Exact Schema Theorems for GP with One-Point and Standard Crossover Operating on Linear Structures and their Application to the Study of the Evolution of Size

Riccardo Poli[1] and Nicholas Freitag McPhee[2]

[1] School of Computer Science, The University of Birmingham
Birmingham, B15 2TT, UK
R.Poli@cs.bham.ac.uk, http://www.cs.bham.ac.uk/~rmp/
[2] Division of Science and Mathematics, University of Minnesota, Morris
Morris, MN, USA
mcphee@mrs.umn.edu, http://www.mrs.umn.edu/~mcphee

Abstract. In this paper, firstly we specialise the exact GP schema theorem for one-point crossover to the case of linear structures of variable length, for example binary strings or programs with arity-1 primitives only. Secondly, we extend this to an exact schema theorem for GP with standard crossover applicable to the case of linear structures. Then we study, both mathematically and numerically, the schema equations and their fixed points for infinite populations for both a constant and a length-related fitness function. This allows us to characterise the bias induced by standard crossover. This is very peculiar. In the case of a constant fitness function, at the fixed-point, structures of any length are present with non-zero probability. However, shorter structures are sampled exponentially much more frequently than longer ones.

1 Introduction

In recent work [6] an exact schema theorem for GP with one-point crossover has been introduced. This gives an exact expression for the expected number of instances of a schema H at generation $t+1$, $E[m(H,t+1)]$, in terms of macroscopic quantities (i.e. properties of schemata, like their fitness or number of instances, as opposed to microscopic properties of the individuals in the population, like their selection probability) measured at generation t.

The theorem has the form $E[m(H,t+1)] = M\alpha(H,t)$, where M is the number of individuals in the population and $\alpha(H,t)$, which we term the *total transmission probability* of H, is the probability that an individual created through the selection/crossover/mutation process samples H [12]. [6] provides an exact value of $\alpha(H,t)$ for a GP system with *one-point crossover* [9, 11]. This operator works by selecting a common crossover point in the parent programs and then swapping the corresponding subtrees, like standard crossover. To account for the possible structural diversity of the two parents, one-point crossover selects the crossover point only in the part of the two trees which have the same topology. This is called the *common region*.

The theory is based on the definition of GP schema proposed in [9] in which a *schema* is a tree composed of functions from the set $\mathcal{F} \cup \{=\}$ and terminals from the

set $\mathcal{T} \cup \{=\}$, where $\mathcal{F}$ and $\mathcal{T}$ are the function set and the terminal set used in a GP run. The symbol $=$ is a "don't care" symbol which stands for a *single* terminal or function. A schema H represents programs having the same shape as H and the same labels for the non-$=$ nodes. In order to be able to represent programs of different sizes and shapes the theory requires also the definition of the concept of hyperschema [7]. A *GP hyperschema* is a rooted tree composed of functions from the set $\mathcal{F} \cup \{=\}$ and terminals from the set $\mathcal{T} \cup \{=, \#\}$. The symbol = is as above, while the symbol # stands for any valid subtree. The notion of hyperschema is a generalisation of both the GP schemata defined above (which are hyperschemata without # symbols) and Rosca's schemata [13] (which are hyperschemata without = symbols).

With these definitions one can prove [6] that the total transmission probability for a GP schema H under one-point crossover and no mutation is:

$$\alpha(H,t) = (1-p_{xo})p(H,t) + p_{xo} \sum_j \sum_k \frac{1}{\mathbf{NC}(G_j,G_k)} \cdot \sum_{i \in C(G_j,G_k)} p(L(H,i) \cap G_j, t) p(U(H,i) \cap G_k, t) \quad (1)$$

where: p_{xo} is the probability of crossover; $p(H,t)$ is the probability of selecting an individual matching the schema H; $G_1, G_2, \cdots$ are all the different schemata that can be built from = nodes only;[1] the indices j and k range over all the different G_i; $C(G_j,G_k)$ is the set of crossover points in the common region between schema G_j and schema G_k; $\mathbf{NC}(G_j,G_k) = |C(G_j,G_k)|$ is the number of nodes in the common region; the index i ranges over all the crossover points in $C(G_j,G_k)$; $L(H,i)$ is the hyperschema obtained by replacing all the nodes on the path between crossover point i and the root node with = nodes, and all the subtrees connected to these nodes with # nodes; $U(H,i)$ is the hyperschema obtained by replacing the subtree below crossover point i with a # node (more details on these definitions can be found in [6, 7]). In fitness proportionate selection, $p(H,t) = \frac{m(H,t)f(H,t)}{M\bar{f}(t)}$ where $m(H,t)$ is the number of programs matching the schema H at generation t, $f(H,t)$ is the mean fitness of the programs matching H, and $\bar{f}(t)$ is the mean fitness of the programs in the population. The hyperschemata $L(H,i)$ and $U(H,i)$ are important because they allow the identification of parents that may lead to the creation of offspring in H: if one crosses over at point i *any* individual in $L(H,i)$ with *any* individual in $U(H,i)$, the resulting offspring is always an instance of H.

In [6] an example was presented which used only unary functions. With this function set only linear trees can be created and, therefore, GP becomes a sort of variable-length GA. In that example, the new schema theorem was applied to a specific schema and a specific population. In this paper we study in much greater depth what happens in a GP system handling linear structures. We start by specialising the schema theorem for one-point crossover to such a case, but without focusing on any particular schema (as was the case in the example in [6]), and by then extending this to an exact schema theorem for GP with standard crossover applicable to the case of linear structures.

[1] Thus the G_i can be seen as cataloguing all the different program shapes.

2 Exact Schema Theory for Linear Structures

When only unary functions are used in GP, schemata (and programs) can only take the form $(h_1(h_2(h_3....(h_{N-1}h_N)....)))$ where $N > 0$, $h_i \in \mathcal{F} \cup \{=\}$ for $1 \leq i < N$, and $h_N \in \mathcal{T} \cup \{=\}$. Therefore, they can be written unambiguously as strings of symbols of the form $h_1h_2h_3....h_{N-1}h_N$.

