5.3. *Let $Q(x) = \sum_{i=0}^{n} a_i x^i$, $P(x) = (Q(x) - Q(1))/(x - 1)$. Then $P(x) = \sum_{i=0}^{n-1} A_i x^i$, with $A_k = \sum_{i=k+1}^{n} a_i$.*

Proof. We show that

$$(x - 1) \sum_{i=0}^{n-1} A_i x^i = \sum_{i=1}^{n} (A_{i-1} - A_i) x^i - A_0 = \sum_{i=1}^{n} a_i x^i - \sum_{i=1}^{n} a_i$$

$$= \sum_{i=0}^{n} a_i x^i - \sum_{i=0}^{n} a_i = Q(x) - Q(1). \quad ∎$$

Consequences:

  (1) If the coefficients of $Q(x)$ are nonnegative, the coefficients of $P(x)$ are also nonnegative; the sequence $\{A_i\}_{i=0}^{n-1}$ is monotone nonincreasing.
  (2) If $a_i = 0$ in $Q(x)$, then $A_{i-1} = A_i$ in $P(x)$.

5.2. Bounds on $Av(p(n))$

A crude upper bound is easy to get. Suppose that $p(0)$ is concentrated on the first d integers. Then obviously, $p(n)$ is concentrated on the first $2^n d$ integers, implying $Av(p(n)) \le 2^n d = O(2^n)$.

In order to establish an exponential lower bound, we need two lemmas, both rather technical. Let $P_n(x)$ denote the polynomial associated with $p(n)$.

Lemma 5.4. *We have*

$$\mathrm{Av}(P_n(x)) = 1 + \tfrac{2}{3}\mathrm{Av}(P_{n-1}(x)) + \tfrac{2}{3}\mathrm{Av}(P'_{n-1}(x)).$$

Proof. It follows that

$\mathrm{Av}(P_n(x))$

$$= \mathrm{Av}(WM(P_{n-1}(x))) = \mathrm{Av}(x[M(P_{n-1}(x))]') \overset{(1)}{=} 1 + \mathrm{Av}([M(P_{n-1}(x))]')$$

$$= 1 + \frac{[M(P_{n-1}(x))]'' \mid_1}{[M(P_{n-1}(x))]' \mid_1} \overset{(2)}{=} 1 + \frac{\tfrac{1}{3}[2P_{n-1}(1)P''_{n-1}(1) + 2P'_{n-1}(1)P'_{n-1}(1)]}{\tfrac{1}{2}[2P_{n-1}(1)P'_{n-1}(1)]}$$

$$= 1 + \frac{2}{3}\mathrm{Av}(P_{n-1}(x)) + \frac{2}{3}\mathrm{Av}(P'_{n-1}(x)),$$

where

(1) since $\mathrm{Av}(P \cdot Q) = \mathrm{Av}(P) + \mathrm{Av}(Q)$
(2) expressing the derivatives $M(P_{n-1}(x))$ in terms of those of $P_{n-1}(x)$, according to Claim 5.2. ∎

In what follows we assume that the initial distribution $p(0)$ is not concentrated on 0.

Lemma 5.5. $\mathrm{Av}(P'_n(x)) \geq \tfrac{2}{3}\mathrm{Av}(P_n(x))$ *for all* $n \geq 3$.

Proof. Let us represent $P_n(x)$ as $P_n(x) = S(x) + T(x)$, where $T(x)$ is the *tail* of $P_n(x)$, and consists of terms of degree ≥ 5, and $S(x)$ is the *head* of $P_n(x)$, consisting of terms of degree ≤ 4. By Claim 4.2, in order to show $\mathrm{Av}(P'_n(x)) \geq \tfrac{2}{3}\mathrm{Av}(P_n(x))$, it suffices to establish $\mathrm{Av}(S'(x)) \geq \tfrac{2}{3}\mathrm{Av}(S(x))$ and $\mathrm{Av}(T'(x)) \geq \tfrac{2}{3}\mathrm{Av}(T(x))$.

