Probabilistic Schema Theorems without Expectation, 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
Phone: +44-121-414-3739

Technical Report: CSRP-99-3
January 1999 (revised April and December 1999)[*]

Abstract

In this paper we first develop a new form of schema theorem in which expectations are not present. This theorem allows 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 use this version of the schema theorem backwards, i.e. to predict the past from the future. Assuming that at least one solution is found at one generation, this allows us to find the conditions (at the previous generation) under which such a solution will indeed be found with a given probability. This allows us to obtain a recursive version of the schema theorem. This schema theorem allows one to find under which conditions on the initial generation the GA will converge to a solution on the hypothesis that building block and population fitnesses are known. These results are important because for the first time they make explicit the relation between population size, schema fitness and probability of convergence over multiple generations.

1 Introduction

Since John Holland's seminal work in the mid seventies and his well known schema theorem (see (Holland 1992) and (Goldberg 1989)), schemata are traditionally used to

[*]This report has been revised for several reasons. Firstly, we wanted to simplify the mathematical notation used in the previous versions of the manuscript to make the work easier to understand. Secondly, in one of the steps of one of the proofs independence of two events was assumed, but later it was realised that we could not prove nor disprove the correctness of this hypothesis. So, the step was modified so as not to rely on the hypothesis of independence and the changes were propagated to the rest of the manuscript. This has changed some of the equations, but not the nature of the results presented in earlier versions of this report. Finally, an example has been provided to clarify strengths and weaknesses of the results presented in the paper.

explain why GAs and more recently GP (Poli and Langdon 1997*b*, Poli and Langdon 1998, Rosca 1997, Poli 1999*c*) work. Schemata are similarity templates representing entire groups of points in the search space. The schema theorem describes how schemata are expected to propagate generation after generation under the effects of selection, crossover and mutation.[1]

The usefulness of schemata has been widely criticised (see for example (Altenberg 1995, Macready and Wolpert 1996, Fogel and Ghozeil 1997, Fogel and Ghozeil 1998, Fogel and Ghozeil 1999)), and quite a number of researchers nowadays seem to believe that the schema theorem is nothing more than a trivial tautology of no use whatsoever. However, as correctly stated in (Radcliffe 1997) the problem with the schema theorem is probably not the theorem itself, rather its over-interpretations.

Recently, the attention of GA theorists has moved away from schemata to land onto Markov chains (Nix and Vose 1992, Davis and Principe 1993, Rudolph 1997*c*). These are very accurate models and have been very useful to obtain theoretical results on the convergence properties of GAs (for example, on the expected time to hit a solution (De Jong *et al.* 1995, Rudolph 1997*a*) or on GA asymptotic convergence (Nix and Vose 1992, Rudolph 1994, Rudolph 1997*b*)). However, although results based on Markov chains are very important in principle, very few useful recipes for practical GA users have been drawn from these fine grain stochastic models.

Somewhere in the middle is the approach of researchers, such as David Goldberg and his students, who don't reject the notion of schemata and the schema theorem. Rather they use these within probabilistic calculations. These are of a much coarser nature with respect to Markov chains and can only provide approximate results. The reasons behind this approach are: a) "there is a no free lunch theorem for GA theory" (Goldberg 1998) (i.e. the more general a theory is the less important it is for practical applications) and, b) from an engineering point of view it is sufficient to do as much theory as it is necessary to get a reasonably good approximation of the behaviour of a GA in practical situations. This middle ground approach has also been quite productive. For example, it has lead to formulating population sizing equations which are compact and reasonably easy to understand (Goldberg *et al.* 1991) and to produce recipes on how to properly set the parameters of a GA (Goldberg *et al.* 1993).

Very recently Stephens and Waelbroeck (Stephens and Waelbroeck 1997, Stephens and Waelbroeck 1999) have produced a new schema theorem which, unlike previous results which concentrated on schema survival and disruption, makes the effects and the mechanisms of schema creation explicit. This theorem gives an exact formulation (rather than a lower bound) for the expected number of instances of a schema at the next generation. Stephens and Waelbroeck used this result as a starting point for a number of other important results on the behaviour of a GA over multiple generations in the assumption of infinite populations. This supports our and other researchers' belief that schema theories have been deemed as useless too prematurely.

In our recent work on GP schemata (Poli and Langdon 1997*b*, Poli and Langdon 1997*a*,

[1] In an alternative interpretation schemata are seen as subsets of the search space, and the schema theorem is interpreted as a description of the expected number of elements of the population belonging to such subsets.

Poli and Langdon 1998) we started asking ourselves whether Markov chain theory is the only way to develop an exact theory for GA convergence. We also wondered whether the schema theorem we had just developed was indeed a useless result. One of the criticisms of the schema theorem is that it only gives a *lower bound* on the *expected value* of the number of individuals sampling a given schema at the next generation. So, one question we wanted to investigate immediately was how reliable such an estimate was. This is why in (Poli *et al.* 1998) we analysed the impact of variance on schema transmission and we obtained a schema-variance theorem which is general and applicable to GAs as well as GP. Encouraged by these results, we then decided to see how far we could go in studying GA convergence using the schema theorem and information on schema variance. This paper presents some of the results of this effort.

The paper is organised as follows. After describing the assumptions on which our work is based (Section 2), we develop a new form of schema theorem in which expectations are not present, in Section 3. This theorem allows 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 use this version of the schema theorem backwards to predict the past from the future. This allows us to find the conditions at one generation under which a solution will be found at the next generation. We discuss a general strategy to use this result in proving GA convergence in Section 5 and we formalise this strategy obtaining a conditional recursive version of the schema theorem (Section 6). This can be used to find under which conditions on the initial generation the GA will converge in constant time in the assumption that population and building-block fitnesses are known. In Section 7 we apply this theorem to a sample problem to illustrate its strengths and weaknesses. The results reported in this paper allow us to make explicit the relation between the probability of finding a solution, schema fitness and population size as discussed by means of example in Section 8. We draw some conclusions 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).

We define the *total transmission probability* for a schema H, $\alpha(H,t)$, as the probability that, at generation t, every time we create/copy (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 is the population size, $m(H,t+1)$ is the number of copies of the schema H at generation $t+1$ and $E[\cdot]$ is the expectation operator.