In order to make the specialisation of Equation 1 to the linear case easier we represent repeated symbols in a string using the power notation where x^y means x repeated y times. For example, the schema `11100000===1` can be written as $1^3 0^5 (=)^3 1$. Since in this case all trees are linear, the space of program shapes can be enumerated by $\{G_n\}$ where G_n is $(=)^n$ for $n > 0$. Given this, the common region between shapes G_j and G_k is simply the shorter of the two schemata, and the size of the common region, $\mathbf{NC}(G_j, G_k)$, is simply $\min(j, k)$. Therefore, the set of crossover points in the common region, $C(G_j, G_k)$, can be identified with the set of indices $\{0, 1, ..., \min(j, k)-1\}$ where the index 0 represent a crossover point before the first symbol in a string (the root node). In this linear representation the hyperschemata $L(H, i)$ and $U(H, i)$ are particularly simple: $U(H, i)$ is $h_1...h_i\#$ and $L(H, i)$ is $(=)^i h_{i+1}...h_N$, where $0 \leq i < N$ (with the convention that for $i = 0$ the hyperschema $h_1...h_i\#$ is simply $\#$). So,

$$U(H,i) \cap G_k \equiv \begin{cases} h_1...h_i(=)^{k-i} & \text{if } i < k, \\ \emptyset & \text{otherwise,} \end{cases}$$

and

$$L(H,i) \cap G_j \equiv \begin{cases} (=)^i h_{i+1}...h_N & \text{if } j = N, \\ \emptyset & \text{otherwise,} \end{cases}$$

where $\emptyset$ is the empty set. Therefore, the summation in j in Equation 1 disappears because $p(L(H, i) \cap G_j, t) = 0$ for all $j \neq N$. As a result of these simplifications, one can transform Equation 1 into:

Theorem 1. *The total transmission probability for a linear GP schema of the form $h_1...h_N$ under one-point crossover and no mutation is*

$$\alpha(h_1...h_N, t) = (1 - p_{xo})p(h_1...h_N, t) + p_{xo} \cdot \sum_k \frac{1}{\min(N,k)} \cdot \sum_{i=0}^{\min(N,k)-1} p(h_1...h_i(=)^{k-i}, t)p((=)^i h_{i+1}...h_N, t). \quad (2)$$

This result can then be extended to obtain the following:

Theorem 2. *The total transmission probability for a linear GP schema of the form $h_1...h_N$ under standard crossover with uniform selection of the crossover points and no mutation is*

$$\alpha(h_1...h_N, t) = (1 - p_{xo})p(h_1...h_N, t) + p_{xo} \sum_k \frac{1}{k} \sum_{i=0}^{\min(N,k)-1} p(h_1...h_i(=)^{k-i}, t) \sum_{n=N-i}^{\infty} \frac{p((=)^{n-N+i} h_{i+1}...h_N, t)}{n}. \quad (3)$$

Proof. The theorem can be derived as a special case of the more general result reported in [8]. However, here we provide an direct proof which shows how this result can be derived by modifying Equation 2.

Equation 2 clearly indicates that one-point crossover with a given crossover point i can create new instances of the schema $h_1...h_N$ only if selection picks up a first parent whose first i nodes match $h_1...h_i$ and a second parent whose last $N-i$ nodes match $h_{i+1}...h_N$. Given that one-point crossover forces the selection of a common crossover point, this means that the second parent must be always of length N. However, if one used standard crossover, one could create instances of $h_1...h_N$ even if the length of the second parent is different from N, provided that the last $N-i$ nodes of the second parent match $h_{i+1}...h_N$ and the second crossover point excised such nodes. If the second crossover point is at position l, this can happen if selection picks up a second parent matching $(=)^l h_{i+1}...h_N$ for any value of $l \geq 0$. Thus, an extra summation needs to be added to Equation 2 to deal with standard crossover. The probability of choosing crossover point l in a second parent matching $(=)^l h_{i+1}...h_N$ is $1/(l+N-i)$. Therefore, after the change of variable $n = l+N-j$ one obtains that the probability of obtaining the subschema $h_{i+1}...h_N$ from the second parent is $\sum_{n=N-i}^{\infty} \frac{p((=)^{n-N+i}h_{i+1}...h_N,t)}{n}$. This should replace the term $p((=)^i h_{i+1}...h_N,t)$ in Equation 2. Since standard crossover does not limit the crossover point in the first parent to belong to the common region, the probability of selecting crossover point i needs to change from $1/\min(N,k)$ to $1/k$. This completes the proof of the theorem. □

Equation 3 is in a form which makes it easy to see the similarities with Equation 2. However, the reader might find it easier to understand the same result rewritten as in the following:

Corollary 1. *The total transmission probability for a linear GP schema of the form $h_1...h_N$ under standard crossover with uniform selection of the crossover points and no mutation can be written in the following equivalent forms:*

$$\alpha(h_1...h_N,t) = (1-p_{xo})p(h_1...h_N,t) + p_{xo}\sum_{i=0}^{N-1}\sum_{k>i}\sum_{n\geq N-i}\frac{1}{kn}p(h_1...h_i(=)^{k-i},t)p((=)^{n-N+i}h_{i+1}...h_N,t), \quad (4)$$

$$\alpha(h_1...h_N,t) = (1-p_{xo})p(h_1...h_N,t) + p_{xo}\sum_{i=0}^{N-1}\sum_{k>0}\sum_{n\geq 0}\frac{1}{(k+i)(n-i+N)}p(h_1...h_i(=)^{k},t)p((=)^{n}h_{i+1}...h_N,t). \quad (5)$$

Equation 4 makes the idea of summing over the set of possible crossover points clearer. Equation 5 makes the idea of summing over varying lengths of "don't care" symbols clearer.

