Modelling Crossover-Induced Linkage in Genetic Algorithms

Adam Prügel-Bennett
ECS, University of Southampton, SO17 1BJ, United Kingdom

4th January 2001

Abstract

The dynamics of a genetic algorithm undergoing ranking selection, mutation, and two-point crossover for the ones-counting problem is studied using a statistical mechanics approach. This approach has been used previously to study this problem, but with uniform crossover. Two-point crossover induces additional linkage between nearby loci which changes the dynamics significantly. To account for this linkage, the evolution of the auto-correlation function is incorporated into a model of the dynamics. This complicates the analysis and requires several additional approximations to be made. Nevertheless, the model we derive is shown to capture the main features of the dynamics and is in good agreement with simulations.

1 Introduction

Recombination or crossover can be a very significant factor in determining the performance of a genetic algorithm (GA) or other evolutionary algorithm. By allowing the genetic strings (chromosomes) to recombine, the GA can perform a more global exploration of the solution space than it would using small mutations alone. Combining solutions may, however, also break up 'building blocks' and produce a deterioration of the solution. If this disruption is too large, the use of crossover can be harmful. Frequently, the critical part of designing an effective GA is to find an efficient crossover operator that produces good mixing with little disruption. It is, therefore, important to understand in detail what crossover does.

In this paper we examine the evolution of a GA that incorporates two-point crossover and compare this to a GA using uniform crossover. Two-point crossover causes a spatial correlation along the genetic strings—so called linkage[1]—which is not present in uniform crossover. Previous studies have shown that multipoint crossover significantly slows down the asymptotic rate of mixing in comparison to uniform crossover [1, 2]. Although this analysis gives some hint of the difference between multipoint and uniform crossover, it

[1]The term linkage is used in two ways. In this context it refers to the relation between alleles on nearby sites. However, we will also use the term linkage disequilibria as a measures of how far away the alleles are from a totally mixed state. Linkage disequilibria is produced by selection and reduced by crossover

leaves unanswered the question of how these different crossovers affect the dynamics of a GA. Here, we focus on how the crossover operator interacts with other operators, specifically with selection and mutation. For this purpose, we consider a GA acting on the ones-counting problem with ranking selection, mutation, and crossover. Figure 1 shows the evolution of a GA, using two-point and uniform crossover, for the ones-counting problem with $L = 64$ sites. Fitness is measured so that random strings have, on average, a fitness of zero, while the fittest possible string has a fitness of L. Using uniform crossover, the GA finds good solutions slightly faster than with two-point crossover, although GAs, using either crossover, eventually reach equilibrium states with a virtually identical average fitness. The small difference between these two types of crossover is peculiar to this problem. Nevertheless, we would like to understand, in detail, the cause of this discrepancy.