In a binary GA the total transmission probability is given by the following equation (which can be obtained either by simplifying the results in (Stephens and Waelbroeck

1997, Stephens and Waelbroeck 1999) or by decomposing the probability that H be created into the sum of the probabilities that this will happen for each possible crossover point):

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

where p_{xo} is the crossover probability, $p(x,t)$ is the selection probability of a schema x at generation t, $L(H,i)$ is the schema obtained by replacing with "don't care" symbols all the elements of H from position $i+1$ to position N, $R(H,i)$ is the schema obtained by replacing with "don't care" symbols all the elements of H from position 1 to position i, and i varies over the valid crossover points.[2] For example, if $H =$`**1111`, $L(H,1) =$`******`, $R(H,1) =$`**1111`, $L(H,3) =$`**1***`, $R(H,3) =$`***111`. Equation 2 can be generalised for genetic programming with one-point crossover (Poli and Langdon 1998) when the appropriate definition of schema is used (Poli 1999*c*, Poli 1999*a*), which makes it possible to extend most of the results in this paper to GP.

For any given problem there may be many solutions. The objective in this work is to find conditions which guarantee that a GA will find at least one of them with a given probability (perhaps in multiple runs). This is what is meant by *GA convergence* in this paper. Let us denote such an unknown solution with $S = b_1b_2\cdots b_N$.

3 Probabilistic Schema Theorems without Expected Values

3.1 Two Strategies to Remove the Expectation Operator

In previous work (Poli *et al.* 1998) 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 $\alpha(H,t)$. 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.

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

where M is the population size.

Therefore, if we know the value of α, we can calculate exactly the probability $\pi(\alpha,x)$ that the schema H will have at least x instances at generation $t+1$. This is:

$$\begin{aligned}\pi(\alpha,x) &= \Pr\{m(H,t+1)\geq x\} = \\ &= \sum_{k=x}^{M}\Pr\{m(H,t+1)=k\} = \\ &= \frac{\Gamma(M)M\alpha^x\,(1-\alpha)^M\, hypg([1,x-M],[1+x],\frac{\alpha}{\alpha-1})}{\Gamma(M-x)\,(M-x)\,\Gamma(x)x\,(1-\alpha)^x}\end{aligned} \tag{3}$$

[2]The symbol L stands for "left part of", while R stands for "right part of".

where *hypg* is the hypergeometric probability distribution defined as follows

$$hypg(n,d,z) = \sum_{k=0}^{\infty} \frac{\prod_{i=1}^{j} \frac{\Gamma(n_i+k)}{\Gamma(n_i)} z^k}{\prod_{i=1}^{m} \frac{\Gamma(d_i+k)}{\Gamma(d_i)} k!}$$

where j is the number of terms in the list $n = [n_1, n_2, \ldots]$ and m is the number of terms in the list $d = [d_1, d_2, \ldots]$, Γ is a generalisation of the factorial function defined as

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt,$$

α is a shorthand notation for $\alpha(H,t)$ and z is a complex number. From this we obtain:

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

$$\Pr\{m(H,t+1) \geq x\} = \pi(\alpha(H,t), x),$$

where $\pi(\cdot)$ is defined in Equation 3, $\alpha(\cdot)$ is defined in Equation 2 and the probability of selection of a generic schema K is $p(K,t) = \frac{m(K,t)}{M} \frac{f(K,t)}{\bar{f}(t)}$, 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.

This theorem can be used in several practical ways. For example, if one knows the population fitness and the fitnesses and number of instances of the schemata in Equation 2, one can compute the exact value of $\alpha(H,t)$. When this is known, one can compute the probability of the event $\{m(H,t+1) \geq x\}$ for any given value of x. Alternatively, if for a given value of α we want that $\Pr\{m(H,t+1) \geq x\} = y$, y being a prefixed probability value, then the theorem allows us to compute the necessary value of x. For example, if we fix $y = 0.99$, solving the equation $\pi(\alpha(H,t), x) = 0.99$ for x would give us a very good probabilistic lower bound for $m(H,t+1)$.

Another way in which the theorem can be used is to find conditions on α under which some prefixed values of x and y can be obtained. This is very important since it allows us to use the schema theorem to find sufficient conditions for convergence of a GA, as will be shown later. Unfortunately, there is one problem with this idea: Equation 3 is not easily solvable for α

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):

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

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

Since $m(H,t+1)$ is binomially distributed,

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

and

$$StdDev[m(H,t+1)] = \sqrt{M\alpha(H,t)\,[1-\alpha(H,t)]}\,.$$

By substituting these equations into Chebyshev's inequality we obtain:

$$\Pr\left\{|m(H,t+1) - M\alpha| \le k\sqrt{M\alpha(1-\alpha)}\right\} \ge 1 - \frac{1}{k^2}$$

where α is a shorthand notation for $\alpha(H,t)$. Since,

$$\begin{aligned} & \Pr\left\{m(H,t+1) > M\alpha - k\sqrt{M\alpha(1-\alpha)}\right\} \\ \ge \; & \Pr\left\{|m(H,t+1) - M\alpha| \le k\sqrt{M\alpha(1-\alpha)}\right\} \end{aligned}$$

we obtain

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

$$\Pr\{m(H,t+1) > M\alpha(H,t) - k\sqrt{M\alpha(H,t)(1-\alpha(H,t))}\} \ge 1 - \frac{1}{k^2} \tag{4}$$

for any fixed $k > 0$, with the same meaning of the symbols as in Theorem 1.[3]

This version of the probabilistic schema theorem can be used in exactly the same ways as the previous version. For example, given x and α we can compute the probability y that $m(H,t+1) > x$ by solving the equation $M\alpha - k\sqrt{M\alpha(1-\alpha)} = x$ for k and then substituting the result into $y = 1 - k^{-2}$. Alternatively, given y and α we can compute $k = \frac{1}{\sqrt{1-y}}$, and then compute x from the previous equation. Also, unlike Theorem 1, this theorem allows us 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

$$M\alpha - k\sqrt{M\alpha(1-\alpha)} = x \tag{5}$$

for α, which gives the solution

$$\alpha'(x) \stackrel{def}{=} \frac{1}{2}\,\frac{M(k^2+2x) + k\sqrt{M^2k^2 + 4\,Mx(M-x)}}{M(k^2+M)}, \tag{6}$$

and then substituting $k = \frac{1}{\sqrt{1-y}}$ into the result. A potential problem with Equation 6 is that the r.h.s. is a complicated function of the population size M: a parameter which we want to determine later on. Fortunately, this dependency can be simplified as described in Section 4.