3 Evolution of Size in Linear Systems

Equations 2 and 3 can be used to study, among other things, the evolution of size in linear GP/GA systems. This is because they can be specialised to describe the transmis-

sion probability of schemata of the form $(=)^N$. For one-point crossover (Equation 2) one obtains:

$$\begin{aligned}\alpha((=)^N,t) &= (1-p_{xo})p((=)^N,t) + p_{xo}\sum_k \sum_{i=0}^{\min(N,k)-1} \frac{p((=)^k,t)p((=)^N,t)}{\min(N,k)} \\ &= (1-p_{xo})p((=)^N,t) + p_{xo}p((=)^N,t)\sum_k p((=)^k,t) \\ &= p((=)^N,t), \end{aligned} \tag{6}$$

where we exploited the fact that $\sum_k p((=)^k,t) = 1$ since $\bigcup_k (=)^k$ represents the entire search space and $(=)^k \cap (=)^j$ is the empty set for any $k \neq j$. This result indicates that length evolves under one-point crossover as if selection only was acting (p_{xo}, for example, has no effect). So, one-point crossover is totally unbiased with respect to program length. This is made particularly clear if one assumes a flat fitness landscape in which $f(H,t) = \bar{f}(t)$ for all H. In these conditions all the dynamics in the system must be caused by crossover or by sampling effects. For a flat landscape under fitness proportionate selection Equation 6 becomes $\alpha((=)^N,t) = m((=)^N,t)/M$ for finite populations, and $\alpha((=)^N,t) = \alpha((=)^N,t-1)$ in the infinite population limit. This is because, in the infinite population case, the quantity $\alpha(H,t)$ can be interpreted in two entirely equivalent ways: as the total transmission probability of the schema H at generation t or as the proportion of individuals in the population in H at generation $t+1$. In this second interpretation, we conventionally define $\alpha(H,-1)$ as the proportion of programs in H at generation 0. Clearly, $\alpha(H,-1)$ is entirely determined by the initialisation procedure adopted.

The equation $\alpha((=)^N,t) = \alpha((=)^N,t-1)$ obtained for the infinite population case is particularly important because it shows that when one-point crossover alone is acting, any initial distribution of lengths, $\alpha((=)^N,-1)$, is a fixed point for the system.

For standard crossover (Equation 3) one obtains:

$$\begin{aligned}\alpha((=)^N,t) = (1-p_{xo})p((=)^N,t) + \\ p_{xo}\sum_k \sum_{i=0}^{\min(N,k)-1} \frac{p((=)^k,t)}{k} \sum_{n=N-i}^{\infty} \frac{p((=)^n,t)}{n},\end{aligned} \tag{7}$$

which can be transformed into

$$\begin{aligned}\alpha((=)^N,t) = (1-p_{xo})p((=)^N,t) + p_{xo}\sum_k \frac{p((=)^k,t)}{k}\cdot \\ \sum_{n=\max(1,N-k+1)}^{\infty} \frac{p((=)^n,t)}{n}\min(N,k,n,k+n-N).\end{aligned} \tag{8}$$

Alternative formulations for this equation can be obtained specialising Corollary 1. For example, from Equation 4 one obtains:

$$\alpha((=)^N,t) = (1-p_{xo})p((=)^N,t) + p_{xo}\sum_{i=0}^{N-1}\sum_{k>i}\frac{p((=)^k,t)}{k}\sum_{n\geq N-i}\frac{p((=)^n,t)}{n}.$$

An alternative way of expressing $\alpha((=)^N, t)$ for standard crossover in terms of *microscopic* quantities is the following:

$$\alpha((=)^N, t) = (1 - p_{xo})p((=)^N, t) + \tag{9}$$
$$p_{xo} \sum_{z_1 \in P} \sum_{z_2 \in P} \sum_{x_1=0}^{N(z_1)-1} \sum_{x_2=0}^{N(z_2)-1} \frac{p(z_1, t)p(z_2, t)}{N(z_1)N(z_2)} \delta(x_1 + N(z_2) - x_2 = N)$$

in which P is the population at generation t, z_1 and z_2 vary over all the possible parents (i.e. the members of P) while x_1 and x_2 vary over all the possible crossover points in z_1 and z_2, respectively. This equation explicitly includes one term for each of the possible ways in which offspring can be created from the parents in the population for all possible crossover points. In the equation, the term $\frac{p(z_1,t)p(z_2,t)}{N(z_1)N(z_2)}$ represents the probability that each of such event be the case, while the term $\delta(x_1 + N(z_2) - x_2 = N)$ makes sure that only the probabilities of the events that lead to the creation of an offspring of length N are included in the sum.

These equations show that for standard crossover not every initial distribution of program lengths is a fixed point (for an infinite population) even if one considers the case of a flat landscape. For example, if one started at generation 0 with only programs of length X, i.e. $\alpha((=)^x, -1) = \delta(x = X)$, assuming $p_{xo} = 1$ one would obtain the following distribution of lengths at generation 1:

$$\alpha((=)^N, 0) = \max\left(\frac{X - |X - N|}{X^2}, 0\right). \tag{10}$$

This implies that $\alpha((=)^X, 0) = 1/X$ which is in general different from $\alpha((=)^X, -1) = 1$ (except for the trivial case in which $X = 1$).[2] So, standard crossover imposes its own specific bias on the distribution of lengths, although we expect that such a bias will have no influence on the average length of programs, since on average the subtrees/substrings swapped by crossover are of the same size. This conjecture can actually be proven mathematically obtaining the following:

Theorem 3. *The mean size of the programs at generation $t+1$, $\mu(t+1)$, in a linear GP system with standard crossover, uniform selection of the crossover points, no mutation and an infinite population is*

$$\mu(t+1) = \sum_N N p((=)^N, t) \tag{11}$$

with the same meaning of the symbols as in previous theorems.

