Genetic Fitness Optimization Using Rapidly Mixing Markov Chains

Paul Vitányi*
CWI and University of Amsterdam

Abstract. A notion of highly probable fitness optimization through evolutionary computing runs on small size populations in a very general setting is proposed. This has applications to evolutionary learning. Based on rapidly mixing Markov chains, the approach pertains to most types of evolutionary genetic algorithms, genetic programming and the like. For systems having associated rapidly mixing Markov chains and appropriate stationary distributions the new method finds optimal programs (individuals) with probability almost 1. Algorithmically, the novel approach prescribes a strategy of executing many short computation runs, rather than one long computation run. Given an arbitrary evolutionary program it may be infeasible to determine whether its associated matrix is rapidly mixing. In our proposed structured evolutionary program discipline, the development of the program and the guaranty of the rapidly mixing property go hand in hand. We conclude with a tentative toy example.

1 Introduction

Theoretical performance analysis of genetic computing often uses unbounded or exponential population sizes [1, 10, 15, 16]. Results obtained in this way may not be directly applicable to real practical problems where we always have to deal with a bounded (small) population size [4, 14].

Considering small population sizes it is at once obvious that the size and constitution of the population may have a major impact on the evolutionary development of the population. We aim to establish a fast feasible speed of convergence to a distribution of populations from which we can obtain by Monte Carlo sampling an optimal type individual with high probability.

The method we propose is applicable to a wide range of genetic computing models which includes genetic algorithms on strings and genetic programming on trees, and so forth. The computational properties are analyzed in terms of a finite Markov chain where the states correspond to finite populations. The transition probability between two states is induced by the selection, reproduction, and fitness rules, [10, 16, 7].

* Partially supported by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, and by NWO through NFI Project ALADDIN under Contract number NF 62-376. Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl; WWW: http://www.cwi.nl/p̃aulv/

Under mild conditions guarantying ergodicity the Markov chain converges to a stationary distribution over the set of states, and hence over the set of reachable populations. From this stationary distribution we can sample a set of populations. If populations with optimal individuals have high enough probability of occurring, then it is almost certain that we can locate such an individual.

We analyze the speed of convergence using recent results, [11, 2, 3] in rapid mixing of Markov chains, to obtain an overall feasible process (for example, polynomially many runs of polynomially long evolutions of small size populations.)

Outline of a Discipline of Genetic Optimization. An evolutionary computation process with positive mutation rates corresponds to the development of an ergodic Markov chain whose states are the populations and which converges to a desired stationary probability distribution over the sample space of which the elements are populations, [10]. In each single run of the evolutionary computation we find with some probability a particular population. Repeat this process so as to obtain a large enough sample of populations drawn from the stationary distribution, and subsequently determine the fittest individuals from each such population. We shall show that if proper conditions can be guarantied, then this process finds a global optimally fit individual with probability almost one. Our analysis shows that for certain genetic computations using a large number of short runs is provably a good strategy as opposed to one long run. In practice, several researchers observed earlier that it pays to restart on a new population when the evolution takes a unpromising direction, for example [6, 4]. [2] To the author's knowledge, we provide the first formal method of genetic fitness optimization (applicable to restricted classes of GA, GP, and related optimization problems) together with a rigorous analysis demonstrating that this strategy is *guarantied* to work with *high probability*, rather than intuitive heuristic or ad hoc arguments.

The efficiency of this technique in any application depends crucially on the rate of convergence of the Markov chain. Since the number of states is typically very very large, the chain should reach equilibrium after each particular evolution has only explored a tiny fraction of the state space. Chains with this property are called rapidly mixing.

Towards Structured Genetic Computing. To actually use the method we have to find a structured methodology to set up the genetic system (selection, reproduction, fitness) such that the resulting Markov chain is rapidly mixing, and, moreover, such that the types with sufficiently high fitness will be obtained by Monte Carlo sampling with sufficiently high probability from the (close to) final stationary state distribution. What we have in mind is a design methodology to develop a genetic system satisfying these requirements from the specifications of the problem statement. This is a tall order, but on the positive side we recall that similar techniques have been used successfully in combinatorial counting, statistical physics, and combinatorial optimization, [11] and certain quadratic dynamic processes related to genetics of infinite populations [13].

[2] Also J. Koza and L.J. Eshelman have algorithms that specifically restart automatically (GP, CHC, respectively), as do many others.

Genetic Learning. Genetic fitness optimization has direct applications to genetic learning. In genetic learning the individuals in a population represent alternative hypotheses concerning the phenomenon being learned. The fitness of the individual can be related to prediction success, classification success of a test set, and so on.

