
Recursive Conditional Schema Theorem, Convergence and Population Sizing in Genetic Algorithms

Riccardo Poli

School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK
R.Poli@cs.bham.ac.uk

Abstract

In this paper we start by presenting two forms of schema theorem in which expectations are not present. These theorems allow one to predict with a known probability whether the number of instances of a schema at the next generation will be above a given threshold. Then we clarify that in the presence of stochasticity schema theorems should be interpreted as conditional statements and we use a conditional version of schema theorem backwards to predict the past from the future. Assuming that at least x instances of a schema are present in one generation, this allows us to find the conditions (at the previous generation) under which such x instances will indeed be present with a given probability. This suggests a possible strategy to study GA convergence based on schemata. We use this strategy to obtain a recursive version of the schema theorem. Among other uses, this schema theorem allows one to find under which conditions on the initial generation a GA will converge to a solution on the hypothesis that building block and population fitnesses are known. We use these conditions to propose a strategy to attack the population sizing problem. This allows us to make explicit the relation between population size, schema fitness and probability of convergence over multiple generations.

1 INTRODUCTION

Schema theories can be seen as *macroscopic* models of genetic algorithms. What this means is that they state something about the properties of a population at the next generation in terms of macroscopic quantities (like schema fitness, population fitness, number of individuals in a schema, etc.) measured at the current generation. These kinds of models tend to hide the huge number of degrees of freedom of a GA behind their macroscopic quantities (which are typically averages over the population or subsets of it). This typically leads to relatively simple equations which are easy to study and understand. A macroscopic model does not have to be an approximate or worst-case-scenario model, although many schema theorems proposed in the past were so. These properties are in sharp contrast to those shown by *microscopic* models, such as Vose's model [Nix and Vose, 1992, Vose, 1999] (see also [Davis and Principe, 1993, Rudolph, 1997c, Rudolph, 1997a, Rudolph, 1994, Rudolph, 1997b]), which are always exact (at least in predicting the expected behaviour of a GA) but tend to produce equations with enormous numbers of degrees of freedom.

The usefulness of schemata and the schema theorem has been widely criticised (see for example [Chung and Perez, 1994, Altenberg, 1995, Fogel and Ghozeil, 1997, Fogel and Ghozeil, 1998]). While some criticisms are really not justified as discussed in [Radcliffe, 1997, Poli, 2000c] others are reasonable and apply to many schema theories.

One of the criticisms is that schema theorems only give *lower bounds* on the *expected value* of the number of individuals sampling a given schema at the next generation. Therefore, they cannot be used to make predictions over multiple generations.¹ Clearly, there is some truth in this. For these reasons, many researchers nowadays believe that schema theorems are nothing more than trivial tautologies of no use whatsoever (see for example [Vose, 1999, preface]). However, this does not mean that the situation cannot be changed and that all schema theories are useless. As shown by recent work [Stephens and Waelbroeck, 1997, Stephens and Waelbroeck, 1999, Poli, 1999b, Poli, 2000b, Poli, 2000a] schema theories have not been fully exploited nor fully developed. For example, recently Stephens and Waelbroeck [Stephens and Waelbroeck, 1997, Stephens and Waelbroeck, 1999] have produced a new schema theorem which gives an exact formulation (rather than a lower bound) for the expected number of instances of a schema at the next generation in terms of macroscopic quantities.² Stephens and Waelbroeck used this result as a starting point

¹If the population is assumed to be infinite, then the expectation operator can be removed from schema theorems. So, the theorems can be used to make long-term predictions on schema propagation. However, these predictions may become easily very inaccurate due to the fact that typically schema theorems provide only lower bounds.

²The novelty of this result is not that it can predict exactly how many individuals a schema will contain on average in the future. This could be calculated easily with microscopic models, e.g. using Vose's model by explicitly monitoring the number of instances of a given schema in the expected trajectory of the GA using an approach such as the one in [De Jong et al., 1995]. The novelty of Stephens and Waelbroeck's result, which will be presented in a simplified form in the next section, is that it makes explicit how and with which probability higher order schemata can be assembled from lower order ones, and it does this by using only a small number of macroscopic quantities.

for a number of other results on the behaviour of a GA over multiple generations on the assumption of infinite populations.

Encouraged by these recent developments, we decided to investigate the possibility of studying GA convergence using schema theorems and information on schema variance. This paper presents the results of this effort.

The paper is organised as follows. After describing the assumptions on which the work is based (Section 2), two forms of schema theorem in which expectations are not present are introduced, in Section 3. These theorems allow one to predict *with a known probability* whether the number of instances of a schema at the next generation will be above a given threshold. Then (Section 4), we clarify that in the presence of stochasticity schema theorems should be interpreted as conditional statements and we use a conditional version of schema theorem backwards to predict the past from the future. Assuming that at least x instances of a schema are present in one generation, this allows us to find the conditions (at the previous generation) under which such x instances will indeed be present with a given probability. As discussed in Section 5, this suggests a possible strategy to study GA convergence based on schemata. Using this strategy a conditional recursive version of the schema theorem is obtained (Section 6). Among other uses, this schema theorem allows one to find under which conditions on the initial generation the GA will converge to a solution in constant time with a known probability on the hypothesis that building block and population fitnesses are known, as illustrated in Section 7. In Section 8 we use these conditions to propose a strategy to attack the population sizing problem which makes explicit the relation between population size, schema fitness and probability of convergence over multiple generations. We draw some conclusions and identify interesting directions for future work in Section 9.

2 SOME ASSUMPTIONS AND DEFINITIONS

In this work we consider a simple generational binary GA with fitness proportionate selection, one-point crossover and no mutation with a population of M bit strings of length N . The crossover operator produces one child (the one whose left-hand side comes from the first parent).

One of the objectives of this work is to find conditions which guarantee that a GA will find at least one solution with a given probability (perhaps in multiple runs). This is what is meant by *GA convergence* in this paper. Let us denote such a solution with $S = b_1 b_2 \dots b_N$.

We define the *total transmission probability* for a schema H , $\alpha(H, t)$, as the probability that, at generation t , every time we create (through selection, crossover and mutation) a new individual to be inserted in the next generation such an individual will sample H [Poli et al., 1998]. This quantity is important because it allows to write an exact schema theorem of the following form:

$$E[m(H, t + 1)] = M\alpha(H, t), \tag{1}$$

where $m(H, t + 1)$ is the number of copies of the schema H at generation $t + 1$ and $E[\cdot]$ is the expectation operator.

In a binary GA in the absence of mutation the total transmission probability is given by the following equation (which can be obtained by simplifying the results in

[Stephens and Waelbroeck, 1997, Stephens and Waelbroeck, 1999] or, perhaps more simply, as described in the following paragraph):

$$\alpha(H, t) = (1 - p_{xo})p(H, t) + \frac{p_{xo}}{N-1} \sum_{i=1}^{N-1} p(L(H, i), t)p(R(H, i), t) \quad (2)$$

where p_{xo} is the crossover probability, $p(K, t)$ is the selection probability of a schema K at generation t , $L(H, i)$ is the schema obtained by replacing all the elements of H from position $i + 1$ to position N with “don’t care” symbols, $R(H, i)$ is the schema obtained by replacing all the elements of H from position 1 to position i with “don’t care” symbols, and i varies over the valid crossover points.³ For example, if $H = **1111$, then $L(H, 1) = *****$, $R(H, 1) = **1111$, $L(H, 3) = **1***$, and $R(H, 3) = ***111$. If one, for example, wanted to calculate the total transmission probability of the schema $*11$, the previous equation would give:

$$\begin{aligned} \alpha(*11, t) &= (1 - p_{xo})p(*11, t) + \frac{p_{xo}}{2} \left(p(***, t)p(*11, t) + p(*1*, t)p(**1, t) \right) \\ &= \left(1 - \frac{p_{xo}}{2} \right) p(*11, t) + \frac{p_{xo}}{2} p(*1*, t)p(**1, t). \end{aligned}$$