Proof. By definition of mean $\mu(t+1) = \sum_N N\alpha((=)^N, t)$. By substituting Equation 9 into this equation one obtains:

$$\mu(t+1) = (1 - p_{xo}) \sum_N N p((=)^N, t) + p_{xo} \sum_N N \times$$

[2] If all programs include only one node (a terminal), standard crossover, like one-point crossover, cannot produce programs with more than one node.

$$\sum_{z_1 \in P} \sum_{z_2 \in P} \sum_{x_1=0}^{N(z_1)-1} \sum_{x_2=0}^{N(z_2)-1} \frac{p(z_1,t)p(z_2,t)}{N(z_1)N(z_2)} \delta(x_1 + N(z_2) - x_2 = N)$$

$$= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_{z_1 \in P} \sum_{z_2 \in P} \frac{p(z_1,t)p(z_2,t)}{N(z_1)N(z_2)} \times$$

$$\sum_{x_1=0}^{N(z_1)-1} \sum_{x_2=0}^{N(z_2)-1} \sum_N N\delta(x_1 + N(z_2) - x_2 = N).$$

With a few calculations it is possible to show that

$$\sum_{x_1=0}^{N(z_1)-1} \sum_{x_2=0}^{N(z_2)-1} \sum_N N\delta(x_1 + N(z_2) - x_2 = N) = N(z_1)N(z_2)\frac{N(z_1) + N(z_2)}{2}$$

whereby

$$\begin{aligned}
\mu(t+1) &= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_{z_1 \in P} \sum_{z_2 \in P} p(z_1,t)p(z_2,t)\frac{N(z_1) + N(z_2)}{2} \\
&= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_{z_1 \in P} p(z_1,t)\frac{N(z_1)}{2} \sum_{z_2 \in P} p(z_2,t) + \\
&\quad p_{xo} \sum_{z_2 \in P} p(z_2,t)\frac{N(z_2)}{2} \sum_{z_1 \in P} p(z_1,t) \\
&= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_{z \in P} p(z,t)N(z) \\
&= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_k N((=)^k) \cdot \sum_{z \in P \cap (=)^k} p(z,t) \\
&= (1 - p_{xo}) \sum_N Np((=)^N, t) + p_{xo} \sum_k kp((=)^k, t) \\
&= \sum_N Np((=)^N, t).
\end{aligned}$$

□

Corollary 2. *On a flat landscape,*

$$\mu(t+1) = \mu(t). \tag{12}$$

It is very difficult to find the fixed points (or to prove that there are none) for a GP system using standard crossover even on the hypothesis of infinite populations and flat landscapes. However, it is possible to find such fixed points numerically by implementing the schema equations and iterating them as described in the next section. We will show in Section 5 how the information provided by such simulations, along with a substantial amount of luck, has allowed us to identify a mathematical function which is provably a fixed point for the size distribution in a linear GP system with standard crossover operating on a flat landscape.

4 Experimental Results

In our experiments we were interested in studying the evolution of program size and the bias imposed by different crossovers. Since the effects of selection on its own have been already studied in great depth in various studies, we concentrated on the effects produced by crossover, i.e. by setting $p_{xo} = 1$. In the experiments we iterated the schema equations provided in the previous section on the assumption of infinite populations. This corresponds to studying either the exact behaviour of a GP system with an infinite population or the average behaviour of the GP system with a finite population over an infinite number of runs.

In the simulations we kept track of $\alpha((=)^N, t)$ for $N > 0$ such that $\alpha((=)^N, t) > 0$. Since in an infinite population standard crossover nearly doubles the length of the longest program in each generation, the number of N that need to be tracked grows exponentially as a function of t. As a result we restricted our attention to $t \leq 4$. Fortunately, this small number of generations was still sufficient to show the system convergence.

In the experiments we used three different initial conditions: the *one peak* distribution,

$$\alpha((=)^N, -1) = \delta(N = 20)$$

where only programs of length 20 are present, the *two peak* distribution

$$\alpha((=)^N, -1) = \begin{cases} 0.5 & \text{if } N = 14 \text{ or } N = 26, \\ 0 & \text{otherwise,} \end{cases}$$

and the *uniform* distribution

$$\alpha((=)^N, -1) = \begin{cases} 1/39 & \text{if } 0 < N < 40, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

These three distributions are all characterised by the same average lengths of 20.

In addition to a flat fitness landscape, realised by a fitness function which always returned the value 10, we used the following *size-related* fitness function

$$f(h_1 ... h_N) = (20 - |20 - N|)\delta(0 < N < 40) \tag{14}$$

which returns 20 minus the distance between the program length, N, and a target length of 20 for $0 < N < 40$, and 0 for $N \geq 40$.

On the flat fitness landscape one-point crossover behaved as expected. All initial distributions $\alpha((=)^N, -1)$ we tried were fixed points.

The situation was very different for standard crossover. As shown in Figure 1, with the *one-peak* initialisation, after one iteration (i.e. at $t = 1$) we obtained the triangular profile calculated in the previous section.[3] As hypothesised, in later generations the average program size remained constant. However, the distribution of sizes does not remain symmetric. In fact, it quickly approaches a limit distribution, where most of the

[3] In this and the following figures, the plots for generation 0 and 1 are shown with thin lines. The lines become progressively thicker as the number of generations increase. Because of the quick convergence to a fixed point, the plots for $t = 3$ and $t = 4$ are often indistinguishable.

programs were quite short. Exactly the same fixed point distribution was approached when initialising the system using the two-peak and the uniform distributions as shown in Figures 2 and 3. This fixed point distribution closely resembles a gamma distribution, a fact that is explored in more detail in the next section.

As a sanity check to make sure that these results were not an artifact of our numerical simulations, we performed real GP runs with a population of 10,000 individuals initialised with the uniform distribution of lengths in Equation 13. The fitness function was flat. The runs were continued for 250 generations. No depth limit was imposed on the offspring produced by crossover. The distribution of lengths in the last 50 generations was recorded. Figure 4 shows the average proportion of programs of each length in the last 50 generations of one run (middle line) along with the 2-standard-deviation wide confidence intervals. It is easy to see how close the average distribution of lengths is to the fixed point distribution obtained by iterating the schema equations (Figure 3, thick line).

In the schema-equation experiments we obtained different fixed point distributions only when we used initial conditions with a different average size, suggesting that the fixed point distribution for standard crossover is only a function of the average size, and independent of the actual shape of the initial distribution. This family of fix-point distributions characterises the search bias imposed by standard crossover when acting on linear structures: in the absence of other biases standard crossover will tend to more heavily sample the space of smaller-than-average programs and will be unable to focus its search on programs of a particular size. This means that if selection prefers longer-than-average programs or programs of a certain length, standard crossover may negatively bias the search. This does not happen with one-point crossover, provided that the initial population includes sufficient variety of shapes.