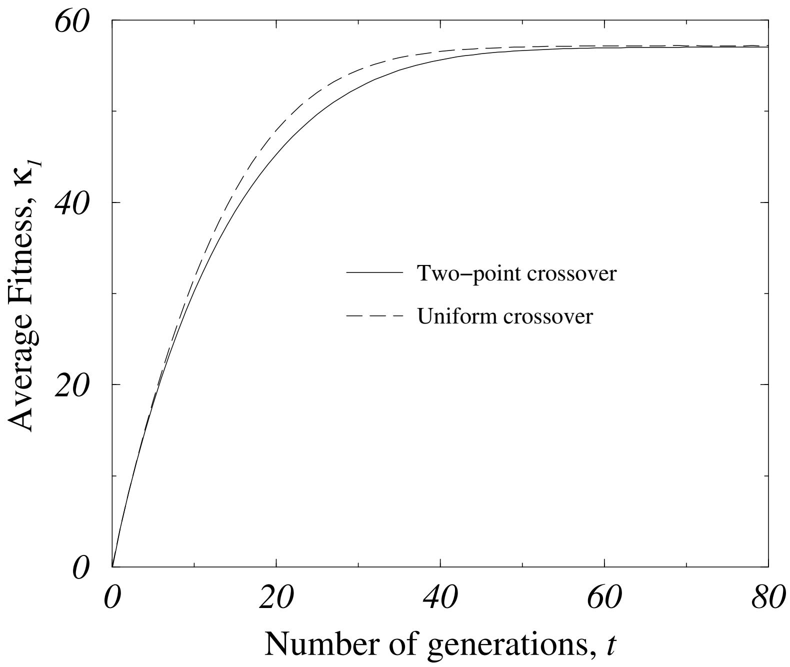

Figure 1: The evolution of the average fitness for a GA acting on the ones-counting problem, with a string length of 64, a population size of 100, and a mutation rate of 1/64. The dashed line shows the behaviour with uniform crossover and the solid line with two-point crossover. The curves have been averaged of 10 000 runs. Errors in the mean are less than the line width.

Historically, single-point and two-point crossover were preferred over uniform crossover as they were believed to be less disruptive [3, 4]. However, empirical studies showed that uniform crossover often outperformed single-point and two-point crossover on many problems [5, 6]. As a result, uniform crossover has become much more frequently. It is now recognized that the crossover operator should be tailored to the problem (see, for example, [7, 8]). However, in choosing a crossover operator, the same issue arises, that is, to balance the degree of disruption against the amount of mixing produced by crossover. In this paper we look in detail at the mixing issue. For ones-counting, crossover causes no disruption on average. This is rather unusual, but it allows us to concentrate just on the mixing effect of crossover.

To perform this study, we will model the dynamics of a GA using a statistical mechanics approach. This has been carried out for a large number of different models [9–23], however, in none of these studies was the spatial correlations caused by a multipoint crossover considered. Including these spatial correlations has proved to be a significant complication (even though its effects, as shown in Figure 1, are surprisingly minor). To model the spatial correlations, we introduce an auto-correlation function measuring the average correlation between pairs of string variables (alleles) at different sites (loci). We then follow the evolution of this auto-correlation function together with other macroscopic variables as we apply selection, mutation, and two-point crossover. The effect of selection is particularly intricate and forms the bulk of the paper. We end up with an approximate model of the dynamics in terms of macroscopic properties of the system. By comparing this with the dynamics under uniform crossover, we can see how crossover-induced linkage slows down the search. In brief, crossover aids search by transforming genetic variation into an increased fitness variance. This is necessary as selection exploits the variance in the fitness to increase the average fitness in the population, but in so doing reduces the fitness variance. However, selection reduces the genetic variation much less than the fitness variance (since, two strings with the same fitness can have a very different genetic makeup). The spatial correlations built up by two-point crossover reduces the genetic variation and so reduces the effectiveness of crossover. The detailed analysis of this process is the subject of this paper.

The rest of the paper is organized as follows. In the next section, we describe the problem, introduce notation, and give an overview of the modelling process we use. In Section 3, we tackle the effect of the operators on the macroscopic variables used to model the system. Comparison between the model we have derived and simulations are given in Section 4. Finally, in Section 5 we discuss our results. Some of the more technical sections have been relegated to the appendices.

2 Problem and Modelling Technique

Problem and GA

The ones-counting problem can be formulated as follows. We consider solutions to be strings consisting of L variables. We denote the strings by a vector $\mathbf{S} = (S_1, S_2, \ldots, S_L)$, where the variables S_i are either plus or minus one. The fitness of the string is given by

$$F(\mathbf{S}) = \sum_{i=1}^{L} S_i. \tag{1}$$

The fittest string therefore consists of a string of all ones and has a fitness L. Random strings with half the alleles being $+1$ and the other half -1 will have a fitness of zero.

We use a classic (generational) GA to find good solution to the ones-counting problem. We start from a random population of P strings, $\mathcal{P}(0) = \{\mathbf{S}^\alpha(0)\}_{\alpha=1}^{P}$, where $\mathbf{S}^\alpha(t)$ denotes the member α at time t. The populations is evolved by applying selection, mutation, and crossover

$$\mathcal{P}(t) \xrightarrow{sel} \mathcal{P}^s(t) \xrightarrow{mut} \mathcal{P}^{sm}(t) \xrightarrow{cross} \mathcal{P}^{smc}(t) = \mathcal{P}(t+1). \tag{2}$$

We use ranking selection in which we assign a rank of 0 to the least-fit member of the population, a rank 1 to the next least-fit member, and so on. The fittest member therefore has a rank $P-1$. Roulette wheel selection is used to choose a new population, $\mathcal{P}^s(t)$, with each member being chosen in proportion to its rank. The population size is kept fixed at each stage of the evolution. The individuals in population, $\mathcal{P}^s(t)$, are then mutated by changing the variables (i.e., $S_i^\alpha \to -S_i^\alpha$) with a probability u.