2 Formal Model

Assume, there are r possible types of individuals, say a set $\Omega = \{1, \ldots, r\}$. Such individuals can be strings, trees or whatever—our discussion is so general that the precise objects don't matter. The genetic system tries to solve an optimization problem in the following sense. Each individual in Ω is graded in terms of how well it solves the problem the genetic system is supposed to solve, expressed as a function f which maps Ω to some grading set G. For example, G can be the real interval $[0,1]$. With $f(i)$ be the *fitness* of type i, the normalized fitness of individual i is

$$\hat{f}(i) = \frac{f(i)}{\sum_{j \in \Omega} f(j)}.$$

To fix thoughts, we use fitness proportional selection where selection of individuals from a population is according to probability related to the product of frequency of occurrence and fitness. That is, in a population $P = (P(1), \ldots, P(r))$ of size n, where type i occurs with frequency $P(i) \geq 0$ with $\sum_{i=1}^{r} P(i) = n$, we have probability $p(i)$ to select individual i (with replacement) for the cross-over defined by

$$p(i) = \frac{f(i)P(i)}{\sum_{j \in \Omega} f(j)P(j)}.$$

Define a Markov chain $\mathcal{M}$ with states P consisting of nonnegative integer r-vectors of which the individual entries sum up to the population size n. The number of states of $\mathcal{M}$ is

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

see [10]. The associated transition matrix $Q = (Q_{i,j})$ is a $N \times N$ matrix where the entry $Q_{i,j}$ is the probability that the kth generation will be P^j given that the $(k-1)$st generation is P^i.

A general closed form expression for transition probabilities for simple GA's is derived in [10] and its asymptotics to steady state distributions as population size increases is determined. In [7] it is observed that the mentioned closed form expression allows expression of 'expected waiting time until global optimum is encountered for the first time', 'expected waiting time for first optimum within some error tolerance of global optimum', and 'variance in such measures from run to run', and so on, but no further analysis is provided. Instead, initial experimental work is reported.

Rather than using the general closed form expression, for our purpose it is more useful to focus on particular illustrative example GAs. This involves no loss of generality, since the special types of selection, cross-over, mutation, and generation of the next population which we consider here is for the sake of definiteness and generalize (within the level of detail we need for the remainder of this paper) to any reasonable other evolutionary computing procedure, be it GA, GP, or other types.

Let us calculate the transition probabilities $Q_{i,j}$ for a very simple selection process. This consists of sampling two individuals from a population P, remove these two individuals from P, let them produce two offspring which are inserted again in the population resulting in a population P'.

The transition probability of

$$P \rightarrow P'$$

with $P'(i) = P(i) - 1, P'(j) = P(j) - 1, P'(k) = P(k) + 1, P'(h) = P'(h) + 1$ (replacing a pair of individuals i, j by k, h) is given by

$$A(i,j,k,h) = p(i)p(j)B(i,j,k,h). \tag{2}$$

Here $B(i,j,k,h)$ incorporates both the mutation probability and the cross-over probability of producing k and h from i and j. (Other known selection, mutation, and generation rules lead to similar transition probabilities between states $P \rightarrow P'$ of the Markov chain.) We desire $\mathcal{M}$ to be ergodic. To ensure this we assume that the mutation probability of obtaining k and h from i and j is positive, even without any cross-over. Choose $B(i,j,k,h) > 0$ such that $\mathcal{M}$ is ergodic.

3 Sample Size Versus Divergence of Evolutionary Trajectories

The phenomenon of the influence of population size and sample size on the drifting apart of evolutionary trajectories can be illustrated in a simplified setting ignoring fitness selection (without loss of generality). For convenience therefore, we ignore the actual populations and deal with the occurrence probability of types rather than with the number of occurrences of types in a population.

3.1 Infinite Sample

Given a distribution p and a transition matrix $B(i,j,k,h)$, let the transformation $p' = g(p)$ be defined by

$$p'(h) = \sum_{i,j,k} p(i)p(j)B(i,j,k,h) \tag{3}$$

Consider a (not necessarily finite) population of individuals, each individual being of some type $i \in \{1, \ldots, r\}$. Let $p(i)$ be the probability of selecting an individual of type i. When individuals of type i mate with individuals of type j, then this produces individuals k and h with probability $B(i,j,k,h)$. Assuming that a mating of i and j must result in some offspring k and h means that

$\sum_{k,h} B(i,j,k,h) = 1$. The resulting probability of h is $p'(h)$ above. We transform a probability distribution p in a probability distribution p' where the probabilities of the types i are given by $p'(i)$. This implies that

$$\sum_{i,j} p(i)p(j)B(i,j,k,h) = p'(k)p'(h)$$

$$\sum_{k,h} p'(k)p'(h) = 1.$$

A distribution ρ is called an *equilibrium distribution* (with respect to transformation g) if $g(\rho) = \rho$. In [1, 10] for simple GA with fitness selection, and [12] for more general quadratic dynamical systems but without fitness selection, the following convergence property is derived.

Theorem 1. *The sequence $p^0, p^1, \ldots$ with $p^t = g^t(p^0)$ $(t \geq 0)$ converges to an equilibrium distribution $\lim_{t\to\infty} p^t = \rho$.*

As is easy to see, if the populations involved are infinitely large, then essentially evolution develops deterministically according to Equation 3. But if the populations are very small, then chance selections can cause great divergence of evolution of populations. It is useful to look at a quantification of this distinction. Neither [10] or [12] explores in an explicit quantitative manner the divergence of trajectories of individual runs based on population sizes. They rather focus on the issue that as the population size grows, the divergence of possible trajectories gets progressively smaller. In the limit, for infinite populations, the generations in the run converge to the expected trajectory for smaller populations. Clearly, if all trajectories are in a small envelope around the expected trajectory, then the expected trajectory is a good predictor for what happens with an individual run. If moreover the expected trajectory corresponds to the infinite population trajectory, as in the system analyzed in [10], then the analysis of the infinite system tells us what to expect from our individual bounded population evolution.