This effect becomes evident when the size-related fitness function in Equation 14 is used. With this fitness function programs of length 20 are maximally fit. However, due to the biases of standard crossover, the fixed point distribution obtained with different initial conditions (see Figures 5, 6 and 7) all share the same features: inability to focus on the maximally fit length and bias towards short programs.

For comparison, Figure 8 shows the behaviour of one-point crossover on the same function with uniform initial conditions. This corroborates the theoretical results in Section 3 which indicated that one-point crossover is a transparent (i.e. unbiased) operator as far as program lengths are concerned.

5 Fixed-point Size Distribution for Standard Crossover on a Flat Landscape

As noted in the previous section, the fixed-point distribution of lengths approached in the experiments with a flat landscape under standard crossover strongly resembles a gamma distribution. This observation prompted us to try to verify mathematically whether this is indeed the case.

A gamma distribution has the following form:

$$g(a, b, x) = \frac{x^{a-1} e^{-x/b}}{\Gamma(a) b^a} \tag{15}$$

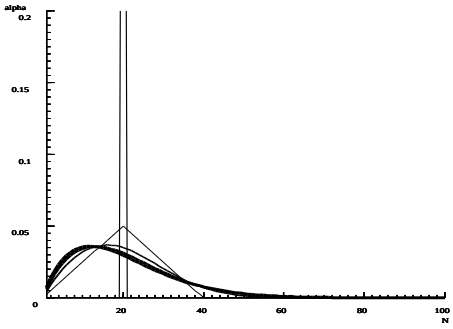

Fig. 1. Total transmission probability for standard crossover on a flat landscape with one-peak initial conditions. The first four generations are shown.

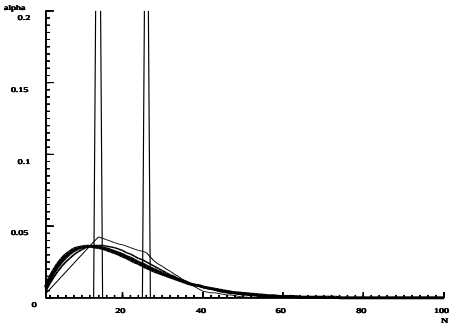

Fig. 2. Total transmission probability for standard crossover on a flat landscape with two-peak initial conditions. The first four generations are shown.

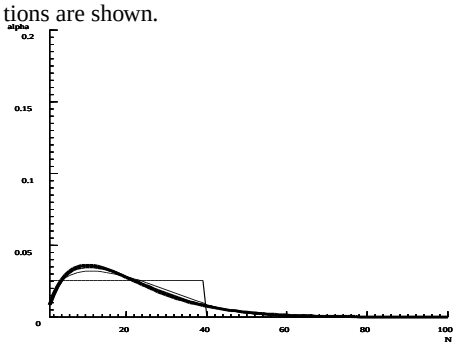

Fig. 3. Total transmission probability for standard crossover on a flat landscape with uniform initial conditions. The first four generations are shown.

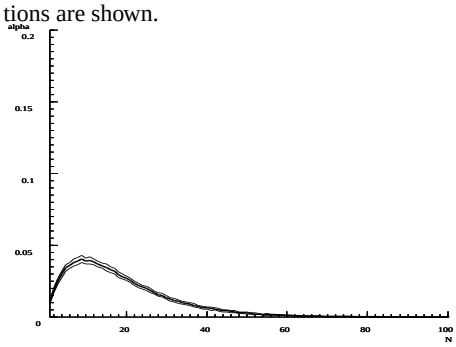

Fig. 4. Proportion of programs of each length in a real GP run for standard crossover on a flat landscape with uniform initial conditions.

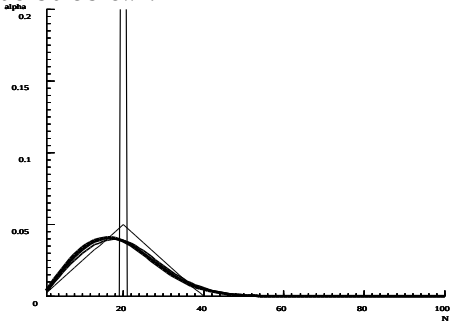

Fig. 5. Total transmission probability for standard crossover on a size-related landscape (Equation 14) with one-point initial conditions. The first four generations are shown.

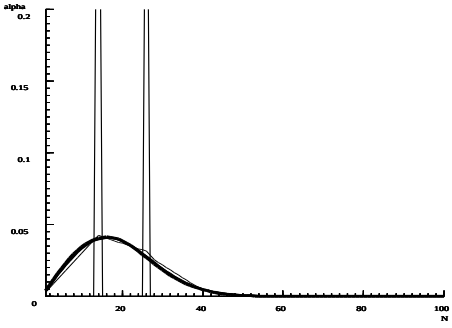

Fig. 6. Total transmission probability for standard crossover on a size-related landscape with two-point initial conditions. The first four generations are shown.

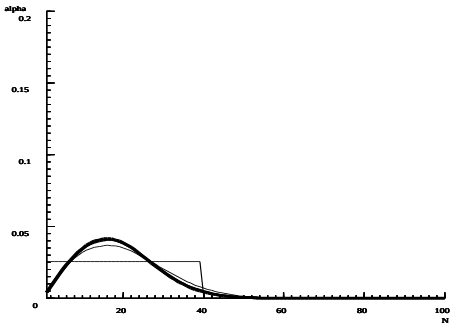

Fig. 7. Total transmission probability for standard crossover on a size-related landscape with uniform initial conditions. The first four generations are shown.

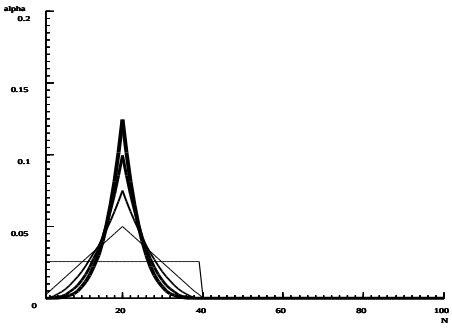

Fig. 8. Total transmission probability for one-point crossover on a size-related landscape with uniform initial conditions. The first four generations are shown.