[3] Elsewhere we have defined another version of this theorem which provides both an upper and a lower bound for $m(H,t+1)$ (Poli 1999*b*).

3.2 Conditional Schema Theorems

The schema theorems described in this and the previous section and in other work are valid in 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:

$$E[m(H,t+1)|\alpha(H,t)=a] = Ma, \tag{7}$$

which provides information on the conditional expected value of the number of instances of a schema at the next generation, i.e. the expected value of $m(H,t{+}1)$ in the assumption that $\alpha(H,t)=a$, a being an arbitrary constant in [0,1].

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 $k>0$

$$\Pr\{m(H,t+1) > Ma - k\sqrt{Ma(1-a)}|\alpha(H,t)=a\} \geq 1-\frac{1}{k^2} \tag{8}$$

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 bond for $m(H,t+1)$ valid in the assumption that $\alpha(H,t)=a$.

3.3 Other Strategies to Remove the Expectation Operator

It is worth noting that Chebychev inequality tends to provide very 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 (many thanks to Günter Rudolf for pointing this out to us). 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 prohibitively complicated when using such inequalities. We intend to investigate this issue in future research.

4 Using the Schema Theorem Backwards

Equation 8 can be written as:

$$\Pr\{m(H,t+1) > x|\alpha(H,t)=\alpha'(x)\} \geq 1-\frac{1}{k^2} \tag{9}$$

where $\alpha'(x)$, defined in Equation 6, is the solution of Equation 5. The l.h.s. of Equation 5 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 $\alpha \geq k^2/(M+k^2)$. It is easy to show that $\alpha' \geq k^2/(M+k^2)$ (since $x \in [0, M]$). Therefore, the l.h.s. of the equation is an increasing function of α. As a consequence its inverse, $\alpha'(x)$, is a continuous increasing function of x.

From these properties it follows that $\forall \epsilon \in [0, 1-\alpha'(x)), \exists \delta$ such that $\alpha'(x) + \epsilon = \alpha'(x+\delta)$. Therefore,

$$\begin{aligned}
1 - \frac{1}{k^2} &\leq \Pr\{m(H,t+1) > x + \delta | \alpha(H,t) = \alpha'(x+\delta)\} && (10)\\
&= \Pr\{m(H,t+1) > x + \delta | \alpha(H,t) = \alpha'(x) + \epsilon\} && (11)\\
&\leq \Pr\{m(H,t+1) > x | \alpha(H,t) = \alpha'(x) + \epsilon\} && (12)
\end{aligned}$$

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

$$\Pr\{m(H,t+1) > x | \alpha(H,t) \geq \alpha'(x) + \epsilon\} \geq 1 - \frac{1}{k^2}$$

This equation is quite useful because it allows us to replace the complicated lower bound for α, α', with a simpler one provided that the latter is bigger than the former. The more restrictive requirement for α can be found considering that:

$$\begin{aligned}
\alpha'(x) &= \frac{1}{2} \frac{M(k^2 + 2x) + k\sqrt{M^2k^2 + 4\,Mx(M-x)}}{M(k^2+M)}\\
&< \frac{1}{2} \frac{M(k^2 + 2x) + k\sqrt{M^2k^2 + 4\,M^2x}}{M^2}
\end{aligned}$$

whereby (after additional simplifications) we obtain

$$\alpha'(x) < \alpha''(x) \stackrel{def}{=} \frac{\psi(k,x)}{M} \tag{13}$$

where

$$\psi(k,x) = \frac{1}{2}\,(k^2 + 2\,x + k\sqrt{k^2 + 4\,x}). \tag{14}$$

So, the previous calculations guarantee that

$$\Pr\{m(H,t+1) > x | \alpha(H,t) \geq \alpha''(x)\} \geq 1 - 1/k^2.$$

By substituting Equation 2 into the previous equation we obtain:

$$\Pr\{m(H,t+1) > x | (1-p_{xo})p(H,t) + \frac{p_{xo}}{N-1}\sum_{i=1}^{N-1} p(L(H,i),t)p(R(H,i),t) \geq \frac{\psi(k,x)}{M}\} \geq 1 - 1/k^2. \tag{15}$$

Since we have assumed to use fitness proportionate selection, $p(K,t) = \frac{m(K,t)}{M}\frac{f(K,t)}{\bar{f}(t)}$, whereby

Theorem 4 Conditional Probabilistic Schema Theorem (Expanded Weak Form). *For a schema H under fitness proportionate selection, one-point crossover applied with p_{xo} probability and no mutation,*

$$\begin{aligned}\Pr\{m(H,t+1) > x | (1-p_{xo})\frac{m(H,t)f(H,t)}{\bar{f}(t)} + \frac{p_{xo}}{(N-1)M\bar{f}^2(t)}\cdot \\ \cdot \sum_{i=1}^{N-1}[m(L(H,i),t)f(L(H,i),t)\cdot \\ \cdot\, m(R(H,i),t)f(R(H,i),t)] \geq \psi(k,x)\} \geq 1-\frac{1}{k^2}.\end{aligned} \quad (16)$$

For simplicity in the rest of the paper we will consider only the case of $p_{xo}=1$ expressed by the following

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

$$\begin{aligned}\Pr\{m(H,t+1) > x | \frac{1}{(N-1)M\bar{f}^2(t)}\cdot \\ \cdot \sum_{i=1}^{N-1}[m(L(H,i),t)f(L(H,i),t)\cdot \\ \cdot\, m(R(H,i),t)f(R(H,i),t)] \geq \psi(k,x)\} \geq 1-\frac{1}{k^2}\end{aligned} \quad (17)$$

It is important note that 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 $i=1,\cdots,N-1$ in the previous two equations are stochastic variables.

This result shows that Theorem 3 can be used backwards in time. This allows us to transform a constraint ($m(H,t+1)>x$) on one variable at generation $t+1$ into a constraint on several variables at the previous generation, as we will discuss in the next section.