In contrast, in Section 4 we give an example using bounded size populations where the *expected trajectory* is *completely different* from *all individual trajectories.* If the individual trajectories of bounded populations diverge wildly, the expected trajectory may not predict anything about what happens to an individual run, as the analysis below shows. We analyze quantitatively in some detail bounds on this divergence based on population sizes. In the next sections we introduce a methodology based on rapidly mixing Markov chains. This methodology guaranties that enough probability is concentrated on trajectories of 'fast' runs ending with small populations containing globally optimal programs. Those optimal programs are then retrieved using Monte Carlo sampling of such populations and evaluating the fitness of the programs they contain.

3.2 Large Sample

If we draw an infinite sample from the distribution p then we produce by Equation 3 a new population distributed according to p'. Let us now analyze the case where we draw a sample of cardinality s. If s is large enough then we can approximate p by the resulting frequencies within an ϵ fraction of each probability $p(i)p(j)$ (for $p(i)p(j)$ bounded away from 0). Quantitatively this works out as follows.

Let there be r types of individuals in Ω. If $s(i,j)$ is an outcome of the random variable measuring number of outcome pairs i,j in s trials then by Chernoff's bound, see for example [8],

$$\Pr(|s(i,j) - p(i)p(j)s| > \epsilon s) < 2e^{-\epsilon^2 s/4p(i)p(j)(1-p(i)p(j))}.$$

If $p'()$ is the next probability distribution as defined above, and $\hat{p}'()$ is the probability distribution we obtain on the basis of the outcome $s(i,j)$ in drawing s examples, for all types $h \in \Omega$,

$$\begin{aligned} |p'(h) - \hat{p}'(h)| &\leq \sum_{i,j,k} |s(i,j)/s - p(i)p(j)| B(i,j,k,h) \\ &\leq \epsilon \sum_{i,j,k} B(i,j,k,h) = \epsilon r, \end{aligned}$$

with an exponentially (in s) vanishing probability of error. Let $\epsilon = \epsilon'/r$ and sample size $s(r) \geq r^3 p(i)p(j)$ to ensure that $\epsilon^2 s/p(i)p(j) > r$, for all $i,j \in \Omega$. Then, with vanishing probability (at most e^{-r}) the next generation probability $p'(h)$ will be at least ϵ'-far from probability $\hat{p}'(h)$ for all r-many types h. [3]

From this result it may be possible to estimate how fast the trajectory of a population of such sample size possibly strays away from the evolution of the infinite system. This may be the subject of a future paper. In a more restricted setting of a quadratic cross-over system with $\Omega = \{0,1\}^l$ reference [13] shows that the probability distribution of an infinite quadratic cross-over system (without fitness selection) stays for the duration of an evolution of t generations in an appropriate sense close to that of a population of size $O(n^2 t)$ initially drawn randomly from the infinite population.

3.3 Small Sample

We draw *one* pair of individuals from p and do a cross-over to obtain a single new pair of individuals. Then $p(i)p(j)B(i,j,k,h)$ is the probability of obtaining k,h. In a computational run of a genetic algorithm this means that we have distribution p before the cross-over and with probability $p(i)p(j)B(i,j,k,h)$ obtain a distribution $\hat{p}'$ after the cross-over, resulting from first drawing i,j to mate, and then producing k,h and updating p accordingly to obtain $\hat{p}'$. If we assume that we have a fixed population of size n where each type in Ω has probability $1/r$ and the mating consumes two elements of types i,j and produces two elements, one each of types k,h, then there are at most r^4 distributions which can be obtained from p this way by Equation 3.

[3] Cubic results appearing in [5, 4] are cubic in the population size n and refer to different issues.

Repeating this procedure, we potentially obtain in t steps up to r^{4t} distributions. For sensitive systems this means that we very quickly realize all N possible different distributions.

Each such distribution $\hat{p}'$ is obtained with some probability $p(i)p(j)B(i,j,k,h)$. Now the right hand side of Equation 3 gives the *expectation* that an individual of type h results from the mating of two individuals of p and gets added to the population. It says therefore something about the expected increase of h-type individuals but does not give the value of any particular new $\hat{p}'(h)$.

4 Divergence of Trajectories of Individual Runs

In the large sample case above we can make the error in distribution entries exponentially small for polynomial size samples. However, the error increases exponentially fast with the generations again, loosing all accuracy in a polynomial number of generations.

In the small sample case above we explicitly consider the exponential explosion of different future possibilities, rather in a Markov chain format. There remains the task of measuring the specific properties of the thus produced ensemble of runs—like average and variance.

The analysis of [10, 12] does not deal with individual runs of a genetic algorithm, but rather with the sequence of expectations over all individual runs of the system. To give an analogy, if we look at the expected outcome of the tth coin flip in a sequence of independent flips of a fair coin, then $p(0) = p(1) = 1/2$. However, in any individual sequence of outcomes the nth outcome is either 0 or 1.