It should be noted that Equation 2 is in a considerably different form with respect to the equivalent results in [Stephens and Waelbroeck, 1997, Stephens and Waelbroeck, 1999]. This is because we developed it using our own notation and following the simpler approach described below. Let us assume that while producing each individual for a new generation one flips a biased coin to decide whether to apply selection only (probability $1 - p_{xo}$) or selection followed by crossover (probability p_{xo}). If selection only is applied, then there is a probability $p(H, t)$ that the new individual created sample H (hence the first term in Equation 2). If instead selection followed by crossover is selected, let us imagine that we first choose the crossover point and then the parents (which is entirely equivalent to choosing first the parents and then the crossover point). When selecting the crossover point, one has to choose randomly one of the $N - 1$ crossover points each of which has a probability $1/(N - 1)$ of being selected. Once this decision has been made, one has to select two parents. Then crossover is executed. This will result in an individual that samples H only if the first parent has the correct left-hand side (with respect to the crossover point) *and* the second parent has the correct right-hand side. These two events are independent because each parent is selected with an independent Bernoulli trial. So, the probability of the joint event is the product of the probabilities of the two events. Assuming that crossover point i has been selected, the first parent has the correct left-hand side if it belongs to $L(H, i)$ while the second parent has the correct right-hand side if it belongs to $R(H, i)$. The probabilities of these events are $p(L(H, i), t)$ and $p(R(H, i), t)$, respectively (whereby the terms in the summation in Equation 2, the summation being there because there are $N - 1$ possible crossover points). Combining the probabilities all these events one obtains Equation 2.

³The symbol L stands for “left part of”, while R stands for “right part of”.

3 PROBABILISTIC SCHEMA THEOREMS WITHOUT EXPECTED VALUES

In previous work [Poli et al., 1998, Poli, 1999b] we emphasised that the process of propagation of a schema from generation t to generation $t + 1$ can be seen as a Bernoulli trial with success probability $\alpha(H, t)$ (this is why Equation 1 is so simple). Therefore, the number of successes (i.e. the number of strings matching the schema H at generation $t + 1$, $m(H, t + 1)$) is binomially distributed, i.e.

$$\Pr\{m(H, t + 1) = k\} = \binom{M}{k} [\alpha(H, t)]^k [1 - \alpha(H, t)]^{M-k}.$$

This is really not surprising. In fact it is a simple extension of the ideas, originally formulated mathematically in [Wright, 1931], at the basis of the well-known Wright-Fisher model of reproduction for a gene in a finite population with non-overlapping generations.

So, if we know the value of α , we can calculate exactly the probability that the schema H will have at least x instances at generation $t + 1$:

Theorem 1 Probabilistic Schema Theorem (Strong Form). *For a schema H under fitness proportionate selection, one-point crossover applied with probability p_{co} and no mutation,*

$$\Pr\{m(H, t + 1) \geq x\} = \sum_{k=x}^M \binom{M}{k} [\alpha(H, t)]^k [1 - \alpha(H, t)]^{M-k},$$

where, $\alpha(\cdot)$ is defined in Equation 2 and the probability of selection of a generic schema K is $p(K, t) = \frac{m(K, t) f(K, t)}{M \bar{f}(t)}$, where $f(K, t)$ is the average fitness of the individuals sampling K in the population at generation t , while $\bar{f}(t)$ is the average fitness of the individuals in the population at generation t .

In theory this theorem could be used to find conditions on α under which, for some prefixed value of x , the r.h.s. of the previous equation takes a value y . This is very important since it is the first step to find sufficient conditions for the conditional convergence of a GA, as shown later. Unfortunately, there is one problem with this idea: although the equation $\sum_{k=x}^M \binom{M}{k} [\alpha(H, t)]^k [1 - \alpha(H, t)]^{M-k} = y$ can be solved for α , as reported in [Poli, 1999a], its solution is expressed in terms of Γ functions and the hypergeometric probability distribution. So, it is really not easy to handle.

As briefly discussed in [Poli, 1999b], one way to remove this problem is *not* to fully exploit our knowledge that the probability distribution of $m(H, t + 1)$ is binomial when computing $\Pr\{m(H, t + 1) \geq x\}$. Instead we could use Chebyshev's inequality [Spiegel, 1975],

$$\Pr\{|X - \mu| < k\sigma\} \geq 1 - \frac{1}{k^2},$$

where X is a stochastic variable (with *any* probability distribution), $\mu = E[X]$ is the mean of X and $\sigma = \sqrt{E[(X - \mu)^2]}$ is its standard deviation.

Since $m(H, t + 1)$ is binomially distributed, $\mu = E[m(H, t + 1)] = M\alpha(H, t)$ and $\sigma = \sqrt{M\alpha(H, t)[1 - \alpha(H, t)]}$. By substituting these equations into Chebyshev's inequality we obtain:

Theorem 2 Probabilistic Schema Theorem (Weak Form). *For a schema H under fitness proportionate selection, one-point crossover applied with probability p_{co} and no mutation,*

$$\Pr\{m(H, t + 1) > M\alpha(H, t) - k\sqrt{M\alpha(H, t)(1 - \alpha(H, t))}\} \geq 1 - \frac{1}{k^2} \quad (3)$$

for any fixed $k > 0$, with the same meaning of the symbols as in Theorem 1.

Unlike Theorem 1, this theorem provides an easy way to compute a value for α such that $m(H, t + 1) > x$ with a probability not smaller than a prefixed constant y , by first solving the equation

$$M\alpha - k\sqrt{M\alpha(1 - \alpha)} = x \quad (4)$$

for α (as described in the following section) and then substituting $k = \frac{1}{\sqrt{1-y}}$ into the result.

It is well-known that for most probability distributions Chebychev inequality tends to provide overly large bounds, particularly for large values of k . Other inequalities exist which provide tighter bounds. Examples of these are the one-sided Chebychev inequality, and the Chernoff–Hoeffding bounds [Chernoff, 1952, Hoeffding, 1963, Schmidt et al., 1992] which provide bounds for the probability tails of sums of binary random variables. These inequalities can all lead to interesting new schema theorems. Unfortunately, the left-hand sides of these inequalities (i.e. the bound for the probability) are not constant, but depend on the expected value of the variable for which we want to estimate the probability tail. This seems to suggest that the calculations necessary to compute the probability of convergence of a GA might become quite complicated when using such inequalities. We intend to investigate this issue in future research.

Finally, it is important to stress that both Theorem 1 and Theorem 2 could be modified to provide upper bounds and confidence intervals. (An extension of Theorem 2 in this direction is described in [Poli, 1999b].) Since in this paper we are interested in the probability of finding solutions (rather than the probability of failing to find solutions), we deemed more important to concentrate our attention on lower bounds for such a probability. Nonetheless, it seems possible to extend some of the results in this paper to the case of upper bounds.

4 CONDITIONAL SCHEMA THEOREMS

The schema theorems described in the previous sections and in other work are valid on the assumption that the value of $\alpha(H, t)$ is a constant. If instead α is a random variable, the theorems need appropriate modifications.

For example, Equation 1 needs to be interpreted as:

$$E[m(H, t + 1)|\alpha(H, t) = a] = Ma, \quad (5)$$

a being an arbitrary constant in $[0,1]$, which provides information on the conditional expected value of the number of instances of a schema at the next generation. So, if one wanted to know the true expected value of $m(H, t + 1)$ the following integration would have to be performed:

$$E[m(H, t + 1)] = \int E[m(H, t + 1)|\alpha(H, t) = a] \text{pdf}(a) da,$$

where $\text{pdf}(a)$ is the probability density function of $\alpha(H, t)$. (A more extensive discussion on the validity of schema theorems in the presence of stochastic effects is presented in [Poli, 2000c].)

Likewise, the weak form of the schema theorem becomes:

Theorem 3 Conditional Probabilistic Schema Theorem (Weak Form).