5 A Possible Route to Prove GA Convergence

Equation 17 is valid for any schema H, for any generation t and for any value of x, including $H=S$ (a solution), $t=T-1$ (T being an arbitrary positive integer) and $x=0$. For these assignments, $m(S,T)>0$ (i.e. our GA will find a solution at generation T) with probability $1-1/k^2$, *if the conditioning event in Equation 17 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 17 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 $P\cdot(1-1/k^2)$ (assuming independence, see below). 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 constraint expressed by Equation 17 into a set of simpler but sufficient constraints of the form $m(L(H,i),t) > \mathcal{M}_{L(H,i),t}$ and $m(R(H,i),t) > \mathcal{M}_{R(H,i),t}$ where $\mathcal{M}_{L(H,i),t}$ and $\mathcal{M}_{R(H,i),t}$ are appropriate constants so that if all these simpler constraints are satisfied then also the conditioning event in Equation 17 is satisfied. Then we could apply Equation 17 recursively to each of the schemata $L(H,i)$ and $R(H,i)$, obtaining $2\times(N-1)$ constraints like the one in Equation 17 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) 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 constraint in Equation 17 at generation t would become $[2\times(N-1)]^2$ constraints at generation $t-2$. Assuming that each is satisfied with a probability of at least P'' then, $P' \geq (P'')^{2(N-1)}$, whereby $P' \geq \left((P'')^{2(N-1)}\right)^{2(N-1)} = (P'')^{[2(N-1)]^2}$. 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 constraints 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 constraint in Equation 17 into a set of simpler but sufficient constraints of the form mentioned above. Unfortunately, as had to be expected, this is not an easy thing to do either, because schema fitnesses and population fitness are present in Equation 17. 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 (see for example the discussion on variance adjustments to the schema theorem in (Goldberg and Rudnick 1991)), and we have started to explore them 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 the strategy described in this section to find fitness-dependent convergence results, i.e. 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 17:

[4] Some of these constraints would actually coincide, leading to a smaller number of constraints.

$$
\begin{aligned}
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}| \\
&\quad \frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \left[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t} \right] \geq \psi(k, \mathcal{M}_{H,t+1}), \\
&\quad \bar{f}(t) = f_t, \\
&\quad f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \\
&\quad f(L(H,2),t) = f_{L(H,2),t}, f(R(H,2),t) = f_{R(H,2),t}, \\
&\quad \vdots \\
&\quad f(L(H,N-1),t) = f_{L(H,N-1),t}, f(R(H,N-1),t) = f_{R(H,N-1),t}\} \geq 1 - \frac{1}{k^2}.
\end{aligned} \tag{18}
$$

This equation can be interpreted as a special case of Equation 17 obtained by restricting ourselves to considering the specific values f_t, $f_{L(H,1),t}$, $f_{R(H,1),t}$, $f_{L(H,2),t}$, $f_{R(H,2),t}$, $\ldots$ $f_{L(H,N-1),t}$, $f_{R(H,N-1),t}$ for population and schema fitnesses and by renaming the constant x with the symbol $\mathcal{M}_{H,t+1}$. It is easy to convince oneself of the correctness of this equation, by noticing that Chebychev inequality guarantees that $Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}\} \geq 1 - \frac{1}{k^2}$ in any world in which $M\alpha(H,t) \geq \psi(k, \mathcal{M}_{H,t+1})$ independently of the value of the variables on which α depends.

In the following section will use this equation to derive a recursive form of schema theorem which represents an implementation of the strategy to obtain GA convergence results described in this section.

6 Recursive Conditional Schema Theorem

When the joint event $\{\bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots, f(L(H,N-1),t) = f_{L(H,N-1),t}, f(R(H,N-1),t) = f_{R(H,N-1),t}\}$ happens the event $\{m(H,t+1) > \mathcal{M}_{H,t+1}\}$ can only happen in two mutually exclusive situations: either when $\{\frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \left[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t} \right] \geq \psi(k, \mathcal{M}_{H,t+1})\}$ or when $\{\frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \left[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t} \right] < \psi(k, \mathcal{M}_{H,t+1})\}$. As a consequence,

$$
\begin{aligned}
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}| \\
&\quad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} = \\
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}, \\
&\quad \frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \left[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t} \right] \geq \psi(k, \mathcal{M}_{H,t+1})| \\
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} + \\
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}, \\
&\quad \frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \left[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t} \right] < \psi(k, \mathcal{M}_{H,t+1})|
\end{aligned}
$$

$$
\begin{aligned}
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} \geq \\
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}, \\
&\quad \frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \Big[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1}) | \\
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} = \\
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1} | \\
&\quad \frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \Big[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1}), \\
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} \cdot \\
&Pr\{\frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \Big[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1}) | \\
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} \geq \\
&\left(1 - \frac{1}{k^2}\right) \cdot Pr\{\frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \Big[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1}) | \\
&\qquad\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\},
\end{aligned}
$$

where in the last inequality we used Equation 18, i.e. a version of the schema theorem.

Under any conditioning events the probability of the event $\{\frac{1}{(N-1)Mf_t^2} \cdot \sum_{i=1}^{N-1} \Big[m(L(H,i),t) f_{L(H,i),t} m(R(H,i),t) f_{R(H,i),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1})\}$ is necessarily greater than or equal to the probability of the event $\{\frac{1}{(N-1)Mf_t^2} \Big[m(L(H,i_H),t) f_{L(H,i_H),t} m(R(H,i_H),t) f_{R(H,i_H),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1})\}$ for any choice of the constant $i_H \in \{1, \ldots, N-1\}$, provided that the fitness function is non-negative.[5] Therefore, from the previous equations we obtain:

$$
\begin{aligned}
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1} | \\
&\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\} \geq \\
&\left(1 - \frac{1}{k^2}\right) \cdot Pr\{\frac{m(L(H,i_H),t) f_{L(H,i_H),t} m(R(H,i_H),t) f_{R(H,i_H),t}}{(N-1)Mf_t^2} \geq \psi(k, \mathcal{M}_{H,t+1}) | \\
&\qquad\qquad \bar{f}(t) = f_t, f(L(H,1),t) = f_{L(H,1),t}, f(R(H,1),t) = f_{R(H,1),t}, \ldots\}.
\end{aligned}
$$

Since the probability of the event $\{\frac{1}{(N-1)Mf_t^2} \Big[m(L(H,i_H),t) f_{L(H,i_H),t} m(R(H,i_H),t) f_{R(H,i_H),t}\Big] \geq \psi(k, \mathcal{M}_{H,t+1})\}$ being true does not depend on the value taken by the fitness of the schemata $L(H,i)$ and $R(H,i)$ for $i \neq i_H$, we could rewrite Equation 18 and all calculations produced in this section making all events conditional only to $\{\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\}$ obtaining

$$
Pr\{m(H,t+1) > \mathcal{M}_{H,t+1} |
$$

[5]The 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 remaining equations present in this section may be very pessimistic. However, at this stage we are not interested in the accuracy of our bounds.