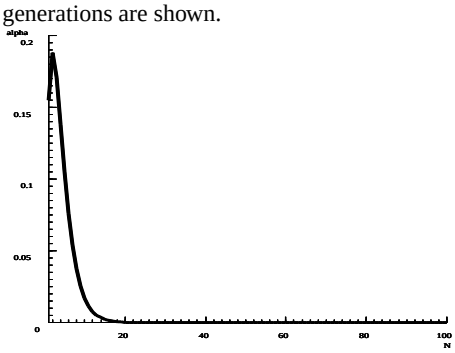

Fig. 9. Total transmission probability for standard crossover on a flat landscape with initial conditions $g_d(2, N)$.

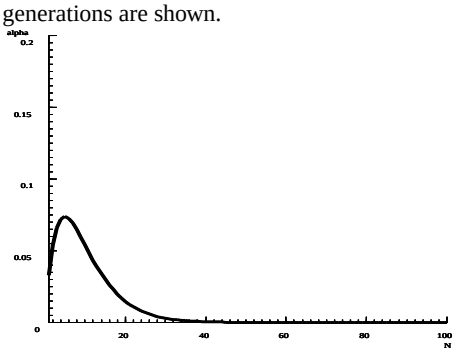

Fig. 10. Total transmission probability for standard crossover on a flat landscape with initial conditions $g_d(5, N)$.

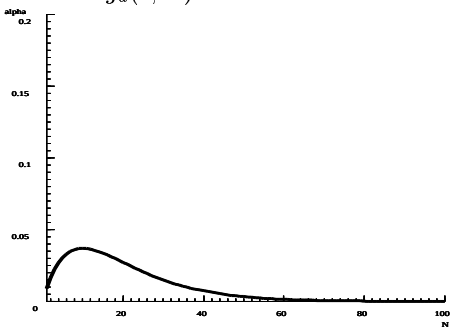

Fig. 11. Total transmission probability for standard crossover on a flat landscape with initial conditions $g_d(10, N)$.

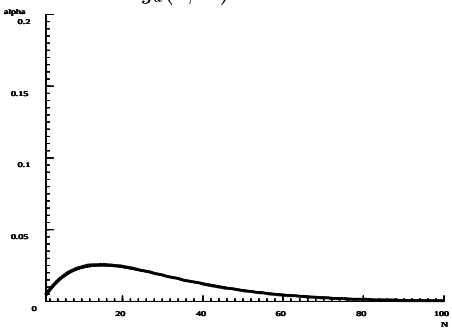

Fig. 12. Total transmission probability for standard crossover on a flat landscape with initial conditions $g_d(15, N)$.

where $\Gamma(a)$ is a generalisation of the factorial function[4] defined by

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

The gamma distribution has two parameters a and b that change its shape. For values of $a > 1$, $g(a, b, 0) = 0$. Also $\lim_{x\to\infty} g(a, b, x) = 0$ and the distribution has its maximum when $x = b(a-1)$. The mean of a gamma distribution is $\mu = ab$. So, the maximum of the distribution is shifted with respect to the mean by a distance b (i.e. the maximum is when $x = \mu - b$).

In order to verify if $g(a, b, x)$ is a fixed-point distribution for standard crossover on a flat landscape, firstly one has to specialise Equation 8 by setting $p(H, t) = \alpha(H, t-1)$. This produces

$$\alpha((=)^N, t) = (1 - p_{xo})\alpha((=)^N, t-1) + p_{xo} \sum_k \frac{\alpha((=)^k, t-1)}{k} \cdot \quad (16)$$
$$\sum_{n=\max(1,N-k+1)}^{\infty} \frac{\alpha((=)^n, t-1)}{n} \min(N, k, n, k+n-N).$$

Then one assumes that the system is in the hypothesised fixed point and substitutes $\alpha((=)^x, t-1) = g(a, b, x)$ in this equation. If $\alpha((=)^x, t-1) = g(a, b, x)$ is indeed a fixed-point for the system, then the r.h.s. resulting from this substitution must be equivalent to $g(a, b, N)$.

If one uses the general form for $\alpha((=)^x, t-1) = g(a, b, x)$, it is actually very difficult to check whether this is a fixed point for Equation 16. However, the experiments in the previous section suggested that a good value for the parameter a would be $a = 2$, which gives

$$g(2, b, x) = \frac{x e^{-x/b}}{b^2}.$$

This function is intuitively very appealing as a potential fixed point because it transforms terms of the form $\frac{\alpha((=)^x, t-1)}{x}$ in Equation 16 into exponentials of the form $\frac{e^{-x/b}}{b^2}$. Indeed, assuming $\alpha((=)^x, t-1) = g(2, b, x)$ allows one to simplify the r.h.s. of Equation 16 dramatically, obtaining:

$$\alpha((=)^N, t) = \left(\frac{N e^{-N/b}}{b^2}\right) \left(\frac{1}{b^2(1 - e^{-\frac{1}{b}})(e^{\frac{1}{b}} - 1)}\right). \quad (17)$$

So,

$$\alpha((=)^N, t) = \alpha((=)^N, t-1) \left(\frac{1}{b^2(1 - e^{-\frac{1}{b}})(e^{\frac{1}{b}} - 1)}\right). \quad (18)$$

[4] For integer $a > 0$, $\Gamma(a) = (a-1)!$.

Since, in general, $b^2(1-e^{-\frac{1}{b}})(e^{\frac{1}{b}}-1) \neq 1$, we can conclude from this result that the assumed gamma distribution is not a fixed point for the system.[5] This had to be expected because the gamma distribution is a continuous probability distribution. While $\int_{x\geq 0} g(a,b,x)dx = 1$, it is not true that $\sum_{x\geq 0} g(a,b,x) = 1$. So, sampling $g(a,b,x)$ does not produce a discrete probability distribution. This is why $\alpha((=)^k) = g(2,a,k)$ cannot be a fixed point. However, the distribution obtained by sampling $g(2,b,x)$ can be transformed into a probability distribution by simply normalising each sample, obtaining:

$$g_d(b,x) = xe^{-x/b}(1-e^{-\frac{1}{b}})(e^{\frac{1}{b}}-1). \tag{19}$$

Naturally, $g_d(b,0) = 0$ and $\lim_{x\to\infty} g_d(b,x) = 0$ and the distribution has a maximum for $x = b$, like $g(2,b,x)$. The mean of the discrete gamma distribution is $\mu = (1+e^{-\frac{1}{b}})/(1-e^{-\frac{1}{b}})$, which is only slightly different from (and quickly converges to) $2b$. So, the maximum of the distribution is shifted with respect to the mean by a distance $x = \mu - b \approx b$.