Crossover is performed by first randomly pairing each member of the population $\mathcal{P}^{sm}(t)$. From each pair of parents we produce two offspring. To remove complications caused by the ends of the string we assume periodic boundary conditions. That is, we think of the string as arranged in a ring. Two-point crossover is then performed by selecting a locus at random and taking half the string starting at that locus from the first parent and the other half of the string from the second parent. This is sketched schematically in Figure 2.

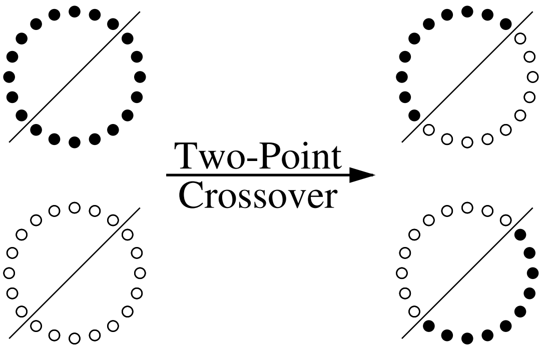

Figure 2: Illustration of two point crossover.

Macroscopic Variables

We use statistical properties of the population to model the evolution of the population. These statistical properties, or macroscopic variables, describe the distribution of fitness, the correlation between strings, and measure the linkage caused by crossover.

Cumulants are used to describe the distribution of fitnesses. The more cumulants we use, the more accurate the model, but also the more complicated it becomes. We obtain a reasonably accurate model using just the first two cumulants—that is, the average fitness, K_1, and variance in the distribution of fitnesses, K_2,

$$K_1 = \langle F_\alpha \rangle_{\mathcal{P}} \tag{3}$$

$$K_2 = \left\langle F_\alpha^2 \right\rangle_{\mathcal{P}} - \langle F_\alpha F_\beta \rangle_{\mathcal{P}}, \tag{4}$$

where the average is over all members of the population. Note that we use the convention that the indices are always kept distinct when averaging over the population, for example,

$$\langle F_\alpha F_\beta \rangle_{\mathcal{P}} = \frac{1}{P(P-1)} \sum_{\alpha \neq \beta} F_\alpha F_\beta.$$

This gives the unbiased estimates of the cumulants for an infinite population. This simplifies the analysis, but does not mean that we are making an infinite population approximation. We include all finite population corrections; we just choose to use unbiased estimates of the cumulants as modelling variables. The average 'genetic' correlation, Q, is defined as

$$Q = \langle Q_{\alpha\beta} \rangle_{\mathcal{P}}\,, \tag{5}$$

where $Q_{\alpha\beta}$ is the pair correlation

$$Q_{\alpha\beta} = \sum_{i=1}^{L} S_i^{\alpha} S_i^{\beta}. \tag{6}$$

Cumulants and correlations have been used to model the dynamics of a GA (see, for example, [9–13, 19, 22]).

To describe the linkage produced by two-point crossover we have to introduce functions measuring the auto-correlation in a single string and cross-correlation between pairs of strings. The auto-correlation is defined as

$$U(k) = \langle U_{\alpha}(k) \rangle_{\mathcal{P}}\,, \qquad\qquad U_{\alpha}(k) = \sum_{i=1}^{L} S_i^{\alpha} S_{i+k}^{\alpha}, \tag{7}$$

where we use the convention that the site subscript satisfies circular boundary conditions (i.e., $S_{i+L} = S_i$). The cross-correlation is defined as

$$V(k) = \langle V_{\alpha\beta}(k) \rangle_{\mathcal{P}}\,, \qquad\qquad V_{\alpha\beta}(k) = \sum_{i=1}^{L} S_i^{\alpha} S_{i+k}^{\beta}. \tag{8}$$