$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} \geq$$
$$\left(1 - \frac{1}{k^2}\right) \cdot Pr\{m(L(H,i_H),t)m(R(H,i_H),t) \geq \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}} |$$
$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\}.$$

The event on the right-hand side of this equation is not of the same form as the event on the left-hand side of the equation. So, it is not possible to use this result recursively. However, further simplifications can be obtained by noticing that if $m(L(H,i_H),t)$ and $m(R(H,i_H),t)$ are both greater than two suitably large constants, $\mathcal{M}_{L(H,i_H),t}$ and $\mathcal{M}_{R(H,i_H),t}$, the product $m(L(H,i_H),t)m(R(H,i_H),t)$ will always be greater than $\frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}$ (although obviously this can happen also when either $m(L(H,i_H),t)$ or $m(R(H,i_H),t)$ are not greater than $\mathcal{M}_{L(H,i_H),t}$ and $\mathcal{M}_{R(H,i_H),t}$). In order for this to happen the two constants need to be such that $\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}$. Therefore,

$$Pr\{m(L(H,i_H),t)m(R(H,i_H),t) \geq \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}} |$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} \geq$$
$$Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}, m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t} |$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\}$$

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

$$\Pr\{m(H,t+1) > \mathcal{M}_{H,t+1} |$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} \geq$$
$$\left(1 - \frac{1}{k^2}\right) \cdot Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}, m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t} |$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\}.$$

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 $\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}, m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}\}$. If the population at generation t was fixed, these two events would 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). Then we would have that $\Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}, m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}\} = \Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}\} \cdot \Pr\{m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}\}$. However, in this work we do not assume that the population is fixed (this would be equivalent to considering α as a constant rather than a stochastic variable). So, we cannot make any assumption on the independence of the events $\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}\}$ and $\{m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}\}$. Fortunately, there is one 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 be trivially 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:

$$\begin{aligned}
&Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}, m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}| \\
&\qquad \mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}, \\
&\qquad \bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} \geq \\
&Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}| \\
&\qquad \mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}, \\
&\qquad \bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} + \\
&Pr\{m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}| \\
&\qquad \mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}, \\
&\qquad \bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} - 1.
\end{aligned}$$

This in turn leads to the following:

Theorem 6 Conditional Recursive Schema Theorem. *For a schema H under fitness proportionate selection, one-point crossover applied with 100% probability and no mutation,*

$$\begin{aligned}
&Pr\{m(H,t+1) > \mathcal{M}_{H,t+1}| \\
&\qquad \mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}}, \\
&\qquad \bar{f}(t) = f_t, f(L(H,i_H),t) = f_{L(H,i_H),t}, f(R(H,i_H),t) = f_{R(H,i_H),t}\} \geq
\end{aligned}$$

$$\left(1-\frac{1}{k^2}\right)\cdot\Big(\Pr\{m(L(H,i_H),t) > \mathcal{M}_{L(H,i_H),t}|$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t)=f_t, f(L(H,i_H),t)=f_{L(H,i_H),t}, f(R(H,i_H),t)=f_{R(H,i_H),t}\} +$$
$$\Pr\{m(R(H,i_H),t) > \mathcal{M}_{R(H,i_H),t}|$$
$$\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} > \frac{\psi(k,\mathcal{M}_{H,t+1})(N-1)Mf_t^2}{f_{L(H,i_H),t}f_{R(H,i_H),t}},$$
$$\bar{f}(t)=f_t, f(L(H,i_H),t)=f_{L(H,i_H),t}, f(R(H,i_H),t)=f_{R(H,i_H),t}\} - 1\Big).$$

This theorem represents the result we wanted. Indeed, it is recursive in the sense that with appropriate additional conditioning events (necessary to restrict the fitness of the building blocks of the building blocks[6] of the schema H and to make sure appropriate constants are used in the events dealing with their numbers) the theorem can be applied again to the events in the right-hand side of the previous equation, and then again the right-hand side of the resulting expressions and so on.

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 a better lower bound would be 0, whereby the theorem would simply state an obvious property of all probabilities.[7] Despite this weakness, this 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 the current generation. In order to illustrate this, let us consider an example.

7 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 $H=S=b_1b_2b_3b_4$ in the population by generation 3, in 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 previous theorem with $t=2$ and $\mathcal{M}_{H,t+1}=\mathcal{M}_{b_1b_2b_3b_4,3}=0$.

If we set $k=2$, we obtain $\psi(k,\mathcal{M}_{H,t+1})(N-1)M = 12M$. So, if we choose $i_H = i_{b_1b_2b_3b_4} = 2$, in order for the theorem to be applicable we need to make sure that $\mathcal{M}_{L(H,i_H),t}\mathcal{M}_{R(H,i_H),t} = \mathcal{M}_{\mathtt{b_1b_2}**,2}\mathcal{M}_{**\mathtt{b_3b_4},2} > \frac{12Mf_2^2}{f_{\mathtt{b_1b_2}**,2}f_{**\mathtt{b_3b_4},2}}$. A possible way to do that is to set $\mathcal{M}_{\mathtt{b_1b_2}**,2} = \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b_1b_2}**,2}}$ and $\mathcal{M}_{**\mathtt{b_3b_4},2} = \frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b_3b_4},2}}$, which, substituted into the theorem, lead to the following result:

$$\Pr\{m(\mathtt{b_1b_2b_3b_4},3)>0|\bar{f}(2)=f_2, f(\mathtt{b_1b_2}**,2)=f_{\mathtt{b_1b_2}**,2}, f(**\mathtt{b_3b_4},2)=f_{**\mathtt{b_3b_4},2}\} \geq$$

[6] Not a typo.