For a schema H under fitness proportionate selection, one-point crossover applied with probability p_{xo} and no mutation, and for any fixed $k > 0$

$$\Pr\{m(H, t + 1) > Ma - k\sqrt{Ma(1-a)}|\alpha(H, t) = a\} \geq 1 - \frac{1}{k^2} \quad (6)$$

where a is an arbitrary number in $[0,1]$ and the other symbols have the same meaning as in Theorem 1.

This theorem provides a probabilistic lower bound for $m(H, t + 1)$ valid on the assumption that $\alpha(H, t) = a$. This can be transformed into:

Theorem 4 Conditional Probabilistic Schema Theorem (Expanded Weak Form).

For a schema H under fitness proportionate selection, one-point crossover applied with probability p_{xo} and no mutation,

$$\Pr\left\{m(H, t + 1) > x \left| (1 - p_{xo}) \frac{m(H, t)f(H, t)}{Mf(t)} + \frac{p_{xo}}{(N-1)M^2f^2(t)} \cdot \sum_{i=1}^{N-1} [m(L(H, i), t)f(L(H, i), t)m(R(H, i), t)f(R(H, i), t)] \geq \tilde{\alpha}(k, x, M) \right.\right\} \geq 1 - \frac{1}{k^2} \quad (7)$$

where

$$\tilde{\alpha}(k, x, M) = \frac{1}{2} \frac{M(k^2 + 2x) + k\sqrt{M^2k^2 + 4Mx(M-x)}}{M(k^2 + M)}. \quad (8)$$

Proof. The l.h.s. of Equation 4 is continuous, differentiable, has always a positive second derivative w.r.t. α and is zero for $\alpha = 0$ and $\alpha = k^2/(M + k^2)$. So, its minimum is between these two values, and it is therefore an increasing function of α for $\alpha \geq k^2/(M + k^2)$.

We are really interested only in the case in which $\alpha \geq k^2/(M + k^2)$ since $m(H, t + 1) \in \{0, 1, \dots, M\} \forall H, \forall t$ whereby only non-negative values of x make sense in Equation 4. Therefore, the l.h.s. of the equation is invertible (i.e. Equation 4 can be solved for x) and its inverse (w.r.t. x), $\tilde{\alpha}(k, x, M)$ (see Equation 8), is a continuous increasing function of x . This allows one to transform Equation 6 into

$$\Pr\{m(H, t + 1) > x|\alpha(H, t) = \tilde{\alpha}(k, x, M)\} \geq 1 - \frac{1}{k^2}. \quad (9)$$

From the properties of $\tilde{\alpha}(k, x, M)$ it follows that $\forall \epsilon \in [0, 1 - \tilde{\alpha}(k, x, M)] \exists \delta$ such that $\tilde{\alpha}(k, x, M) + \epsilon = \tilde{\alpha}(k, x + \delta, M)$. Therefore,

$$\begin{aligned} & \Pr\{m(H, t + 1) > x | \alpha(H, t) = \tilde{\alpha}(k, x, M) + \epsilon\} \\ & \geq \Pr\{m(H, t + 1) > x + \delta | \alpha(H, t) = \tilde{\alpha}(k, x, M) + \epsilon\} \\ & = \Pr\{m(H, t + 1) > x + \delta | \alpha(H, t) = \tilde{\alpha}(k, x + \delta, M)\} \\ & \geq 1 - \frac{1}{k^2} \end{aligned}$$

Since this is true for all valid values of ϵ , it follows that

$$\Pr\{m(H, t + 1) > x | 1 \geq \alpha(H, t) \geq \tilde{\alpha}(k, x, M)\} \geq 1 - \frac{1}{k^2}.$$

In this equation the condition $1 \geq \alpha(H, t)$ may be omitted since $\alpha(H, t)$ represents a probability, and so it cannot be meaningfully bigger than 1.

The proof is completed by substituting Equation 2 into the previous equation and considering that in fitness proportionate selection $p(K, t) = \frac{m(K, t)}{M} \frac{f(K, t)}{\bar{f}(t)}$. \square

For simplicity *in the rest of the paper it will be assumed that $p_{x_0} = 1$* in which case the theorem becomes

$$\Pr \left\{ m(H, t + 1) > x \left| \frac{1}{(N - 1)M^2 \bar{f}^2(t)} \cdot \sum_{i=1}^{N-1} [m(L(H, i), t) f(L(H, i), t) m(R(H, i), t) f(R(H, i), t)] \geq \tilde{\alpha}(k, x, M) \right. \right\} \geq 1 - \frac{1}{k^2} \quad (10)$$

5 A POSSIBLE ROUTE TO PROVING GA CONVERGENCE

Equation 10 is valid for any generation t , for any schema H and for any value of x , including $H = S$ (a solution) and $x = 0$. For these assignments, $m(S, t) > 0$ (i.e. the GA will find a solution at generation t) with probability $1 - 1/k^2$ (or higher), *if the conditioning event in Equation 10 is true at generation $t - 1$* . So, the equation indicates a condition that the potential building blocks of S need to satisfy at the penultimate generation in order for the GA to converge with a given probability.

Since a GA is a stochastic algorithm, in general it is impossible to guarantee that the condition in Equation 10 be satisfied. It is only possible to ensure that the probability of it being satisfied be say P (or at least P). This does not change the situation too much: it only means that $m(S, t) > 0$ with a probability of at least $P \cdot (1 - 1/k^2)$. If P and/or k are small this probability will be small. However, if one can perform multiple runs, the probability of finding at least a solution in R runs, $1 - [1 - P \cdot (1 - 1/k^2)]^R$, can be made arbitrarily large by increasing R .

So, if we knew P we would have a proof of convergence for GAs. The question is how to compute P . The following is a possible route to doing this (other alternatives exist, but we will not consider them in this paper).

Suppose we could transform the condition expressed by Equation 10 into a set of simpler but sufficient conditions of the form $m(L(H, i), t) > \mathcal{M}_{L(H, i), t}$ and $m(R(H, i), t) > \mathcal{M}_{R(H, i), t}$ (for $i = 1, \dots, N - 1$) where $\mathcal{M}_{L(H, i), t}$ and $\mathcal{M}_{R(H, i), t}$ are appropriate constants so that if all these simpler conditions are satisfied then also the conditioning event in Equation 10 is satisfied. Then we could apply Equation 10 recursively to each of the schemata $L(H, i)$ and $R(H, i)$, obtaining $2 \times (N - 1)$ conditions like the one in Equation 10 but for generation $t - 1$.⁴ Assuming that each is satisfied with a probability of at least P' and that all these events are independent (which may not be the case, see below) then $P \geq (P')^{2(N-1)}$. Now the problem would be to compute P' . However, exactly the same procedure just used for P could be used to compute P' . So, the condition in Equation 10 at generation t would become $[2 \times (N - 1)]^2$ conditions at generation $t - 2$. Assuming that each is satisfied with a probability of at least P'' then, $P' \geq (P'')^{2(N-1)}$, whereby $P' \geq ((P'')^{2(N-1)})^{2(N-1)} = (P'')^{[2(N-1)]^2}$. Now the problem would be to compute P'' .

This process could continue until quantities at generations 1 were involved. These are normally easily computable, thus allowing the completion of a GA convergence proof. Potentially this would involve a huge number of simple conditions to be satisfied at generation 1. However, this would not be the only complication. In order to compute a correct lower bound for P it would be necessary to compute the probabilities of being true of complex events which are the intersection of many non-independent events. This would not be easy to do.

Despite these difficulties all this might work, *if* we could transform the condition in Equation 10 into a set of simpler but sufficient conditions of the form mentioned above. Unfortunately, as one had to expect, this is not an easy thing to do either, because schema fitnesses and population fitness are present in Equation 10. These make the problem of computing P in its general form even harder to tackle mathematically.