The correlations $U(k)$, $V(k)$ and Q are not independent, for example, $U_{\alpha}(k) = V_{\alpha,\alpha}(k)$, $V(0) = Q$, and $U(0) = L$. In addition the correlation functions have some important symmetries: $U(k)$ and $V(k)$ are periodic (i.e., $U(k) = U(L+k)$ and $V(k) = V(L+k)$) and even (i.e., $U(k) = U(-k)$ and $V(k) = V(-k)$). To measure the degree of linkage we consider the difference between the auto- and cross-correlation functions

$$W(k) = U(k) - V(k) = \left\langle \sum_{i=1}^{L} S_i^{\alpha} (S_{i+k}^{\alpha} - S_{i+k}^{\beta}) \right\rangle_{\mathcal{P}}. \tag{9}$$

The novelty of this paper is in calculating how $W(k)$ evolves under selection, mutation, and crossover. The effects of mutation and crossover can be computed exactly, but with selection we are forced to make approximations. This is often the case in modelling evolution because computing selection requires more information that we can practically keep. We are therefore forced to use some inference procedure to supply this information. Here, we use a maximum entropy ansatz. Maximum entropy has been used to model the dynamics of a GA [10, 24], but in these cases it was used to infer different information than used here.

The linkage function $W(k)$ plays an important role because it is intimately related to the second cumulant and the correlation

$$K_2 \;=\; \left\langle F_{\alpha}^2 \right\rangle_{\mathcal{P}} - \langle F_{\alpha} F_{\beta} \rangle_{\mathcal{P}}$$

$$
\begin{aligned}
&= \left\langle \left(\sum_i S_i^\alpha \right)^2 \right\rangle_{\mathcal{P}} - \left\langle \left(\sum_i S_i^\alpha \right) \left(\sum_i S_i^\beta \right) \right\rangle_{\mathcal{P}} \\
&= L - Q + \sum_{k=1}^{L-1} W(k). \qquad (10)
\end{aligned}
$$

We use this identity to compute the effect of crossover on the second cumulant. To describe the linkage function exactly requires a lot of information (in fact, $\lceil L/2 \rceil$ numbers). This is not in the spirit of modelling where we try to find a description that uses a number of parameters that does not grow with the size of the system. We therefore expand $W(k)$ in terms of a set of orthogonal polynomials, $\psi_n(k)$,

$$
W(k) = \sum_{n=0} w_n \psi_n(k), \qquad (11)
$$

where $\psi_n(k)$ are n-th order polynomials in

$$
[L - 2k]_L \equiv (L - 2k) \bmod L
$$

satisfying

$$
\sum_{k=1}^{L-1} \psi_n(k) \psi_m(k) = \delta_{m,n}.
$$

The first component,

$$
\psi_0(k) = \frac{1}{\sqrt{L-1}},
$$

measures the average linkage, while the higher components measure the spatial variation in $W(k)$ (see Appendix A). This expansion respects the periodic and even symmetries of $W(k)$. We need only keep the first few components of $W(k)$ in order to obtain an accurate model of the dynamics. The expansion has the property that the term on the right-hand side of (10) depends only on w_0 since

$$
\sum_{k=1}^{L-1} W(k) = \sqrt{L-1}\, w_0. \qquad (12)
$$

We shall treat the constant term w_0 and the higher terms w_n ($n > 0$) differently.

The auto- and cross-correlation function are related directly to the fitness, since

$$
\begin{aligned}
F_\alpha^2 &= L + \sum_{k=1}^{L-1} U_\alpha(k) = L + \sqrt{L-1}\, u_0^\alpha \\
F_\alpha F_\beta &= Q_{\alpha\beta} + \sum_{k=1}^{L-1} V_{\alpha\beta}(k) = L + \sqrt{L-1}\, v_0^{\alpha\beta},
\end{aligned}
$$

where

$$
\begin{aligned}
U_\alpha(k) &= \sum_{n=1} u_n^\alpha \, \psi_n(k) \\
V_{\alpha\beta}(k) &= \sum_{n=1} v_n^{\alpha\beta} \, \psi_n(k).
\end{aligned}
$$

Only the zeroth component of the correlation functions are correlated explicitly with the fitness, although there is a subtle relationship between the fitness and higher components. This arises because the fit individuals have few bad alleles ($S_i = -1$) and so tend to be more uniform with smaller high order components of $U_\alpha(k)$ than less fit individuals.