[7] It seems possible to obtain a much stronger result by computing the joint probability distribution of $m(L(H,i),t)$ and $m(R(H,i),t)$ and using it to replace the Bonferroni inequality with a stronger bound. However, we have not explored the viability of this approach fully, yet.

$$0.75 \cdot \Big(\Pr\{m(\mathtt{b}_1\mathtt{b}_2**,2) > \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}} | \bar{f}(2)=f_2, f(\mathtt{b}_1\mathtt{b}_2**,2)=f_{\mathtt{b}_1\mathtt{b}_2**,2}, f(**\mathtt{b}_3\mathtt{b}_4,2)=f_{**\mathtt{b}_3\mathtt{b}_4,2}\} +$$
$$\Pr\{m(**\mathtt{b}_3\mathtt{b}_4,2) > \frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b}_3\mathtt{b}_4,2}} | \bar{f}(2)=f_2, f(\mathtt{b}_1\mathtt{b}_2**,2)=f_{\mathtt{b}_1\mathtt{b}_2**,2}, f(**\mathtt{b}_3\mathtt{b}_4,2)=f_{**\mathtt{b}_3\mathtt{b}_4,2}\} - 1\Big).$$

This 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 (more on this later). In any case the recursive schema theorem presented earlier can be applied again to such probabilities.

For example let us calculate a lower bound for the probability that there will be at least $\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}}$ instances of the schema $H = b_1b_2**$ in the population at generation 2, in 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 previous theorem with $t=1$ and $\mathcal{M}_{H,t+1} = \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}}$.

If we set $k=2$ again,[8] we obtain $\psi(k,\mathcal{M}_{H,t+1})(N-1)M = 3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})$. So, if we choose $i_H = 1$, $\mathcal{M}_{b_1***,2} = \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{b_1***,1}}$ and $\mathcal{M}_{*b_2**,2} = \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{*b_2**,1}}$, which, substituted into the conditional recursive schema theorem, lead to the following result:

$$\begin{aligned}
&\Pr\{m(b_1b_2**,2) > \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}} | \bar{f}(1)=f_1, f(\mathtt{b}_1***,1)=f_{\mathtt{b}_1***,1}, f(*\mathtt{b}_2**,1)=f_{*\mathtt{b}_2**,1},\\
&\quad \bar{f}(2)=f_2, f(\mathtt{b}_1\mathtt{b}_2**,2)=f_{\mathtt{b}_1\mathtt{b}_2**,2}, f(**\mathtt{b}_3\mathtt{b}_4,2)=f_{**\mathtt{b}_3\mathtt{b}_4,2}\} \geq\\
&0.75\cdot\Big(\Pr\{m(\mathtt{b}_1***,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{b_1***,1}} |\\
&\quad \bar{f}(1)=f_1, f(\mathtt{b}_1***,1)=f_{\mathtt{b}_1***,1}, f(*\mathtt{b}_2**,1)=f_{*\mathtt{b}_2**,1},\\
&\quad \bar{f}(2)=f_2, f(\mathtt{b}_1\mathtt{b}_2**,2)=f_{\mathtt{b}_1\mathtt{b}_2**,2}, f(**\mathtt{b}_3\mathtt{b}_4,2)=f_{**\mathtt{b}_3\mathtt{b}_4,2}\} +\\
&\quad\Pr\{m(*\mathtt{b}_2**,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{*b_2**,1}} |\\
&\quad \bar{f}(1)=f_1, f(\mathtt{b}_1***,1)=f_{\mathtt{b}_1***,1}, f(*\mathtt{b}_2**,1)=f_{*\mathtt{b}_2**,1},\\
&\quad \bar{f}(2)=f_2, f(\mathtt{b}_1\mathtt{b}_2**,2)=f_{\mathtt{b}_1\mathtt{b}_2**,2}, f(**\mathtt{b}_3\mathtt{b}_4,2)=f_{**\mathtt{b}_3\mathtt{b}_4,2}\} - 1\Big).
\end{aligned}$$

A similar equation holds for the schema $H = **b_3b_4$, for which we will assume to use $i_H = 3$. 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:

$$\begin{aligned}
&\Pr\{m(b_1b_2b_3b_4,3) > 0|\mathcal{F}\} \geq\\
&0.75\cdot\Big(0.75\cdot\Big(\Pr\{m(\mathtt{b}_1***,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{b_1***,1}}|\mathcal{F}\}+
\end{aligned}$$

[8]Nothing prevents one from using different k's for different schemata when applying the recursive schema theorem.

$$\Pr\{m(*\mathtt{b}_2**,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{*b_2**,1}}|\mathcal{F}\}-1\Big) +$$
$$0.75\cdot\Big(\Pr\{m(**\mathtt{b}_3*,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b}_3\mathtt{b}_4,2}})+1}\frac{f_1}{f_{**b_3*,1}}|\mathcal{F}\} +$$
$$\Pr\{m(***\mathtt{b}_4,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b}_3\mathtt{b}_4,2}})+1}\frac{f_1}{f_{***b_4,1}}|\mathcal{F}\}-1\Big)-1\Big),$$

whereby

$$\begin{aligned}
&\Pr\{m(b_1b_2b_3b_4,3)>0|\mathcal{F}\}\geq\\
&0.5625\cdot\Big(\Pr\{m(\mathtt{b}_1***,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{b_1***,1}}|\mathcal{F}\}+\\
&\Pr\{m(*\mathtt{b}_2**,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{\mathtt{b}_1\mathtt{b}_2**,2}})+1}\frac{f_1}{f_{*b_2**,1}}|\mathcal{F}\}+\\
&\Pr\{m(**\mathtt{b}_3*,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b}_3\mathtt{b}_4,2}})+1}\frac{f_1}{f_{**b_3*,1}}|\mathcal{F}\}+\\
&\Pr\{m(***\mathtt{b}_4,1) > \sqrt{3M\psi(2,\frac{\sqrt{12M+1}f_2}{f_{**\mathtt{b}_3\mathtt{b}_4,2}})+1}\frac{f_1}{f_{***b_4,1}}|\mathcal{F}\}\Big)-1.875
\end{aligned}\tag{19}$$

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. In all other cases we get 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 (more on this in the following section). Therefore, it is not impossible for the events in the right-hand side of the equation to be all true. In this case the lower bound 0.375 on the probability of success can be increased by performing multiple runs (as indicated in Section 5. For example, it can be increased to more than 99% by performing $R = 10$ runs.