However, the expectation says not too much about what actually happens. To see this, consider a *dictatorial coin* which gives a first outcome 0 or 1 with fair odds. However, afterwards it always gives the same outcome. So it either produces an all 0 run or an all 1 run with equal probabilities. The expectation of obtaining a 0 at the tth trial is 1/2. However, in actual fact at the tth $(t > 1)$ trial we have either probability 1 or probability 0 for outcome 0. In terms of the above formalism, initially, $p(0) = p(1) = 1/2$.

Consider the following similar example in quadratic systems format. Let the B transformation be given by (a more complicated example can also satisfy the requirement that B be symmetric, locally reversible, and a-periodic as required in [12])

$i\ j\ k\ h$	$B(i,j,k,h)$
0 0 0 0	1
0 1 0 0	1
1 0 1 1	1
1 1 1 1	1
· · · ·	0

Then, we have

$$p \rightarrow p' \rightarrow p'' \rightarrow \cdots$$

with

$$p'(0) = p(0)p(0)B(0,0,0,0) + p(0)p(1)B(0,1,0,0) = 1/2$$
$$p'(1) = p(1)p(0)B(1,0,1,1) + p(1)p(1)B(1,1,1,1) = 1/2$$

But if we interpret the system in an algorithmic manner, then the following scenario is perhaps more realistic. Initially, the population is $P^0 = \{0,1\}$ with $p(0) = p(1) = 1/2$. Draw two elements from this population with replacement. Execute a cross-over according to the B matrix. We then obtain either a population $P^1 = \{0,0\}$ with $p^1(0) = 1, p^1(1) = 0$, or a population $P^2 = \{1,1\}$ with $p^2(0) = 0, p^2(1) = 1$, with probability 1/2 for obtaining P^1 or P^2:

$$p^0 \to p^1 \to p^1 \to \cdots$$
or
$$p^0 \to p^2 \to p^2 \to \cdots$$

(Another possibility is that we draw two elements from the population without replacement. So we draw either the sequence 0, 1 or the sequence 1, 0 with equal probability. For this example these outcomes lead again by B to the two different populations with associated distributions.)

Now either of the resulting P^1 or P^2 with associated distributions will forever remain entirely stable under B transformations. We have not derived an explicit population associated with the distribution p'. In fact, the interpretation of p' is

$$p'(0) = (p^1(0) + p^2(0))/2$$
$$p'(1) = (p^1(1) + p^2(1))/2$$

5 Towards a Discipline of Evolutionary Programming

The upshot of the considerations so far is that with a limited size populations the variation in evolutions and resulting populations is very great. In practice we always deal with very limited size populations such as say 500 individuals. The question arises how to overcome the problem that an individual evolution can become trapped in an undesirable niche of populations with non-optimal individuals. The answer is that we need to randomize over the evolutions so that we inspect a sample of evolutions resulting in a set of populations, one of which almost surely contains a global optimally fit individual. The latter is easy if the set of inspected evolutions is large enough, say almost as large as the set of individuals to be evaluated with respect to fitness, thus begging the question.

However, it turns out that we can make such an approach feasible in the rigorous sense that one constructs the evolutionary system so that its associated Markov chain satisfies certain conditions, which in turn guaranty that both with probability almost one a global optimally fit individual occurs in one of the evolved populations, and the entire set of evolutions required takes only feasible time. We aim for two properties.

1. The stationary distribution of populations the associated Markov chain of the evolutionary process converges to concentrate a sufficient amount of probability on populations containing maximally fit individuals.
2. The Markov chain of the evolutionary process converges sufficiently fast to the stationary distribution: it is *rapidly mixing*, see Appendix A for precise definitions. Here it is only important to know that the evolution of such a chain approximates the stationary distribution suitably fast close enough.

The question of rapid mixing, property 2, can be satisfied by having the evolutionary system satisfy some structural properties. Such properties can, at least in principle if not in practice, always be taken care of while implementing the evolutionary system by choosing the selection rules, cross-over operator, and mutation rules appropriately. These requirements are covered in the next section.

The question of probability concentration, property 1, is more subtle, and it is not yet clear how to generally go about it, even in principle. However, similar approaches in approximating hard combinatorial optimization problems (approximating the permanent which is deterministically a #P-hard problem, that is, at least as hard as NP-complete problems) have successfully resolved this issue.

5.1 Structural Requirements of the Discipline

We have delegated the technical section on rapidly mixing property to Appendix A. The precise definitions can perhaps be ignored at first reading until after the main approach has become clear in outline.

Our task in designing evolutionary systems turns out to be two fold. We need to design the system such that for the associated Markov chain

1. the second largest eigenvalue $\lambda_{\max}$ is suitably bounded away far enough from 1 so that the Markov chain is rapidly mixing (Definition 10 of the Appendix A); and
2. the stationary distribution π gives probability greater than $1/q(n)$, where $q(n)$ is polynomial, to states s which contain nonzero frequencies of the fittest types.