A number of strategies are possible to find bounds for these fitnesses. For example one could use the ideas in the discussion on variance adjustments to the schema theorem in [Goldberg and Rudnick, 1991a, Goldberg and Rudnick, 1991b]. Another possibility would be to exploit something like Theorem 5 in [Altenberg, 1995] which gives the expected fitness distribution at the next generation. Similarly, perhaps one could use a statistical mechanics approach [Prügel-Bennett and Shapiro, 1994] to predict schema and population fitnesses. We have started to explore these ideas in extensions to the work presented in this paper. However, in the following we will not attempt to get rid of the population and schema fitnesses from our results. Instead we will use a variant of the strategy described in this section (which does not require assumptions on the independence of the events mentioned above) to find fitness-dependent convergence results. That is, we will find a lower bound for the *conditional probability of convergence* given a set of schema fitnesses. To do that we will use a different formulation of Equation 10.

In Equation 10 the quantities $\bar{f}(t)$, $m(L(H, i), t)$, $f(L(H, i), t)$, $m(R(H, i), t)$, $f(R(H, i), t)$ (for $i = 1, \dots, N - 1$) are stochastic variables. However, this equation can be specialised to the case in which we restrict ourselves to considering specific values for some (or all) such variables. When this is done, some additional conditioning events need to be added to the equation. For example, if we assume that the values of all the fitness-related variables

⁴Some of these conditions would actually coincide, leading to a smaller number of conditions.

$\bar{f}(t)$, $f(L(H, i), t)$, $f(R(H, i), t)$ (for $i = 1, \dots, N - 1$) are known, Equation 10 should be transformed into:

$$\Pr \left\{ m(H, t + 1) > x \mid \begin{aligned} & \frac{\sum_{i=1}^{N-1} [m(L(H, i), t) \langle f(L(H, i), t) \rangle m(R(H, i), t) \langle f(R(H, i), t) \rangle]}{(N-1)M^2 \langle \bar{f}(t) \rangle^2} \geq \bar{\alpha}(k, x, M), \\ & \bar{f}(t) = \langle \bar{f}(t) \rangle, f(L(H, 1), t) = \langle f(L(H, 1), t) \rangle, f(R(H, 1), t) = \langle f(R(H, 1), t) \rangle, \dots, \\ & f(L(H, N-1), t) = \langle f(L(H, N-1), t) \rangle, f(R(H, N-1), t) = \langle f(R(H, N-1), t) \rangle \end{aligned} \right\} \geq 1 - \frac{1}{k^2}, \quad (11)$$

where we used the following *notation*: if X is any random variable then $\langle X \rangle$ is taken to be a particular explicit value of X .⁵

It is easy to convince oneself of the correctness of this kind of specialisations of Equation 10, by noticing that Chebychev inequality guarantees that $\Pr\{m(H, t + 1) > x\} \geq 1 - \frac{1}{k^2}$ in any world in which $\alpha(H, t) \geq \bar{\alpha}(k, x, M)$ independently of the value of the variables on which α depends.

6 RECURSIVE CONDITIONAL SCHEMA THEOREM

By using the strategy described in the previous section and a specialisation of Equation 10 we obtain the following

Theorem 5 Conditional Recursive Schema Theorem. *For a schema H under fitness proportionate selection, one-point crossover applied with 100% probability and no mutation,*

$$\Pr\{m(H, t + 1) > \mathcal{M}_{H, t+1} \mid \mu_\iota, \phi_\iota\} \geq \left(1 - \frac{1}{k^2}\right) \left(\Pr\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t} \mid \mu_\iota, \phi_\iota\} + \Pr\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t} \mid \mu_\iota, \phi_\iota\} - 1\right)$$

where

$$\mu_\iota = \left\{ \mathcal{M}_{L(H, \iota), t} \mathcal{M}_{R(H, \iota), t} > \frac{\bar{\alpha}(k, \mathcal{M}_{H, t+1}, M)(N-1)M^2 \langle \bar{f}(t) \rangle^2}{\langle f(L(H, \iota), t) \rangle \langle f(R(H, \iota), t) \rangle} \right\}$$

and

$$\phi_\iota = \{\bar{f}(t) = \langle \bar{f}(t) \rangle, f(L(H, \iota), t) = \langle f(L(H, \iota), t) \rangle, f(R(H, \iota), t) = \langle f(R(H, \iota), t) \rangle\},$$

for any choice of the constants $\iota \in \{1, \dots, N - 1\}$, $\mathcal{M}_{H, t+1} \in [0, M]$, $\mathcal{M}_{L(H, \iota), t} \in [0, M]$ and $\mathcal{M}_{R(H, \iota), t} \in [0, M]$.

⁵If the value of a random variable $\langle X \rangle$ appears more than once in an equation, all the instances are assumed to represent the same number.

Proof. For brevity in this proof we will use the definition

$$\sigma_\iota = \frac{1}{(N-1)M^2\langle \bar{f}(t) \rangle^2} \cdot \left(m(L(H, \iota), t)\langle f(L(H, \iota), t) \rangle m(R(H, \iota), t)\langle f(R(H, \iota), t) \rangle + \sum_{i=1, i \neq \iota}^{N-1} m(L(H, i), t)f(L(H, i), t)m(R(H, i), t)f(R(H, i), t) \right).$$

When the joint event $\phi_\iota = \{\bar{f}(t) = \langle \bar{f}(t) \rangle, f(L(H, \iota), t) = \langle f(L(H, \iota), t) \rangle, f(R(H, \iota), t) = \langle f(R(H, \iota), t) \rangle\}$ happens, the event $\{m(H, t+1) > \mathcal{M}_{H, t+1}\}$ can only happen in two mutually exclusive situations: either when $\{\sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M)\}$ or when $\{\sigma_\iota < \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M)\}$. As a consequence,

$$\begin{aligned} \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1} | \phi_\iota\} &= \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1}, \sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) | \phi_\iota\} + \\ &\quad \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1}, \sigma_\iota < \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) | \phi_\iota\} \\ &\geq \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1}, \sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) | \phi_\iota\} \\ &= \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1} | \sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M), \phi_\iota\} \cdot \\ &\quad \Pr\{\sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) | \phi_\iota\} \\ &\geq \left(1 - \frac{1}{k^2}\right) \cdot \Pr\{\sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) | \phi_\iota\}, \end{aligned}$$

where in the last inequality we used a specialisation of Equation 10, i.e. a version of schema theorem, in which the values of $\bar{f}(t)$, $f(L(H, \iota), t)$ and $f(R(H, \iota), t)$ are assumed to be known.

Under any conditioning events the probability of the event $\{\sigma_\iota \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M)\}$ is necessarily greater than or equal to the probability of the event $\left\{\frac{1}{(N-1)M^2\langle \bar{f}(t) \rangle^2} [m(L(H, \iota), t)\langle f(L(H, \iota), t) \rangle] m(R(H, \iota), t)\langle f(R(H, \iota), t) \rangle] \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M)\right\}$ for any choice of the constant $\iota \in \{1, \dots, N-1\}$, provided that the fitness function is non-negative.⁶ Therefore, from the previous equations we obtain:

$$\begin{aligned} \Pr\{m(H, t+1) > \mathcal{M}_{H, t+1} | \phi_\iota\} &\geq \\ \left(1 - \frac{1}{k^2}\right) \Pr\left\{\frac{m(L(H, \iota), t)\langle f(L(H, \iota), t) \rangle m(R(H, \iota), t)\langle f(R(H, \iota), t) \rangle}{(N-1)M^2\langle \bar{f}(t) \rangle^2} \geq \tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M) \mid \phi_\iota\right\}. \end{aligned}$$

This can be rewritten as

$$\Pr\{m(H, t+1) > \mathcal{M}_{H, t+1} | \phi_\iota\} \geq \left(1 - \frac{1}{k^2}\right) \Pr\left\{m(L(H, \iota), t)m(R(H, \iota), t) \geq \frac{\tilde{\alpha}(k, \mathcal{M}_{H, t+1}, M)(N-1)M^2\langle \bar{f}(t) \rangle^2}{\langle f(L(H, \iota), t) \rangle \langle f(R(H, \iota), t) \rangle} \mid \phi_\iota\right\}.$$