This discrete gamma distribution can be shown mathematically to be a fixed point. So, if

$$\alpha((=)^N, t-1) = Ne^{-N/b}(1-e^{-\frac{1}{b}})(e^{\frac{1}{b}}-1), \tag{20}$$

then $\alpha((=)^N, t+T) = \alpha((=)^N, t-1)$ for any positive value of b and T. This can be reformulated in terms of the mean μ of $g_d(b,N)$ by setting $b = -1/\ln((\mu-1)/(\mu+1))$ in $g_d(b,N)$, obtaining

$$\alpha((=)^N, t-1) = Ne^{N\ln(\frac{\mu-1}{\mu+1})}\left(\frac{\mu+1}{\mu-1}-1\right)\left(1-\frac{\mu-1}{\mu+1}\right), \tag{21}$$

which can be written more simply

$$\alpha((=)^N, t-1) = Nr^{N-1}(r-1)^2, \tag{22}$$

where $r = (\mu-1)/(\mu+1)$. (For an alternative derivation of this and related results see [14].)

This result was also corroborated numerically by initialising the system at $\alpha((=)^N, -1) = g_d(b,N)$ for different values of b and iterating the schema equations for a few generations, as shown in Figures 9–12 where the plots for the first four generations coincide perfectly.

As noted before, on average standard crossover inserts and removes the same amount of genetic material. Therefore, on a flat landscape and with an infinite population no change in mean length can occur. So, given any initial distribution of lengths $\alpha((=)^N, -1)$, if one assumes that there are no fixed points other than the family of discrete gamma functions $g_d(b,x)$, then, by setting $\mu = \sum_N N\alpha((=)^N, -1)$ (the mean

[5] The relative difference between $\alpha((=)^N, t)$ and $\alpha((=)^N, t-1)$, $\Delta(b) = \frac{b^2(1-e^{-\frac{1}{b}})(e^{\frac{1}{b}}-1)-1}{b^2(1-e^{-\frac{1}{b}})(e^{\frac{1}{b}}-1)}$, is always very small and decreases very quickly as b increases. For example, $\Delta(1) \approx 0.08$, $\Delta(2) \approx 0.02$, $\Delta(5) \approx 0.003$, $\Delta(10) \approx 0.0008$, $\Delta(20) \approx 0.0002$ and $\Delta(100) \approx 8\times 10^{-6}$. So, for most practical purposes a gamma distribution can be considered to be a fixed point for the evolution of size under standard crossover on a flat landscape.

length of the programs in the initial generation), Equation 21 gives the fixed point to which GP will converge.

At this stage we are unable to prove that there are no other fixed points in the system. So, we cannot guarantee that the length distribution in GP with standard crossover on a flat landscape will always converge towards a discrete gamma distribution or converge at all. The experiments described in the previous section always seemed to do so, which suggests that other fixed points might be unlikely. Also, even if there are other fixed point distributions, it seems likely that they will share important characteristics with $g_d(b,x)$. This is because GP with standard crossover will always be able to produce programs which are much longer than average. So, the only way in which the average length can remain constant is to have more shorter-than-average programs than longer-than-average ones.

6 Search Space Sampling under Standard Crossover

It is important to understand the consequences of the bias described in the previous section. Let us imagine that our linear GP system operating on a flat landscape is at the fixed point $g_d(b,x)$ for some value of b. Since, there are $n(x) = |\mathcal{F}|^{x-1}|\mathcal{T}|$ different programs of length x in the search space, it is possible to compute the average probability $p_{\text{sample}}(x)$ that each of these will be sampled by standard crossover, namely

$$p_{\text{sample}}(x) = g_d(b,x)/n(x). \tag{23}$$

It is easy to study this function and to conclude that, for a flat landscape, standard GP will sample a particular short program much more often than it will sample a particular long one. Figures 13 and 14 show the average sampling probability for programs of a given length for standard crossover on a flat landscape at the fixed points $g_d(b,x)$ for $b = 2, 5, 10, 20, 50, 100, 1000, 100000, 1000000$ assuming $|\mathcal{F}| = |\mathcal{T}| = 2$. In general, an increase in length of one order of magnitude corresponds to a drop in sampling probability of many orders of magnitude and this trend is largely independent of the particular value of b. For example, when $b = 10$ (i.e. when the average program length is about 20), on average (and approximately) GP will resample the same program of length 5 every 1054 crossovers, while it will resample the same program of length 50 every 3.33×10^{17} crossovers with a difference of 13 orders of magnitude! This difference does not change significantly even for much larger values of b. Indeed $\lim_{b\to\infty} p_{\text{sample}}(5)/p_{\text{sample}}(50) = \lim_{b\to\infty} \frac{2^{45}}{10} e^{\frac{45}{b}} = \frac{2^{45}}{10} \approx 3.5 \times 10^{12}$ which is still a huge number. An alternative way of looking at this is to calculate the ratio $\frac{p_{\text{sample}}(x)}{p_{\text{sample}}(x+1)} = \frac{2x}{x+1} e^{\frac{1}{b}} > \frac{2x}{x+1}$, which for large values of x is approximately 2.

Preliminary work suggests that similar results hold for GP with standard crossover operating on trees. If these results are confirmed the widely reported tendency to bloat is particularly remarkable.