8 Population Sizing

In this section we want to discuss how 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 do this continuing the example presented in the previous section.

For the sake of simplicity, let us assume that we initialise the population making sure that $m(0*\cdots*,1) = m(1*\cdots*,1) = m(*0*\cdots*,1) = m(*1*\cdots*,1) = \ldots = m(*\cdots*0,1) = m(*\cdots*1,1) = M/2$.

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 19.[9] To achieve this one would have

[9]This is by no means neither the only nor the best way to size the population. For example, a better way might be to choose M so as to minimise the total computational effort $R\times M\times G$, G being the maximum number of generations allowed in each run. However, the method described in this section is one of the simplest, and the strategy described can be extended to other population sizing methods.

to make sure that each of the four events in the equation are the case. Let us start from the first one of them: $\{m(\mathtt{b_1}***,1) > \sqrt{3M\psi(2, \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b_1b_2}**,2}}) + 1}\frac{f_1}{f_{b_1***,1}}|\mathcal{F}\}$.

Since $m(\mathtt{b_1}***,1) = M/2$, the event happens with probability 1 if

$$\frac{M}{2} > \sqrt{3M\psi(2, \frac{\sqrt{12M+1}f_2}{f_{\mathtt{b_1b_2}**,2}}) + 1}\frac{f_1}{f_{b_1***,1}}.$$

Since we have assumed that f_1, $f_{b_1***,1}$, f_2 and $f_{\mathtt{b_1b_2}**,2}$ are known, the equation can be (numerically) solved for M obtaining a value $M(f_1, f_{b_1***,1}, f_2, f_{\mathtt{b_1b_2}**,2})$. This is the minimum value of M which guarantees that the first event in Equation 19 will happen with probability 1.

The same procedure can be repeated for the other events in Equation 19, obtaining the lower bounds $M(f_1, f_{*b_2**,1}, f_2, f_{\mathtt{b_1b_2}**,2})$, $M(f_1, f_{**b_3*,1}, f_2, f_{**\mathtt{b_3b_4},2})$ and $M(f_1, f_{***b_4,1}, f_2, f_{**\mathtt{b_3b_4},2})$. Therefore, the minimum population size that maximises the right-hand side of Equation 19 is

$$\begin{aligned} M_{min} \quad = \quad & \lceil \max[M(f_1, f_{b_1***,1}, f_2, f_{\mathtt{b_1b_2}**,2}), M(f_1, f_{*b_2**,1}, f_2, f_{\mathtt{b_1b_2}**,2}), \\ & M(f_1, f_{**b_3*,1}, f_2, f_{**\mathtt{b_3b_4},2}), M(f_1, f_{***b_4,1}, f_2, f_{**\mathtt{b_3b_4},2})]\rceil. \end{aligned}$$

Of course, given the known weaknesses of the bounds used to derive the recursive schema theorem and, hence, Equation 19, it has to be expected that the population size proposed by the previous equation will be much larger than necessary. To give a feel for the values suggested by the equation, let us imagine that the ratios between order-1 building block fitness and population fitness ($f_{b_1***,1}/f_1$, $f_{*b_2**,1}/f_1$, etc.) be constant and equal to r_1 and that the ratios between order-2 building block fitness and population fitness ($f_{\mathtt{b_1b_2}**,2}/f_2$ and $f_{**\mathtt{b_3b_4},2}/f_2$) be constant and equal to r_2. The following table shows the values of M_{min} resulting from different values of r_1 and r_2:

	r_2				
r_1	1	2	3	4	5
1	2,353	769	433	301	233
2	193	76	49	37	31
3	49	22	15	13	11
4	19	10	7	6	6
5	10	5	4	4	4

When both fitness ratios are 1 (for example because the fitness landscape is flat) the population size suggested by the previous equation is huge considering that the length of bit-strings in the population is only 4. However, if the building blocks of the solution have well above average fitness, more realistic population sizes are suggested.

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 above depend significantly more on the order-1 building-block fitness ratio r_1 than on r_2. This suggests 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.

9 Conclusions

In this paper we have developed a new form of schema theorem in which expectations are not present. This theorem allows 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.

By utilising in an unusual way this version of the schema theorem backwards in time, i.e. to predict the past from the future, we have been able to obtain 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 on the hypothesis that building block and population fitnesses are known. As an example, in the paper we have shown how such conditions can be derived for a generic 4-bit problem.

The results reported in this paper provide interesting insights for a theoretical understanding of GA's by making explicit the relation between the probability of finding a solution, schema fitness and population size.

These results are important because for the first time they make explicit the relation between population size, schema fitness and probability of convergence over multiple generations. This has allowed us to identify rigorous strategies to size the population and therefore to calculate the computational effort required to solve a given problem using a GA.

All the results in this paper are based on the assumption that the fitness of the building block involved in the process of finding a solution and the population fitness are known at each generation. Since these are fitness-dependent characteristics, our results do not represent a full schema-theorem-based proof of convergence for GAs. In addition, a lot needs to be done to improve the tightness of the lower bounds obtained in the paper. 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.

Despite these limitations, we believe this work shows 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 for useful comments and discussion.

References

Altenberg, Lee (1995). The Schema Theorem and Price's Theorem. In: *Foundations of Genetic Algorithms 3* (L. Darrell Whitley and Michael D. Vose, Eds.). Morgan Kaufmann. Estes Park, Colorado, USA. pp. 23–49.

Chernoff, Herman (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.

Davis, Thomas E. and Jose C. Principe (1993). A markov chain framework for the simple genetic algorithm. *Evolutionary Computation* **1**(3), 269–288.

De Jong, Kenneth A., William M. Spears and Diana F. Gordon (1995). Using Markov chains to analyze GAFOs. In: *Proceedings of the Third Workshop on Foundations of Genetic Algorithms* (L. Darrell Whitley and Michael D. Vose, Eds.). Morgan Kaufmann. San Francisco. pp. 115–138.

Fogel, D. B. and A. Ghozeil (1997). Schema processing under proportional selection in the presence of random effects. *IEEE Transactions on Evolutionary Computation* **1**(4), 290–293.