The event on the right-hand side of this equation is not of the same form as the event on the left-hand side. So, it is not possible to use this result recursively. However, further simplifications can be obtained by noticing that if $m(L(H, \iota), t)$ and $m(R(H, \iota), t)$

⁶The difference between the probabilities of these two events may be very large, particularly if N is large. This means that the bounds provided by the theorem may be very pessimistic. However, at this stage we are not interested in the accuracy of our bounds.

are both greater than two suitably large constants, $\mathcal{M}_{L(H,\iota),t}$ and $\mathcal{M}_{R(H,\iota),t}$, the product $m(L(H,\iota),t)m(R(H,\iota),t)$ will always be greater than $\frac{\tilde{\alpha}(k,\mathcal{M}_{H,t+1},M)(N-1)M^2\langle\bar{f}(t)\rangle^2}{\langle f(L(H,\iota),t)\rangle\langle f(R(H,\iota),t)\rangle}$ (although obviously this can happen also when either $m(L(H,\iota),t)$ or $m(R(H,\iota),t)$ are not greater than $\mathcal{M}_{L(H,\iota),t}$ and $\mathcal{M}_{R(H,\iota),t}$). In order for this to happen the two constants need to be chosen such that the event $\mu_\iota = \left\{ \mathcal{M}_{L(H,\iota),t} \mathcal{M}_{R(H,\iota),t} > \frac{\tilde{\alpha}(k,\mathcal{M}_{H,t+1},M)(N-1)M^2\langle\bar{f}(t)\rangle^2}{\langle f(L(H,\iota),t)\rangle\langle f(R(H,\iota),t)\rangle} \right\}$ is the case. Therefore,

$$\Pr \left\{ m(L(H,\iota),t)m(R(H,\iota),t) \geq \frac{\tilde{\alpha}(k,\mathcal{M}_{H,t+1},M)(N-1)M^2\langle\bar{f}(t)\rangle^2}{\langle f(L(H,\iota),t)\rangle\langle f(R(H,\iota),t)\rangle} \middle| \mu_\iota, \phi_\iota \right\} \geq \Pr \left\{ m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}, m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t} \middle| \mu_\iota, \phi_\iota \right\}$$

where the extra conditioning event μ_ι has been introduced to guarantee that the choice of two constants $\mathcal{M}_{L(H,\iota),t}$ and $\mathcal{M}_{R(H,\iota),t}$ is appropriate. So, by repeating the calculations presented in this proof with this extra conditioning event, we obtain:

$$\begin{aligned} & \Pr\{m(H,t+1) > \mathcal{M}_{H,t+1} \mid \mu_\iota, \phi_\iota\} \\ & \geq \left(1 - \frac{1}{k^2}\right) \Pr\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}, m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t} \mid \mu_\iota, \phi_\iota\}. \end{aligned} \quad (12)$$

In order to obtain a recursive form of schema theorem we now need to simplify (or find a lower bound for) the conditional probability of the joint event $\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}, m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t}\}$.

If the events $\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}\}$ and $\{m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t}\}$ were independent we could write $\Pr\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}, m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t}\} = \Pr\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}\} \cdot \Pr\{m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t}\}$. However, in this paper we prefer to make no assumption on the independence of these events (more on this later). Fortunately, there is a way to compute a lower bound for their joint conditional probability: to use the Bonferroni inequality [Sobel and Uppuluri, 1972]. This states that $\Pr\{A, B\} \geq \Pr\{A\} + \Pr\{B\} - 1$ where A and B are two arbitrary (dependent or independent) events, which can trivially be extended to conditional probabilities obtaining: $\Pr\{A, B|C\} \geq \Pr\{A|C\} + \Pr\{B|C\} - 1$ where C is another arbitrary event. This leads to the following inequality:

$$\begin{aligned} & \Pr\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}, m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t} \mid \mu_\iota, \phi_\iota\} \geq \\ & \Pr\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t} \mid \mu_\iota, \phi_\iota\} + \Pr\{m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t} \mid \mu_\iota, \phi_\iota\} - 1, \end{aligned}$$

which, substituted into Equation 12, completes the proof of the theorem. \square

This theorem is recursive in the sense that with appropriate additional conditioning events similar to μ_ι and ϕ_ι (necessary to restrict the fitness of the building blocks of the building blocks of the schema H and to make sure appropriate constants are used) the theorem can be applied again to the events in its right-hand side, and then again the right-hand side of the resulting expressions and so on. So, this theorem provides information on the long term behaviour of a GA with respect to schema propagation in the same spirit as the well-known argument on the exponential increase/decrease of the number of instances of

a schema [Goldberg, 1989][page 30]. However, unlike such argument, Theorem 5 does not make any assumptions on the fitness of the schema or its building blocks.

It is obvious that this result is useful only when the probabilities on the right-hand side are all bigger than 0.5 (at least on average). If this is not the case the bound would be less than 0, whereby the theorem would simply state an obvious property of all probabilities.

It would be possible to obtain a much stronger result if the events $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ could be shown to be independent. This would allow to replace the Bonferroni inequality with a much tighter bound.⁷ However, at present the author is unable to prove or disprove whether $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ are indeed independent in all cases. So, Bonferroni inequality is the safest choice.

It is perhaps important to discuss why $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ might be independent. As two of the reviewers of this paper suggested, one might think that there is dependence between the two events due to any linkage disequilibrium that exists between the schemata $L(H, \iota)$ and $R(H, \iota)$ (interpreted as bit patterns), i.e. to the fact that $L(H, \iota)$ and $R(H, \iota)$ may tend to occur together among individuals in the population (positive linkage disequilibrium) or may be not found in the same individuals (negative linkage disequilibrium). This would be correct if in order to decide whether $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}, m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ is the case one sampled the population M times counting how many of the individuals selected belong to $L(H, \iota)$ and/or $R(H, \iota)$. However, this is not the correct way of interpreting $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}, m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$. This is because the first event in such a joint event refers to a property of the first parents selected for crossover, while the second event refers to a property of the second parents. So, the correct way to check whether $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}, m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ is the case is to: a) sample the population $2M$ times to produce M first parents and M second parents, b) count how many first parents are in $L(H, \iota)$, c) count how many second parents are in $R(H, \iota)$, and finally d) ask whether $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ are both the case.

If the population at generation t was fixed, these two events would seem to be independent for the simple reason that the first and the second parents in each crossover operation are selected with independent Bernoulli trials (e.g. with independent sweeps of the roulette). However, in this work we do not assume that the population is fixed, rather we see it as stochastic (i.e. we consider α as a stochastic variable). So, if the population instances were generated according to some particular distribution, surely the probability distributions of the stochastic variables $m(L(H, \iota), t)$ and $m(R(H, \iota), t)$ would be functions of the probability distribution of the population. So, some form of dependence between the events $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ would seem possible. However, one might reason, these events are independent for any particular instantiation of the population (and therefore for any particular value of α). So, it is possible that the events $\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t}\}$ and $\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t}\}$ remain independent when the population is treated as a stochastic quantity. We intend to study this hypothesis further in future work.

⁷In the case of independence, the r.h.s. of the theorem would become $(1 - \frac{1}{k^2}) \Pr\{m(L(H, \iota), t) > \mathcal{M}_{L(H, \iota), t} | \mu_\iota, \phi_\iota\} \Pr\{m(R(H, \iota), t) > \mathcal{M}_{R(H, \iota), t} | \mu_\iota, \phi_\iota\}$.

7 CONDITIONAL CONVERGENCE PROBABILITY

Despite its weakness, the previous theorem is important because it can be used to built equations which predict a property of a schema more than one generation in the future on the basis of properties of the building blocks of such a schema at an earlier generation. In particular the theorem can be used to find a lower bound for the conditional probability that the GA will converge to a particular solution S with a known effort. Here, instead of introducing a general but complicated conditional convergence theorem to show this, we prefer to illustrate the basic idea by using an example.