For Item 1 it is required that the matrices are (i) irreducible, and (ii) have nonnegative entries. Since the only matrices we consider are *stochastic* where the entries are transition probabilities, (ii) is in our case easy to satisfy up to the 'suitable' condition. Since we only deal with ergodic matrices, and (i) is required for ergodicity, Item 1 is always satisfied in our case. Ergodicity is immediate if we have a positive mutation probability of transforming i into j for each pair of types i, j. Hence by proper choice of the genetic system leading to suitable transition probabilities inducing a rapidly mixing Markov chain one can satisfy Item 1 in construction of an evolutionary system. It is perhaps less easy to see whether it is feasible to satisfy Item 2 in each particular case, without knowing the optimal individual a priori. However, a similar approach for approximating very hard combinatorial optimization problems, [11], worked out fine.

Assume that we have defined our evolutionary system satisfying Items 1, 2. The program we use is then as follows. Repeat a polynomial number of times:

1. From a start state evolve through a polynomial number of generations;
2. From the population vector select the fittest individual.

Suppose we are dealing with $\Omega = \{0,1\}^l$. Then there are $r = 2^l$ different types. Suppose further we are dealing with populations of size n, say with $n \leq 500$, then the number of possible populations is given by Equation 1 and hence the associated Markov chain has that number N of states.

We repeat the above process a logarithmic number of times in N, that is, a polynomial number of times in terms of r. Then, the probability that we end with a vector which contains the fittest types with frequency 0 vanishes as a negative exponential. Thus, we have discovered a new interesting paradigm.

Paradigm. Running the program longer than a polynomial number of generations will not significantly change the closeness of the state distribution to the stationary distribution in the Markov chain. We can only guaranty that we find a state (vector) containing an optimal fit individual with probability $1/q(n)$. However, polynomially repeating this procedure implies Monte Carlo sampling which almost surely discovers the individual with optimal fitness.

6 A Toy Rapidly Mixing Genetic Algorithm

Consider a toy evolutionary problem as follows. We consider a population of size $\sqrt{l}$ and very simple crossover only and some mutation. This example already illustrates adequately the rapid mixing phenomenon. The genetic algorithm G is defined as follows. The set of all program types is $\Omega = \{0,1\}^l$ with l fixed, even, and large enough for the following analysis to hold. The *fitness* of a program $\omega \in \Omega$ with $\omega = \omega_1\omega_2 \ldots \omega_l$ is given by the function

$$f(\omega) = 1 \text{ if } \sum_{i=1}^{l} = l/2 \text{ and } 1/2 \text{ otherwise .}$$

The *starting population* P^0 at time $t_0 = 0$ contains $\sqrt{l}$ copies of the individual $00\ldots0$; its cardinality (number of elements in P^0) is $\sqrt{l}$. We express the *frequency* of a string ω in a population P by $\#_\omega(P)$. That is, $\#_{00\ldots0}(P^0) = \sqrt{l}$ and $\#_\omega(P^0) = 0$ for $\omega \neq 00\ldots0$

The transition of one population to the next generation (population) is as follows. To avoid problems of periodicity, we add self-loop probability of $1/2$ to each state (that is, population). Note that this also dispenses with the problem of negative eigenvalues. Consequently, there is probability $1/2$ that the state changes using crossover and mutation, and there is probability $1/2$ that it stays the same. The probability $p(\omega)$ of selecting a string ω from a population P is

$$p(\omega) = \frac{\#_\omega(P)f(\omega)}{\sum_{\omega\in\Omega} \#_\omega(P)f(\omega)}. \tag{4}$$

In the selection phase we select two individuals in P, say ω^i, ω^j, according to these probabilities, and with probability 1/2 we perform a crossover and mutation on each (and with probability 1/2 we do nothing). The crossover operator interchanges a single bit of ω^i with the corresponding bit of ω^j. It selects the single bit position with uniform probability $1/l$. Subsequently, we mutate each offspring by flipping a single bit with uniform probability $1/l$ chosen from the positions 1 through l. (If $i = j$ then the cross-over doesn't do anything and the two mutations may result in 0,1, or 2 bit flips of ω_i.) We first prove that G is rapid mixing by reducing it to the following problem.

Consider a system G' where the initial state is a binary l-vector. At each step uniformly at random select a bit position of the current l-vector and flip that bit with fifty-fifty probability to produce the next l-vector. Then G' is a Markov chain where the states are the binary l-vectors.

Lemma 2. *The chain G' is rapid mixing with r.p.d. at most ϵ within $O(l^2(l + \log(1/\epsilon)))$ steps.*

For a proof see [11], pp. 63–66. This system is an almost uniform generator for Ω, using singleton populations, where it suffices to use an arbitrary starting singleton population. In terms of GA's it is single-bit mutation. Our example involves single-bit mutation, single-bit cross-over, and selection. The reader is advised that this is only a cosmetic change to make the example look more like a 'realistic' GA. Our toy example G is essentially the example G' as in Lemma 2. To reduce G to G', consider the vectors in successive generations $P^0, P^1, \ldots$ to maintain their identity. If $P_t = \{\omega^{t,1}, \ldots, \omega^{t,\sqrt{l}}\}$ for $t > 0$ and in the selection phase we select indices i, j, then $\omega^{t+1,k} = \omega^{t,k}$ for $0 \leq k \leq \sqrt{l}$ and $k \neq i, j$, or $\omega^{t+1,h}$ results from $\omega^{t,h}$ (the 'same vector') by at most two bit flips for $h = i, j$.