Start with T. Let $T(x) = \sum_{i>4} b_i x^i$. Then

$$\frac{4}{5}\mathrm{Av}(T(x)) = \frac{4}{5}\frac{\sum_{i>4} i b_i}{\sum_{i>4} b_i} \leq \frac{\sum_{i>4} i(i-1)b_i}{\sum_{1>4} i b_i} = \mathrm{Av}(T'(x)).$$

The inequality follows from Claim 4.2, since for all $i > 4$ the corresponding terms satisfy

$$\frac{i(i-1)b_i}{i b_i} = i - 1 \geq \frac{4}{5}i = \frac{4}{5}\frac{i b_i}{b_i}.$$

The proof for S requires more work. It turns out that, up to a multiplicative factor, $S(x)$ can be explicitly computed. Consider the chain,

$$M(P_{n-2}(x)) \overset{W}{\to} P_{n-1}(x) \to P_{n-1}^2(x) \to \int_1^x P_{n-1}^2(t)\, dt$$

$$\to \frac{\int_1^x P_{n-1}^2(t)\, dt}{x-1} = M(p_{n-1}(x)) \overset{W}{\to} P_n(x).$$

What happens to the first five terms of the generating function along this chain? Observe that the coefficients of $M(P_{n-2}(x))$ do not increase with degree, and that the coefficients of x^0, x^1, x^2 are equal. This is a consequence of Claim 5.3, since

$$M(P_{n-2}(x)) = \frac{\int_1^x [P_{n-2}(t)]^2\, dt}{x-1},$$

and $P_{n-2}(x)$, obtained by application of W ($n \geq 3!$), has $a_0 = 0$. Thus, the head of $M(P_{n-2}(x))$ is proportional to $ax^3 + x^2 + x + 1$, with $1 \geq a \geq 0$. Using, in that order, the expressions for $W(P(x))$ (Claim 4.1), for squaring, for integrating, for dividing by $x-1$ (Claim 5.3), and, finally, again for $W(P(x))$, we obtain

$ax^3 + x^2 + x + 1$

$\quad \to 3ax^3 + 2x^2 + x \to (6a + 4)x^4 + 4x^3 + x^2$

$\quad \to \dfrac{6a+4}{5}x^5 + x^4 + \dfrac{1}{3}x^3$, which is proportional to $(18a + 12)x^5 + 15x^4 + 5x^3$,

$\quad \to Ax^5 + (A + 18a + 12)x^4 + (A + 18a + 27)x^3$

$\quad\quad + (A + 18a + 32)(x^2 + x + 1),$

for some $A \geq 0$. Substituting B for $A + 18a$, we proceed with

$$(B + 12)x^4 + (B + 27)x^3 + (B + 32)(x^2 + x + 1)$$

$$\to (4B + 48)x^4 + (3B + 81)x^3 + (2B + 64)x^2 + (B + 32)x.$$

Thus, up to a multiplicative factor, $S(1) = 10B + 225$; $S'(1) = 30B + 595$; $S''(1) = 70B + 1090$ for some $B \geq 0$. Applying once more Claim 4.2, we conclude that

$$\mathrm{Av}(S'(x)) = \frac{S''(1)}{S'(1)} \geq \frac{2}{3}\frac{S'(1)}{S(1)} = \frac{2}{3}\mathrm{Av}(S(x)). \qquad\blacksquare$$

Now we are in the position to prove the main result of this section:

Theorem 5.6. *For all $n \geq 4$, $\mathrm{Av}(p(n)) > \frac{10}{9}\mathrm{Av}(p(n-1))$.*

Proof. Using the generating functions, our aim is to show that $\mathrm{Av}(P_n(x)) > \frac{10}{9}\mathrm{Av}(P_{n-1}(x))$. By Lemma 5.4, $\mathrm{Av}(P_n(x)) > \frac{2}{3}\mathrm{Av}(P_{n-1}(x)) + \frac{2}{3}\mathrm{Av}(P'_{n-1}(x))$. By Lemma 5.5, $\mathrm{Av}(P'_{n-1}(x)) \geq \frac{2}{3}\mathrm{Av}(P_{n-1}(x))$. Combining the two results, we obtain

$$\mathrm{Av}(P_n(x)) > \frac{2}{3}\mathrm{Av}(P_{n-1}(x)) + \frac{4}{9}\mathrm{Av}(P_{n-1}(x)) = \frac{10}{9}\mathrm{Av}(P_{n-1}(x)). \qquad\blacksquare$$

Consequently,

Theorem 5.7. $\mathrm{Av}(p(n)) > (\frac{10}{9})^{n-4}\mathrm{Av}(p(4)) \geq (\frac{10}{9})^{n-4}\mathrm{Av}(p(0))$.