Fogel, D. B. and A. Ghozeil (1998). The schema theorem and the misallocation of trials in the presence of stochastic effects. In: *Evolutionary Programming VII: Proc. of the 7th Ann. Conf. on Evolutionary Programming* (V.W. Porto, N. Saravanan, D. Waagen and A.E. Eiben, Eds.). Springer. Berlin. pp. 313–321.

Fogel, D. B. and A. Ghozeil (1999). Schema processing, proportional selection, and the misallocation of trials in genetic algorithms. *Information Sciences*.

Goldberg, D. E. (1998). A threefold decomposition of GA problem difficulty and its core. Seminar at the School of Computer Science, The University of Birmingham, UK.

Goldberg, D. E. and Mike Rudnick (1991). Genetic algorithms and the variance of fitness. Technical Report IlliGAL Report No 91001. Department of General Engineering, University of Illinois at Urbana-Champaign.

Goldberg, D. E., K. Deb and D. Thierens (1993). Toward a better understanding of mixing in genetic algorithms. *Journal of the Society of Control Engineers (SICE)* **32**(1), 10–16.

Goldberg, D. E., K. Deb and J. H. Clark (1991). Genetic algorithms, noise, and the sizing of populations. Technical Report IlliGAL Report No 91010. Department of General Engineering, University of Illinois at Urbana-Champaign.

Goldberg, David E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. Reading, Massachusetts.

Hoeffding, Wassily (1963). probability inequalities for sums of bonded random variables. *journal of the American statistical association* **58**(301), 13–30.

Holland, John (1992). *Adaptation in Natural and Artificial Systems*. second ed.. MIT Press. Cambridge, Massachusetts.

Macready, William G. and David H. Wolpert (1996). On 2-armed gaussian bandits and optimization. Sante Fe Institute Working Paper 96-05-009.

Nix, Allen E. and Michael D. Vose (1992). Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence* **5**, 79–88.

Poli, Riccardo (1999*a*). New results in the schema theory for GP with one-point crossover which account for schema creation, survival and disruption. Technical Report CSRP-99-18. University of Birmingham, School of Computer Science.

Poli, Riccardo (1999*b*). Schema theorems without expectations. In: *Proceedings of the Genetic and Evolutionary Computation Conference* (Wolfgang Banzhaf, Jason Daida, Agoston E. Eiben, Max H. Garzon, Vasant Honavar, Mark Jakiela and Robert E. Smith, Eds.). Vol. 1. Morgan Kaufmann. Orlando, Florida, USA. p. 806.

Poli, Riccardo (1999*c*). Schema theory without expectations for GP and GAs with one-point crossover in the presence of schema creation. In: *Foundations of Genetic Programming* (Thomas Haynes, William B. Langdon, Una-May O'Reilly, Riccardo Poli and Justinian Rosca, Eds.). Orlando, Florida, USA.

Poli, Riccardo and W. B. Langdon (1997*a*). An experimental analysis of schema creation, propagation and disruption in genetic programming. In: *Genetic Algorithms: Proceedings of the Seventh International Conference* (Thomas Back, Ed.). Morgan Kaufmann. Michigan State University, East Lansing, MI, USA. pp. 18–25.

Poli, Riccardo and W. B. Langdon (1997*b*). A new schema theory for genetic programming with one-point crossover and point mutation. In: *Genetic Programming 1997: Proceedings of the Second Annual Conference* (John R. Koza, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max Garzon, Hitoshi Iba and Rick L. Riolo, Eds.). Morgan Kaufmann. Stanford University, CA, USA. pp. 278–285.

Poli, Riccardo and William B. Langdon (1998). Schema theory for genetic programming with one-point crossover and point mutation. *Evolutionary Computation* **6**(3), 231–252.

Poli, Riccardo, William B. Langdon and Una-May O'Reilly (1998). Analysis of schema variance and short term extinction likelihoods. In: *Genetic Programming 1998: Proceedings of the Third Annual Conference* (John R. Koza, Wolfgang Banzhaf, Kumar Chellapilla, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max H. Garzon, David E. Goldberg, Hitoshi Iba and Rick Riolo, Eds.). Morgan Kaufmann. University of Wisconsin, Madison, Wisconsin, USA. pp. 284–292.

Radcliffe, Nicholas J. (1997). Schema processing. In: *Handbook of Evolutionary Computation* (T. Baeck, D. B. Fogel and Z. Michalewicz, Eds.). pp. B2.5–1–10. Oxford University Press.

Rosca, Justinian P. (1997). Analysis of complexity drift in genetic programming. In: *Genetic Programming 1997: Proceedings of the Second Annual Conference* (John R. Koza, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max Garzon, Hitoshi Iba and Rick L. Riolo, Eds.). Morgan Kaufmann. Stanford University, CA, USA. pp. 286–294.

Rudolph, Günter (1994). Convergence analysis of canonical genetic algorithm. *IEEE Transactions on Neural Networks* **5**(1), 96–101.

Rudolph, Günter (1997*a*). Genetic algorithms. In: *Handbook of Evolutionary Computation* (T. Baeck, D. B. Fogel and Z. Michalewicz, Eds.). pp. B2.4–20–27. Oxford University Press.

Rudolph, Günter (1997*b*). Modees of stochastic convergence. In: *Handbook of Evolutionary Computation* (T. Baeck, D. B. Fogel and Z. Michalewicz, Eds.). pp. B2.3–1–3. Oxford University Press.

Rudolph, Günter (1997*c*). Stochastic processes. In: *Handbook of Evolutionary Computation* (T. Baeck, D. B. Fogel and Z. Michalewicz, Eds.). pp. B2.2–1–8. Oxford University Press.

Schmidt, J. P., A. Siegel and A. Srinivasan (1992). Chernoff-hoeffding bounds for applications with limited independence. Technical Report 92-1305. department of computer science, Cornell University.

Sobel, M. and V. R. R. Uppuluri (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, Murray R. (1975). *Probability and Statistics*. McGraw-Hill. New York.

Stephens, C. R. and H. Waelbroeck (1997). Effective degrees of freedom in genetic algorithms and the block hypothesis. In: *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)* (Thomas Bäck, Ed.). Morgan Kaufmann. East Lansing.

Stephens, C. R. and H. Waelbroeck (1999). Schemata evolution and building blocks. *Evolutionary Computation* **7**(2), 109–124.