Lemma 3. *Let $\epsilon > 0$ and $T(l) = O(l^{5/2}(l + \log(1/\epsilon)))$. For each $t \geq T(l)$, with probability at least $1 - 1/T(l)$, for each $\omega \in \{0,1\}^{\sqrt{l}}$ every l-vector $\omega^{0,j} \in P^0$ has probability $(1 \pm \epsilon)/2^l$ of being changed into $\omega^{t,j} = \omega$ in t generations of G.*

Proof. For a fraction of at least $1 - 1/t$ of all runs of $t > \sqrt{l}$ steps of a population of $\sqrt{l}$ elements, for each index j the vector $\omega^{\cdot,j}$ is selected with frequency of at least

$$t/(4\sqrt{l}) \pm O(\sqrt{t/\sqrt{l}\log t}) \tag{5}$$

in the selection phases of the generating process. This is shown similar to the statistical analysis of 'block frequencies' of high Kolmogorov complexity strings in [8], Theorem 2.15.

Namely, consider t throws of a $\sqrt{l}$-sided coin, each throw constituting the selection of an index. There are $2^{(t \log l)/2}$ possible sequences x of t outcomes. Hence, the maximal Kolmogorov complexity is given by $C(x|t,l) \leq (t \log l)/2 + O(1)$. Moreover, there is a fraction of at least $1 - 1/t$ sequences x which has $C(x|t,l) \geq (t \log l)/2 - \log t$. Divide each such x in consecutive blocks of length $(\log l)/2$.

Let $\#j(x)$ denote the number of occurrences of each of the $\sqrt{l}$ elementary index j outcomes in x. Then, by [8] p. 132,

$$|\#j(x) - t/\sqrt{l}| \leq \sqrt{\frac{(\log l)/2 + \log t + O(1)}{\sqrt{l} \log e} 4t}.$$

The extra factor '4' in the denominator of Equation 5 is an overestimate accounting for the the following facts: (i) some individuals have fitness 1/2, and the other individuals have fitness 1; (ii) with probability 1/2 the population is not changed at all.

Following the same vector in the successive generations, consider each time it is selected for a cross-over and mutation. At such times, with fifty-fifty probability either nothing is done or the vector incurs (i) a bit flip in a position which was selected uniformly at random because of the cross-over (or no bit flip if the bits in that position of the two parents happened to be the same), followed by (ii) a bit flip in a position selected uniformly at random because of the mutation. From the viewpoint of the individual vector it simply describes a trajectory of length as given in Equation 5 of the singleton l-vector in Lemma 2 Substitute t in Equation 5 by $T(l)$ as in the statement of the lemma. By Lemma 2 the lemma is proven.

Corollary 4. *It follows that G is a rapidly mixing Markov Chain with a uniform stationary distribution.*

Lemma 5. *The probability of finding a population with an optimally fit element in t runs is at least $1 - 2e^{-\alpha t}$ with $\alpha = c/(16(1-c))$, for the fixed constant c given in Equation 6.*

Proof. There are $\binom{l}{l/2} \approx 2^l/\sqrt{\pi l/2}$ strings with fitness 1. Hence a fraction of at most

$$(1 - 1/\sqrt{\pi l/2})^{\sqrt{l}} < e^{-\sqrt{2/\pi}}$$

populations of size $\sqrt{l}$ contain no such strings. This means that a constant fraction of at least

$$c = 1 - e^{-\sqrt{2/\pi}}, \tag{6}$$

of the populations of size $\sqrt{l}$ contain at least one string of fitness 1.

Consider each run of $T(l)$ generations an experiment with a *success* outcome if the final population contains an individual with fitness 1. Let the number of successes in t trials be $s(t)$. Then, with β defined as

$$\beta = \Pr\{|s(t) - ct| > \delta t\}$$

we have

$$\beta < 2e^{-\delta^2 t/(4c(1-c))},$$

by Chernoff's bound. For $\delta = c/2$ we know that the number of successes $s(t) > 0$ with probability at least $1 - \beta$.

Theorem 6 (Rapidly Mixing GA Algorithm). *Let ϵ and $T(l)$ be as in Lemma 2 and let α be as in Lemma 5. Repeat t times: run G for $T(l)$ generations. This procedure uses $O(T(l) \cdot t)$ elementary steps consisting of the generation from one population to the next population. (With $t = l$ this is a low degree polynomial in l and ϵ). The probability of finding an optimal element exceeds*

$$1 - 2e^{-\alpha t},$$

where $\alpha > 0$, that is, with probability of failure which vanishes exponentially fast with rising t.

Proof. By Lemmas 3, 5.

A Appendix: Basics of Markov Chains and Rapid Mixing

A sequence of random variables $(X_i)_{t=0}^{\infty}$ with outcomes in a finite state space $T = \{0, \ldots, N-1\}$ is a *finite state time-homogeneous Markov chain* if for any ordered pair i, j of states the quantity $p_{i,j} = \Pr(X_{t+1} = j | X_t = i)$ called the *transition probability* from state i to state j, is independent of t. If $\mathcal{M}$ is a Markov chain then its associated *transition matrix* is $P = (p_{i,j})_{i,j=0}^{N-1}$. The matrix P is non-negative and *stochastic*, its row sums are all unity.