7 Conclusions

In this paper, an exact schema theorem for genetic programming operating on linear structures using standard crossover has been provided. Using this theorem and the cor-

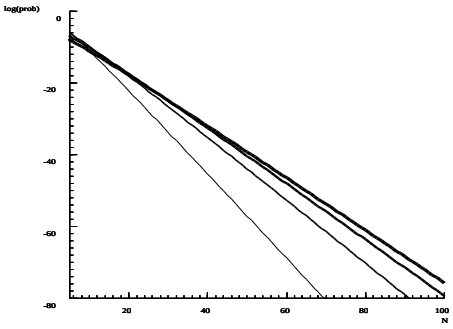

Fig. 13. Plot of the average sampling probability $p_{\text{sample}}(N)$ vs. program length N for standard crossover on a flat landscape at the fixed points $g_d(b, N)$ for $b = 2, 5, 10, 20$ for $|\mathcal{F}| = |\mathcal{T}| = 2$. Thicker plots represent higher values of b. Note the use of a logarithmic scale.

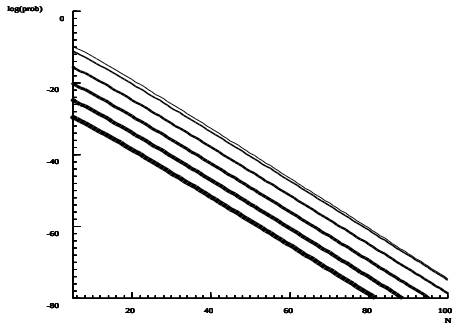

Fig. 14. Plot of the average sampling probability $p_{\text{sample}}(N)$ vs. program length N for standard crossover on a flat landscape at the fixed points $g_d(b, N)$ for $b = 50, 100, 1000, 100000, 1000000$ for $|\mathcal{F}| = |\mathcal{T}| = 2$. Thicker plots represent higher values of b. Note the use of a logarithmic scale.

responding version for one-point crossover, which is also reported, we have been able to study the evolution of size and the biases introduced by the operators both mathematically and in simulations.

This research establishes a link between important areas of theoretical research in GP: the study of the evolution of shape and size [1], the biases imposed by different operators [10] and the theory of schemata [9, 7, 6]. In recent research we have started using these results to study bloat in linear representations [3]. In the future we hope to be able to use this theory to model mathematically and better understand the operator bias as well as the reasons for bloat, intron proliferation and code compression [2, 4, 5] in non-linear structures.

Acknowledgements

The authors would like to thank Bill Langdon, Jonathan Rowe and other members of the EEBIC (Evolutionary and Emergent Behaviour Intelligence and Computation) group at Birmingham for helpful comments and discussion.

The second author would like to extend special thanks to The University of Birmingham School of Computer Science for graciously hosting him during his sabbatical, and various offices and individuals at the University of Minnesota, Morris, for making that sabbatical possible.

References

[1] W. B. Langdon, T. Soule, R. Poli, and J. A. Foster. The evolution of size and shape. In L. Spector, W. B. Langdon, U.-M. O'Reilly, and P. J. Angeline, editors, *Advances in Genetic*

Programming 3, chapter 8, pages 163–190. MIT Press, Cambridge, MA, USA, June 1999.

[2] N. F. McPhee and J. D. Miller. Accurate replication in genetic programming. In L. Eshelman, editor, *Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95)*, pages 303–309, Pittsburgh, PA, USA, 15-19 July 1995. Morgan Kaufmann.

[3] N. F. McPhee and R. Poli. A schema theory analysis of the evolution of size in genetic programming with linear representations. In *Genetic Programming, Proceedings of EuroGP 2001*, LNCS, Milan, 18-20 Apr. 2001. Springer-Verlag.

[4] P. Nordin and W. Banzhaf. Complexity compression and evolution. In L. Eshelman, editor, *Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95)*, pages 310–317, Pittsburgh, PA, USA, 15-19 July 1995. Morgan Kaufmann.

[5] P. Nordin, F. Francone, and W. Banzhaf. Explicitly defined introns and destructive crossover in genetic programming. In J. P. Rosca, editor, *Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, pages 6–22, Tahoe City, California, USA, 9 July 1995.

[6] R. Poli. Exact schema theorem and effective fitness for GP with one-point crossover. In D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.-G. Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 469–476, Las Vegas, July 2000. Morgan Kaufmann.

[7] R. Poli. Hyperschema theory for GP with one-point crossover, building blocks, and some new results in GA theory. In R. Poli, W. Banzhaf, and *et al.*, editors, *Genetic Programming, Proceedings of EuroGP 2000*. Springer-Verlag, 15-16 Apr. 2000.

[8] R. Poli. General schema theory for genetic programming with subtree-swapping crossover. In *Genetic Programming, Proceedings of EuroGP 2001*, LNCS, Milan, 18-20 Apr. 2001. Springer-Verlag.

[9] R. Poli and W. B. Langdon. A new schema theory for genetic programming with one-point crossover and point mutation. In J. R. Koza, K. Deb, M. Dorigo, D. B. Fogel, M. Garzon, H. Iba, and R. L. Riolo, editors, *Genetic Programming 1997: Proceedings of the Second Annual Conference*, pages 278–285, Stanford University, CA, USA, 13-16 July 1997. Morgan Kaufmann.

[10] R. Poli and W. B. Langdon. On the search properties of different crossover operators in genetic programming. In J. R. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba, and R. Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 293–301, University of Wisconsin, Madison, Wisconsin, USA, 22-25 July 1998. Morgan Kaufmann.

[11] R. Poli and W. B. Langdon. Schema theory for genetic programming with one-point crossover and point mutation. *Evolutionary Computation*, 6(3):231–252, 1998.

[12] R. Poli, W. B. Langdon, and U.-M. O'Reilly. Analysis of schema variance and short term extinction likelihoods. In J. R. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba, and R. Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 284–292, University of Wisconsin, Madison, Wisconsin, USA, 22-25 July 1998. Morgan Kaufmann.

[13] J. P. Rosca. Analysis of complexity drift in genetic programming. In J. R. Koza, K. Deb, M. Dorigo, D. B. Fogel, M. Garzon, H. Iba, and R. L. Riolo, editors, *Genetic Programming 1997: Proceedings of the Second Annual Conference*, pages 286–294, Stanford University, CA, USA, 13-16 July 1997. Morgan Kaufmann.

[14] J. E. Rowe and N. F. McPhee. The effects of crossover and mutation operators on variable length linear structures. Technical Report CSRP-01-7, University of Birmingham, School of Computer Science, January 2001.