Suppose $N = 4$ and we want to know a lower bound for the probability that there will be at least one instance of the schema $H = S = b_1 b_2 b_3 b_4$ in the population by generation 3, on the assumption that the fitnesses of the building blocks of H and of the population at previous generations are known. A lower bound for this can be obtained using the theorem with $t = 2$ and $\mathcal{M}_{H,t+1} = \mathcal{M}_{b_1 b_2 b_3 b_4,3} = 0$.

If we set $k = 2$, we obtain $\tilde{\alpha}(k, \mathcal{M}_{H,t+1}, M)(N - 1)M^2 = 3M^2\tilde{\alpha}(2, 0, M)$. So, if we choose $\iota = 2$, in order for the theorem to be applicable we need to make sure that the event $\mu_2 = \left\{ \mathcal{M}_{b_1 b_2 ** , 2} \mathcal{M}_{** b_3 b_4, 2} > \frac{3M^2 \tilde{\alpha}(2, 0, M) \langle \bar{f}(2) \rangle^2}{\langle f(b_1 b_2 ** , 2) \rangle \langle f(** b_3 b_4, 2) \rangle} \right\}$ is the case. There are many ways to do this. A very simple one is to set $\mathcal{M}_{b_1 b_2 ** , 2} = \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(b_1 b_2 ** , 2) \rangle}$ and $\mathcal{M}_{** b_3 b_4, 2} = \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(** b_3 b_4, 2) \rangle}$, which, substituted into the theorem, lead to the following result:

$$\begin{aligned} \Pr\{m(b_1 b_2 b_3 b_4, 3) > 0 | \phi_2\} &\geq \\ 0.75 \cdot \left(\Pr \left\{ m(b_1 b_2 ** , 2) > \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(b_1 b_2 ** , 2) \rangle} \middle| \phi_2 \right\} + \right. \\ &\left. \Pr \left\{ m(** b_3 b_4, 2) > \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(** b_3 b_4, 2) \rangle} \middle| \phi_2 \right\} - 1 \right), \end{aligned}$$

where $\phi_2 = \{\bar{f}(2) = \langle \bar{f}(2) \rangle, f(b_1 b_2 ** , 2) = \langle f(b_1 b_2 ** , 2) \rangle, f(** b_3 b_4, 2) = \langle f(** b_3 b_4, 2) \rangle\}$. It should be noted that we have omitted the conditioning event μ_2 from the previous equation because we have chosen the constants $\mathcal{M}_{b_1 b_2 ** , 2}$ and $\mathcal{M}_{** b_3 b_4, 2}$ in such a way that μ_2 is always the case (effectively considering a special case of the recursive schema theorem given in the previous section).

The previous equation shows quite clearly how pessimistic our bound is. In fact, unless the schema fitnesses are sufficiently bigger than the average fitness of the population, the probabilities of the events on the right-hand side will be 0.

The recursive schema theorem can be applied again to such probabilities. For example let us calculate a lower bound for the probability that there will be at least $\frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(b_1 b_2 ** , 2) \rangle}$ instances of the schema $H = b_1 b_2 **$ in the population at generation 2, on the assumption that the fitnesses of the building blocks of H and of the population at previous generations are known. A lower bound for this can be obtained using the recursive schema theorem with $t = 1$ and $\mathcal{M}_{H,t+1} = \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(b_1 b_2 ** , 2) \rangle}$. If we set $k = 2$ again, we obtain $\tilde{\alpha}(k, \mathcal{M}_{H,t+1}, M)(N - 1)M = 3M^2\tilde{\alpha}(2, \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(b_1 b_2 ** , 2) \rangle}, M)$.

So, if we choose $\iota = 1$, $\mathcal{M}_{b_1***,2} = \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(\mathbf{b}_1***,1)\rangle}}$
and $\mathcal{M}_{*b_2**,2} = \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(*b_2**,1)\rangle}}$, and we substitute these values into the conditional recursive schema theorem, we obtain:

$$\begin{aligned} & \Pr\left\{m(\mathbf{b}_1\mathbf{b}_2**,2) > \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle} \mid \phi_1, \phi_2\right\} \geq \\ & 0.75 \cdot \left(\Pr\left\{m(\mathbf{b}_1***,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(\mathbf{b}_1***,1)\rangle}} \mid \phi_1, \phi_2\right\} + \right. \\ & \left. \Pr\left\{m(*b_2**,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(*b_2**,1)\rangle}} \mid \phi_1, \phi_2\right\} - 1\right), \end{aligned}$$

where $\phi_1 = \{\bar{f}(1) = \langle\bar{f}(1)\rangle, f(\mathbf{b}_1***,1) = \langle f(\mathbf{b}_1***,1)\rangle, f(*b_2**,1) = \langle f(*b_2**,1)\rangle\}$. A similar equation holds for the schema $H = **b_3b_4$, for which we will assume to use $\iota = 3$. By making use of these inequalities and representing the conditioning events on schema and population fitnesses with the symbol \mathcal{F} for brevity, we obtain:

$$\begin{aligned} & \Pr\{m(\mathbf{b}_1\mathbf{b}_2\mathbf{b}_3\mathbf{b}_4,3) > 0 \mid \mathcal{F}\} \geq \tag{13} \\ & 0.5625 \cdot \left(\Pr\left\{m(\mathbf{b}_1***,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(\mathbf{b}_1***,1)\rangle}} \mid \mathcal{F}\right\} + \right. \\ & \Pr\left\{m(*b_2**,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(\mathbf{b}_1\mathbf{b}_2**,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(*b_2**,1)\rangle}} \mid \mathcal{F}\right\} + \\ & \Pr\left\{m(**b_3*,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(**b_3b_4,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(**b_3*,1)\rangle}} \mid \mathcal{F}\right\} + \\ & \left. \Pr\left\{m(***b_4,1) > \sqrt{3M^2\tilde{\alpha}\left(2, \frac{\sqrt{3M^2\tilde{\alpha}(2,0,M)+1}\langle\bar{f}(2)\rangle}{\langle f(**b_3b_4,2)\rangle}, M\right) + 1 \frac{\langle\bar{f}(1)\rangle}{\langle f(***b_4,1)\rangle}} \mid \mathcal{F}\right\}\right) \\ & - 1.875 \end{aligned}$$

So, the lower bound for the probability of convergence at a given generation is a linear combination of the probabilities of having a sufficiently large number of building blocks of order 1 at the initial generation.

The weakness of this result is quite obvious. When all the probabilities on the right-hand side of the equation are 1, the lower bound we obtain is 0.375.⁸ In all other cases we get

⁸In case the events $\{m(L(H,\iota),t) > \mathcal{M}_{L(H,\iota),t}\}$ and $\{m(R(H,\iota),t) > \mathcal{M}_{R(H,\iota),t}\}$ could be shown to be independent, the bound in Equation 13 would be proportional to the product of the probabilities of having a sufficiently large number of building blocks of order 1 at the initial generation. In the example considered in this section, the bound would be 0.5625, with a 50% improvement with respect to the linear bound 0.375.

smaller bounds. In any case it should be noted that some of the quantities present in this equation are under our control since they depend on the initialisation strategy adopted. Therefore, it is not impossible for the events in the right-hand side of the equation to be all the case.

8 POPULATION SIZING

The recursive conditional schema theorem can be used to study the effect of the population size M on the conditional probability of convergence. We will show this continuing the example in the previous section.

For the sake of simplicity, let us assume that we initialise the population making sure that all the building blocks of order 1 have exactly the same number of instances, i.e. $m(0*\dots*, 1) = m(1*\dots*, 1) = m(*0*\dots*, 1) = m(*1*\dots*, 1) = \dots = m(*\dots*0, 1) = m(*\dots*1, 1) = M/2$.

A reasonable way to size the population in the previous example would be to choose M so as to maximise the lower bound in Equation 13.⁹ To achieve this one would have to make sure that each of the four events in the r.h.s. of the equation happen. Let us start from

the first one: $\left\{ m(\mathbf{b}_1 * ** , 1) > \sqrt{3M^2 \tilde{\alpha} \left(2, \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(\mathbf{b}_1 \mathbf{b}_2 ** , 2) \rangle}, M \right) + 1 \frac{\langle \bar{f}(1) \rangle}{\langle f(\mathbf{b}_1 ** * , 1) \rangle}} \middle| \mathcal{F} \right\}$.