For $s \in \mathcal{N}$, the *s-step transition matrix* is the power $P^s = (p_{i,j}^s)$ with $p_{i,j}^s = \Pr(X_{t+s} = j | X_t = i)$, independent of t. Denote the distribution of X_t by the row vector $\pi^t = (\pi_0^t, \ldots, \pi_{N-1}^t)$ with $\pi_i^t = \Pr(X_t = i)$. If π^0 denotes the initial distribution then $\pi^t = \pi^0 P^t$ for all $t \in \mathcal{N}$. Often we have $\pi_i^0 = 1$ for some i (and 0 elsewhere) in which case i is called the *initial state*.

The chain is *ergodic* if there exists a distribution π over T with strictly positive probabilities such that

$$\lim_{s \to \infty} p_{i,j}^s = \pi_j,$$

for all $i, j \in T$. In this case we have that $\pi^t = \pi^0 P^t \to \pi$ pointwise as $t \to \infty$, and the limit is independent of π^0. The *stationary distribution* π is the unique vector satisfying $\pi P = \pi$, where $\sum_i \pi_i = 1$; that is, the unique normalized left eigenvector of P with eigenvalue 1. Necessary and sufficient conditions for ergodicity are that the chain should be *irreducible*, for each pair of states $i, j \in T$ there is an $s \in \mathcal{N}$ such that $p_{i,j}^s > 0$ (j can be reached from i in a finite number of steps); and *aperiodic*, the $\gcd\{s : p_{i,j}^s > 0\} = 1$ for all $i, j \in T$.

An ergodic Markov chain is (*time-*)*reversible* iff either (and hence both) of the following equivalent conditions hold.

- For all $i, j \in T$ we have $p_{i,j}\pi_i = p_{j.i}\pi_j$. That is, in a stationary distribution, the expected number of transitions per unit time from state i to state j and from state j to state i are equal. For any ergodic chain, if π is a positive vector satisfying above condition and the normalization condition $\sum_i \pi_i = 1$, then the chain is reversible and π is its stationary distribution.

- The matrix $D^{1/2}PD^{-1/2}$ is symmetric, where $D^{1/2}$ is the diagonal matrix $\mathrm{diag}(\pi_0^{1/2}, \ldots, \pi_{N-1}^{1/2})$ and $D^{-1/2}$ is its inverse.

For example, if in Equation 2 $A(i,j,k,h) = A(k,h,i,j)$ for all $P \in T$ and all $i,j,k,h \in \Omega$, then the induced Markov chain defined there is reversible. Of course, the converse need not be true.

Consider now the problem of sampling elements from the state space, assumed very large, according to the stationary distribution π. The desired distribution can be realized by picking an arbitrary initial state and simulating the transitions of the Markov chain according to probabilities $p_{i,j}$, which we assume can be computed locally as required. As the number t of simulated steps increases, the distribution of the random variable X_t will approach π. The rate of approach to stationary can be expressed in the following time-dependent measure of deviation from the limit. For any non-empty subset $U \subseteq T$, the *relative pointwise distance* (r.p.d.) over U after t steps is given by

$$\Delta_U(t) = \max_{i,j \in U} \frac{|p_{i,j}^t - \pi_j|}{\pi_j}.$$

This way, $\Delta_U(t)$ is the largest relative distance between π^t and π at any state $j \in U$, maximized over all possible states in U. The parameter U allows us to specify relevant portions of the state space. In case $U = T$ we will omit the subscript and write Δ instead of Δ_U.

The stationary distribution π of an ergodic chain is the left eigenvector of P with associated eigenvalue $\lambda_0 = 1$. Let $\lambda_1, \ldots, \lambda_{N-1}$ with $\lambda_i \in \mathcal{C}$ (the complex numbers) be the remaining eigenvalues (not necessarily distinct) of P. By the standard Perron-Frobenius theory for non-negative matrices these satisfy $|\lambda_i| < 1$ for $1 \le i \le N-1$. The transient behavior of the chain, and hence its rate of convergence, is governed by the magnitude of the eigenvalues λ_i. In the reversible case, the second characterization above implies that the eigenvalues of P are those of the symmetric matrix $D^{1/2}PD^{-1/2}$ and so are all real. This leads to the following clean formulation of above dependence, [11].

Lemma 7. *Let P be the transition matrix of an ergodic reversible Markov chain, π is stationary distribution, and $\lambda_0 = 1, \ldots, \lambda_{N-1}$ its (necessarily real) eigenvalues. Then, for any nonempty subset $U \subseteq T$ and all $t \in \mathcal{N}$ the relative pointswise distance over U satisfies*

$$\Delta_U(t) \le \frac{\lambda_{\max}^t}{\min_{i \in U} \pi_i},$$

where $\lambda_{\max}$ is the largest value in $|\lambda_1|, \ldots, |\lambda_{N-1}|$.

Lemma 8. *With the notation of Lemma 7 the relative pointswise distance over T satisfies*

$$\Delta(t) \ge \lambda_{\max}^t$$

for all even $t \in \mathcal{N}$. Moreover, if all eigenvalues of P are non-negative, then the bound holds for all $t \in \mathcal{N}$.