5.3. Stationary Distributions

Theorem 5.8. *For any $\alpha \geq 0$ there exists a unique stationary distribution with expected fitness α, and that is the geometrical distribution with parameter $\alpha/(1+\alpha)$.*

Proof. Since the present system is time reversible, it falls into the general framework of [9], and thus a stationary distribution π must satisfy

$$\pi_i \pi_j = \pi_k \pi_l \quad \text{for } i+j=k+l.$$

Consequently it must be geometric, with the parameter determined by the average fitness of $p(0)$ (which remains invariant). This can also be verified directly, the generating function of a stationary distribution π must satisfy the integral equation,

$$\frac{\int_1^x f^2(t)\,dt}{x-1} = f(x),$$

and thus also

$$f^2(x) = (x-1)f'(x) + f(x),$$

with boundary conditions $f(1) = 1$, $f'(1) = \alpha$. The solution of this equation is

$$f(x) = \frac{1}{(1+\alpha)-\alpha x} = \sum_{i=0}^{\infty} \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^i x^i,$$

as expected.   ∎

A stronger version of the above theorem can also be shown; if one starts with an arbitrary p with $\mathrm{Av}(p) = \alpha$, and applies M repeatedly, the distribution $M^n p$ converges to the appropriate geometrical distribution. The fact of convergence can be inferred from the entropy rising property of the system, and that for each α there is a *unique* stationary distribution $\pi(\alpha)$ with average fitness α (the reader is referred to [9] for the general theory of such systems). However, the rate of convergence remains an open problem. A new approach for this kind of problem (successfully applied to two classical systems and yielding new proofs) was recently developed in [2]. We believe it can be used for the present system as well, but leave this for a future study.

6. CONCLUSION

This completes the analysis of the three systems discussed. The point made was that GAs can in principle be rigorously analyzed.

Perhaps the most important lesson learned from the obtained results is that the interaction between the mating and the fitness operators is much easier to understand if M has some nice properties, and in particular, does not decrease the average fitness of the distribution. Although the latter property appears to be a reasonable one to require, it is rarely met in applications of GAs. The reason is that such "nice" mating operators do not always exist, and even if they do, finding them would require a good deal of understanding of the structure of the space of solutions.

If one is so lucky as to find a nice M for his problem, it seems to be advisable to apply M sufficiently many times in a row before each application of W. The reason is clear: while the effect of a single mating operator is often rather hard to quantify, a sufficiently long sequence of mating operations applied in a row brings the distribution close to a stationary state, which should have more structure, and is easier to understand.

The second lesson is that one should always seek for a basis in which the mating operator M takes a simple form: finding such a basis is likely to lead to a better understanding of M, and thus of the entire system.

ACKNOWLEDGMENTS

REFERENCES

[1] E. Baum, D. Boneh, and C. Garrett, On genetic algorithms, Proc COLT, 1995, pp. 230–239.

[2] E. Carlen and M. Carvalho, Strict entropy production bounds and stability of the rate of convergence to equilibrium for the Boltzmann equation, preprint, 1992.

[3] R. Cerf, "An asymptotic theory for genetic algorithms," Artificial evolution: Lecture notes in computer science, Springer-Verlag, Berlin/New York, 1996, Vol. 1063, pp. 37–53.

[4] W. Hoeffding, Probability inequalities for sums of bounded random variables, Amer Stat Assoc 27 (1963), 13–29.

[5] J. Holland, Adaptation in natural and artificial systems, University of Michigan Press, Ann Arbor, MI, 1975.

[6] J.R. Koza, Genetic programming, MIT Press, Cambridge, MA, 1992.

[7] Y. Rabani, Y. Rabinovich, and A. Sinclair, A computational view of population genetics, Random Structures Algorithms 12 (1998), 313–334.

[8] Y. Rabinovich, Ph.D. Thesis, Quadratic dynamical systems and theoretical properties of genetic algorithms, Hebrew University, Jerusalem, Dec. 1993.

[9] Y. Rabinovich, A. Sinclair, and A. Wigderson, Quadratic dynamical systems, Proc 33rd IEEE FOCS, 1992, pp. 304–313.

[10] Y. Rabinovich and A. Wigderson, Analysis of a simple genetic algorithm, Proc 4th Int Conf Genetic Algorithms, 1991, pp. 215–221.

[11] G.J.E. Raulings, Ed., Foundation of genetic algorithms, Morgan Kaufmann, San Mateo, CA, 1991.

[12] R.L. Riolo, Survival of the fittest bits, Scientific American 267, July 1992.

[13] A.L. Samuel, "Some studies in machine learning using the game of checkers," Computers and thought, McGraw-Hill, New York, 1963.

[14] D.W. Walkup, Polya sequences, binomial convolution and unions of random sets, J Appl Probab 13 (1976), 76–85.