Since $m(\mathbf{b}_1 * ** , 1) = M/2$, the event happens with probability 1 if

$$\frac{M}{2} > \sqrt{3M^2 \tilde{\alpha} \left(2, \frac{\sqrt{3M^2 \tilde{\alpha}(2, 0, M) + 1} \langle \bar{f}(2) \rangle}{\langle f(\mathbf{b}_1 \mathbf{b}_2 ** , 2) \rangle}, M \right) + 1 \frac{\langle \bar{f}(1) \rangle}{\langle f(\mathbf{b}_1 ** * , 1) \rangle}}.$$

Clearly, we are interested in the smallest value of M for which this inequality is satisfied. Since it is assumed that $\langle \bar{f}(1) \rangle$, $\langle f(\mathbf{b}_1 * ** , 1) \rangle$, $\langle \bar{f}(2) \rangle$ and $\langle f(\mathbf{b}_1 \mathbf{b}_2 ** , 2) \rangle$ are known, such a value of M , let us call it M_1 , can easily be obtained numerically.

The same procedure can be repeated for the other events in Equation 13, obtaining the lower bounds M_2 , M_3 and M_4 . Therefore, the minimum population size that maximises the right-hand side of Equation 13 is

$$M_{min} = \lceil \max(M_1, M_2, M_3, M_4) \rceil.$$

Of course, given the known weaknesses of the bounds used to derive the recursive schema theorem, it has to be expected that M_{min} will be much larger than necessary. To give a feel for the values suggested by the equation, let us imagine that the ratios between building block fitness and population fitness ($\langle f(\mathbf{b}_1 * ** , 1) \rangle / \langle \bar{f}(1) \rangle$, $\langle f(*\mathbf{b}_2 ** , 1) \rangle / \langle \bar{f}(1) \rangle$, $\langle f(\mathbf{b}_1 \mathbf{b}_2 ** , 2) \rangle / \langle \bar{f}(2) \rangle$, $\langle f(* * \mathbf{b}_3 \mathbf{b}_4 , 2) \rangle / \langle \bar{f}(2) \rangle$, etc.) be all equal to r . When the fitness ratio $r = 1$ (for example because the fitness landscape is flat) the population size suggested by the previous equation ($M_{min} = 2,322$) is huge considering that the length of the bit-strings in the population is only 4. The situation is even worse if $r < 1$. However, if the building blocks of the solution have well above average fitness, more realistic population sizes are suggested (e.g. if $r = 3$ one obtains $M_{min} = 6$).

⁹This is by no means neither the only nor the best way to size the population, but it is probably one of the simplest.

Table 1 Population sizes obtained for different fitness ratios for order-1 (r_1) and order-2 (r_2) building blocks.

r_1	r_2						
	0.5	0.75	1	2	3	4	5
0.5	119,896	55,416	32,394	9,378	4,778	3,054	2,204
0.75	24,590	11,566	6,874	2,112	1,130	754	564
1	8,056	3,848	2,322	750	418	288	222
2	556	276	172	62	38	28	24
3	106	52	32	10	6	6	4
4	42	16	6	2	2	2	2
5	42	16	6	2	2	2	2

It is interesting to compare how order-1 and order-2 building block fitnesses influence the population size. Let us imagine that the ratios between order-1 building block fitnesses and population fitness at generation 1 ($\langle f(\mathbf{b}_1 ** *) \rangle / \langle \bar{f}(1) \rangle$, $\langle f(* \mathbf{b}_2 ** *) \rangle / \langle \bar{f}(1) \rangle$, etc.) be constant and equal to r_1 and that the ratios between order-2 building block fitnesses and population fitness at generation 2 ($\langle f(\mathbf{b}_1 \mathbf{b}_2 ** *) \rangle / \langle \bar{f}(2) \rangle$ and $\langle f(** \mathbf{b}_3 \mathbf{b}_4, 2) \rangle / \langle \bar{f}(2) \rangle$) be constant and equal to r_2 . Table 1 shows the values of M_{min} resulting from different values of r_1 and r_2 . The population sizes in the table are all even because of the particular initialisation strategy adopted.

Clearly, the recursive schema theorem presented in this paper will need to be strengthened if we want to use it to size the population in practical applications. However, the procedure indicated in this section demonstrates that in principle this is a viable approach and that useful insights can be obtained already. For example, it is interesting to notice that the population sizes in the table depend significantly more on the order-1/generation-1 building-block fitness ratio r_1 than on the order-2/generation-2 building-block fitness ratio r_2 . This seems to suggest that problems with deceptive attractors for low-order building-blocks may be harder to solve for a GA than problems where deception is present when higher-order building-blocks are assembled. This conjecture will be checked in future work. In the future it would also be very interesting to compare the population sizing equations derived from this approach with those obtained by others (e.g. see [Goldberg et al., 1992]).

9 CONCLUSIONS AND FUTURE WORK

In this paper we have used a form of schema theorem in which expectations are not present in an unusual way, i.e. to predict the past from the future. This has allowed the derivation of a recursive version of the schema theorem which is applicable to the case of finite populations. This schema theorem allows one to find under which conditions on the initial generation the GA will converge to a solution in constant time. As an example, in the paper we have shown how such conditions can be derived for a generic 4-bit problem.

All the results in this paper are based on the assumption that the fitness of the building blocks involved in the process of finding a solution and the population fitness are known

at each generation. Therefore, our results do not represent a full schema-theorem-based proof of convergence for GAs. In future research we intend to explore the possibility of getting rid of schema and population fitnesses by replacing them with appropriate bounds based on the “true” characteristics of the schemata involved such as their static fitness. As indicated in Section 5, several approaches to tackle this problem are possible. If this step is successful, it will allow to identify rigorous strategies to size the population and therefore to calculate the computational effort required to solve a given problem using a GA. This in turn will open the way to a precise definition of “GA-friendly” (“GA-easy”) fitness functions. Such functions would simply be those for which the number of fitness evaluations necessary to find a solution with say 99% probability in multiple runs is smaller (much smaller) than 99% of the effort required by exhaustive search or random search without resampling.

Since the results in this paper are based on Chebychev inequality and Bonferroni bound, they are quite conservative. As a result they tend to considerably overestimate the population size necessary to solve a problem with a known level of performance. This does not mean that they will be useless in predicting on which functions a GA can do well. It simply means that they will over-restrict the set of GA-friendly functions. A lot can be done to improve the tightness of the lower bounds obtained in the paper. When less conservative results became available, more functions could be included in the GA-friendly set.

Not many people nowadays use fixed-size binary GAs with one-point crossover in practical applications. So, the theory presented in this paper, as often happens to all theory, could be thought as being ten or twenty years or so behind practice. However, there is really a considerable scope for extension to more recent operators and representations. For example, by using the crossover-mask-based approach presented in [Altenberg, 1995][Section 3 and Appendix] one could write an equation similar to Equation 2 valid for *any* type of homologous crossover on binary strings. The theory presented in this paper could then be extended for many crossover operators of practical interest. Also, in the exact schema theorem presented in [Stephens and Waelbroeck, 1997, Stephens and Waelbroeck, 1999] point mutation was present. So, it seems possible to extend the results presented in this paper to the case of point mutation (either alone or with some form of crossover). Finally, Stephens and Waelbroeck's theory has been recently generalised in [Poli, 2000b, Poli, 2000a] where an exact expression of $\alpha(H, t)$ for genetic programming with one-point crossover was reported. This is valid for variable-length and non-binary GAs as well as GP and standard GAs. As a result, it seems possible to extend the results presented in this paper to such representations and operators, too. So, although in its current form the theory presented in this paper is somehow behind practice, it is arguable that it might not remain so for long.