Therefore, provided π is not extremely small in any state of interest, the convergence of the reversible chain will be rapid iff $\lambda_{\max}$ is suitably bounded away from 1. Such a chain is called *rapid mixing.*

If we order the eigenvalues $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} > -1$ then $\lambda_{\max} = \max\{\lambda_1, |\lambda_{N-1}|\}$ and the value of λ_{N-1} is significant only if some eigenvalues are negative. The oscillatory behavior associated with negative eigenvalues cannot occur if each state is equipped with sufficiently large self-loop probability. It is enough to have $\min_j p_{j,j} \geq 1/2$. To see this, let I_N denote the $N \times N$ identity matrix and consider the non-negative matrix $2P - I_N$, whose eigenvalues are $\mu_i = 2\lambda_i - 1$. By Perron-Frobenius, $\lambda_i \geq -1$ for all $i \in T$ which implies that $\mu_{N-1} \geq 0$.

Lemma 9. *With the notation of Lemma 7, let the eigenvalues of P be ordered $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} > -1$. Then the modified chain with transition matrix $P' = 1/2(P - I_N)$, with I_N as above, is also ergodic and reversible with the same stationary distribution, and its eigenvalues λ'_i similarly ordered satisfy $\lambda'_{N-1} > 0$ and $\lambda'_{\max} = \lambda'_1 = 1/2(1 + \lambda_1)$.*

Following [11] we define rapid mixing.

Definition 10. Given a family of ergodic Markov chains $\mathcal{M}(x)$ parametrized on strings x over a given alphabet. For each such x, let $\Delta^{(x)}(t)$ denote the r.p.d. of $\mathcal{M}(x)$ over its entire state space after t steps, and define the function $\tau^{(x)}(\epsilon)$ from the positive reals to the natural numbers by

$$\tau^{(x)}(\epsilon) = \min\{t : \Delta^{(x)}(t') \leq \epsilon \text{ for all } t' \geq t\}.$$

We call such a family *rapidly mixing* iff there exist a polynomial bounded function q such that $\tau^{(x)}(\epsilon) \leq q(|x|, \log \epsilon^{-1})$ for all x and $0 < \epsilon \leq 1$.

The question arises whether the approach to rapidly mixing Markov chains can be generalized from *reversible* chains to *non-reversible* chains. This was affirmatively settled in [9] and another treatment was later given in [3]. See the short discussion in [11].

In the applications to evolutionary programming, x will be a problem instance and the state space of $\mathcal{M}(x)$ will include solution sets $R(x)$ of some relation R.

References

1. T.E. Davis and J.C. Principe, A simulated annealing like convergence theory for the simple genetic algorithm, *Proc. 4th Int'l Conf. Genet. Algorithms*, Morgan Kaufmann, 1991, 174-181.
2. P. Diaconis and D. Stroock, Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability*, 1:1(1991), 36-61.
3. J.A. Fill, Eigenvalue bounds on convergence to stationary for nonreversible Markov chains, with an application to the exclusion process, *The Annals of Applied Probability*, 1:1(1991), 62-87.

4. D.E. Goldberg, Sizing populations for serial and parallel genetic algorithms, *Proc. 3rd Int'nl Conf. Genet. Algorithms*, Morgan Kaufmann, 1989, 70-79.
5. J.H. Holland, *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, 1975.
6. K.A. de Jong and W.M. Spears, Using genetic algorithms to solve NP-complete problems, *Proc. 3rd Int'nl Conf. Genet. Algorithms*, Morgan Kaufmann, 1989, 124-132.
7. K.A. de Jong, W.M. Spears, and D.F. Gordon, Using Markov chains to analyze GAFOs, draft version for FOGA94 Proceedings, 1994.
8. M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York, 1993.
9. M. Mihail, Conductance and convergence of Markov chains: a combinatorial treatment of expanders. *Proc. 30th IEEE Symp. Foundat. Comput. Science*, 1989, 526-531.
10. A.E. Nix and M.D. Vose, Modeling genetic algorithms with Markov chains, *Annals of Mathematics and Artificial Intelligence*, 5(1992), 79-88.
11. A. Sinclair, *Algorithms for Random Generation and Counting, A Markov Chain Approach*, Birkhäuser, 1992.
12. Y. Rabinovitch, A. Sinclair and A. Wigderson, Quadratic dynamical systems, *Proc. 33rd IEEE Symp. Foundat. Comput. Science*, 1992.
13. Y. Rabani, Y. Rabinovitch, and A. Sinclair, A computational view of population genetics, *Proc. 27th ACM Symp. Theor. Comput.*, 1995.
14. C.R. Reeves, Using genetic algorithms with small populations, *Proc. 5th Int'nl Conf. Genet. Algorithms*, Morgan Kaufmann, 1993, 92-99.
15. G. Rudolf, Convergence analysis of canonical genetic algorithms, *IEEE Trans. Neural Networks*, 5:1(1994), 96-101.
16. J. Suzuki, A Markov chain analysis on simple genetic algorithms, *IEEE Trans. Systems, Man, and Cybernetics*, 25:4(1995), 655-659.