Despite their current limitations, we believe that the results reported in this paper are important because, unlike previous results, they make explicit the relation between population size, schema fitness and probability of convergence over multiple generations. These and other recent results show that schema theories are potentially very useful in analysing and designing GAs and that the scepticism with which they are dismissed in the evolutionary computation community is becoming less and less justifiable.

Acknowledgements

The author wishes to thank the members of the Evolutionary and Emergent Behaviour Intelligence and Computation (EEBIC) group at Birmingham, Bill Spears, Ken De Jong and Jonathan Rowe for useful comments and discussion. The reviewers of this paper are also thanked warmly for their thorough analysis and helpful comments. Finally, many thanks to Günter Rudolf for pointing out to us the existence of the Chernoff–Hoeffding bounds.

References

- [Altenberg, 1995] Altenberg, L. (1995). The Schema Theorem and Price's Theorem. In Whitley, L. D. and Vose, M. D., editors, *Foundations of Genetic Algorithms 3*, pages 23–49, Estes Park, Colorado, USA. Morgan Kaufmann.
- [Chernoff, 1952] Chernoff, H. (1952). a measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *annals of mathematical statistics*, 23(4):493–507.
- [Chung and Perez, 1994] Chung, S. W. and Perez, R. A. (1994). The schema theorem considered insufficient. In *Proceedings of the Sixth IEEE International Conference on Tools with Artificial Intelligence*, pages 748–751, New Orleans.
- [Davis and Principe, 1993] Davis, T. E. and Principe, J. C. (1993). A Markov chain framework for the simple genetic algorithm. *Evolutionary Computation*, 1(3):269–288.
- [De Jong et al., 1995] De Jong, K. A., Spears, W. M., and Gordon, D. F. (1995). Using Markov chains to analyze GAFOs. In Whitley, L. D. and Vose, M. D., editors, *Foundations of Genetic Algorithms 3*, pages 115–137. Morgan Kaufmann, San Francisco, CA.
- [Fogel and Ghozeil, 1997] Fogel, D. B. and Ghozeil, A. (1997). Schema processing under proportional selection in the presence of random effects. *IEEE Transactions on Evolutionary Computation*, 1(4):290–293.
- [Fogel and Ghozeil, 1998] Fogel, D. B. and Ghozeil, A. (1998). The schema theorem and the misallocation of trials in the presence of stochastic effects. In Porto, V. W., Saravanan, N., Waagen, D., and Eiben, A. E., editors, *Evolutionary Programming VII: Proc. of the 7th Ann. Conf. on Evolutionary Programming*, pages 313–321, Berlin. Springer.
- [Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Massachusetts.
- [Goldberg et al., 1992] Goldberg, D. E., Deb, K., and Clark, J. H. (1992). Accounting for noise in the sizing of populations. In Whitley, D., editor, *Foundations of Genetic Algorithms Workshop (FOGA-92)*, Vail, Colorado.
- [Goldberg and Rudnick, 1991a] Goldberg, D. E. and Rudnick, M. (1991a). Genetic algorithms and the variance of fitness. Technical Report IlliGAL Report No 91001, Department of General Engineering, University of Illinois at Urbana-Champaign.
- [Goldberg and Rudnick, 1991b] Goldberg, D. E. and Rudnick, M. (1991b). Genetic algorithms and the variance of fitness. *Complex systems*, 5:265–278.

- [Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bonded random variables. *Journal of the American Statistical Association*, 58(301):13–30.
- [Nix and Vose, 1992] Nix, A. E. and Vose, M. D. (1992). Modeling genetic algorithms with Markov chains. *Annals of Mathematics and Artificial Intelligence*, 5:79–88.
- [Poli, 1999a] Poli, R. (1999a). Probabilistic schema theorems without expectation, recursive conditional schema theorem, convergence and population sizing in genetic algorithms. Technical Report CSRP-99-3, University of Birmingham, School of Computer Science.
- [Poli, 1999b] Poli, R. (1999b). Schema theorems without expectations. In Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M., and Smith, R. E., editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 1, page 806, Orlando, Florida, USA. Morgan Kaufmann.
- [Poli, 2000a] Poli, R. (2000a). Exact schema theorem and effective fitness for GP with one-point crossover. In Whitley, D., Goldberg, D., Cantu-Paz, E., Spector, L., Parmee, I., and Beyer, H.-G., editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 469–476, Las Vegas. Morgan Kaufmann.
- [Poli, 2000b] Poli, R. (2000b). Hyperschema theory for GP with one-point crossover, building blocks, and some new results in GA theory. In Poli, R., Banzhaf, W., Langdon, W. B., Miller, J. F., Nordin, P., and Fogarty, T. C., editors, *Genetic Programming, Proceedings of EuroGP'2000*, volume 1802 of *LNCIS*, pages 163–180, Edinburgh. Springer-Verlag.
- [Poli, 2000c] Poli, R. (2000c). Why the schema theorem is correct also in the presence of stochastic effects. In *Proceedings of the Congress on Evolutionary Computation (CEC 2000)*, pages 487–492, San Diego, USA.
- [Poli et al., 1998] Poli, R., Langdon, W. B., and O'Reilly, U.-M. (1998). Analysis of schema variance and short term extinction likelihoods. In Koza, J. R., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M. H., Goldberg, D. E., Iba, H., and Riolo, R., editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 284–292, University of Wisconsin, Madison, Wisconsin, USA. Morgan Kaufmann.
- [Prügel-Bennett and Shapiro, 1994] Prügel-Bennett, A. and Shapiro, J. L. (1994). An analysis of genetic algorithms using statistical mechanics. *Physical Review Letters*, 72:1305–1309.
- [Radcliffe, 1997] Radcliffe, N. J. (1997). Schema processing. In Baeck, T., Fogel, D. B., and Michalewicz, Z., editors, *Handbook of Evolutionary Computation*, pages B2.5–1–10. Oxford University Press.
- [Rudolph, 1994] Rudolph, G. (1994). Convergence analysis of canonical genetic algorithm. *IEEE Transactions on Neural Networks*, 5(1):96–101.
- [Rudolph, 1997a] Rudolph, G. (1997a). Genetic algorithms. In Baeck, T., Fogel, D. B., and Michalewicz, Z., editors, *Handbook of Evolutionary Computation*, pages B2.4–20–27. Oxford University Press.
- [Rudolph, 1997b] Rudolph, G. (1997b). Models of stochastic convergence. In Baeck, T., Fogel, D. B., and Michalewicz, Z., editors, *Handbook of Evolutionary Computation*,

pages B2.3–1–3. Oxford University Press.

- [Rudolph, 1997c] Rudolph, G. (1997c). Stochastic processes. In Baeck, T., Fogel, D. B., and Michalewicz, Z., editors, *Handbook of Evolutionary Computation*, pages B2.2–1–8. Oxford University Press.
- [Schmidt et al., 1992] Schmidt, J. P., Siegel, A., and Srinivasan, A. (1992). Chernoff-Hoeffding bounds for applications with limited independence. Technical Report 92-1305, Department of Computer Science, Cornell University.
- [Sobel and Uppuluri, 1972] Sobel, M. and Uppuluri, V. R. R. (1972). On Bonferroni-type inequalities of the same degree for the probability of unions and intersections. *Annals of Mathematical Statistics*, 43(5):1549–1558.
- [Spiegel, 1975] Spiegel, M. R. (1975). *Probability and Statistics*. McGraw-Hill, New York.
- [Stephens and Waelbroeck, 1997] Stephens, C. R. and Waelbroeck, H. (1997). Effective degrees of freedom in genetic algorithms and the block hypothesis. In Bäck, T., editor, *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, pages 34–40, East Lansing. Morgan Kaufmann.
- [Stephens and Waelbroeck, 1999] Stephens, C. R. and Waelbroeck, H. (1999). Schemata evolution and building blocks. *Evolutionary Computation*, 7(2):109–124.
- [Vose, 1999] Vose, M. D. (1999). *The simple genetic algorithm: Foundations and theory*. MIT Press, Cambridge, MA.
- [Wright, 1931] Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.