Modelling Technique

In practice, GAs exhibit fluctuations between different runs. We shall model the average behaviour for an ensemble of GAs, thus averaging out these fluctuations. (We can think of the ensemble as a set of independent runs of the GA using different seeds for the random number generator.) This means that, rather than using the raw macroscopic variables from a single run, we use variables describing the ensemble. In particular, we use the ensemble averages which we denote by a bar. Thus, $\bar{Q}$ is the ensemble average of Q, and so on. Using only ensemble averages is an over simplification, since the evolution of some quantities depend nonlinearly on the macroscopic variables. In such cases the average behaviour will depend on the full distribution of the macroscopic variables. This complicated the analysis considerably. In this paper, we will ignore corrections due to the ensemble distribution. Corrections due to fluctuations have been calculated previously by including variables describing the covariance of the macroscopic variables in the ensemble [13]. This gave very accurate results, but at the cost of complicating the analysis further. However, the corrections due to fluctuations are typically small, particularly where there is crossover which has the effect of reducing the importance of nonlinear terms [25]. Nevertheless, ensemble fluctuations probably set the limit of accuracy that can be obtained in any reasonable calculation using this approach.

We choose to keep only a small number of ensemble variables,

$$\Xi = \{\bar{K}_1, \bar{K}_2, \bar{Q}, \bar{w}_0, \bar{w}_1, \bar{w}_2, \ldots\}.$$

To calculate the dynamics we are forced to infer all other quantities from this small number of ensemble variables. Our task is to calculate the dynamics of the ensemble variables under selection, mutation, and crossover

$$\Xi(t) \xrightarrow{sel} \Xi^s(t) \xrightarrow{mut} \Xi^{sm}(t) \xrightarrow{cross} \Xi^{smc}(t) = \Xi(t+1). \tag{13}$$

Having found how each operator changes our set of ensemble variables, we can then calculate the full dynamics of the ensemble by iterating (13) starting from an initial set of ensemble variables $\Xi(0)$. For a random initial population, all the ensemble variables are zero except $\bar{K}_2$ which is equal to L. The recursion (13) forms a model or caricature of the true dynamics defined in (2). As it involves only a few variables it can provide an understand of the dynamics that is obscured in the real system.

3 Calculating the Dynamics

We have now arrive at the main task of this paper, namely to calculate how the ensemble variables are changed under the operations of the GA. The novel part of this calculation is to compute the change of the linkage function, $\bar{W}(k)$. By far the hardest task is to calculate the effect of selection on $\bar{W}(k)$. This calculation is necessarily approximate because it requires more information than we have available. We resort to a maximum entropy argument to infer this missing information. Simple estimates of the effect of selection, mutation, and crossover are presented in that order. The section is concluded with a discussion on computing corrections to our first approximations.

3.1 Selection

The effect of selection on the cumulants has been computed in other papers. We therefore give only the results. We then turn our attention to the correlations Q, $V(k)$, $U(k)$, and $W(k)$. The result for Q has been obtained previously [22], but we give a new derivation which we believe to be superior to our previous account.

Cumulants under Selection

The effect of ranking selection on the cumulants describing the distribution of fitness is relatively straightforward to calculate. For an infinite population it was calculated in [26]. A useful approximation that includes finite population corrections is given in [22]. The exact results for a finite population have been computed in [23]. We will use the approximation of [22] since the results are sufficiently accurate for our purposes and more easily extended to variations of the selection scheme such as using stochastic universal sampling [27]. Starting from a Gaussian with mean $\bar{K}_1$ and variance $\bar{K}_2$, the effect of selection on the first two cumulants is

$$\text{Sel: } \bar{K}_1 \to \bar{K}_1^s = \bar{K}_1 + \sqrt{\frac{\bar{K}_2}{\pi}} \tag{14}$$

$$\text{Sel: } \bar{K}_2 \to \bar{K}_2^s = r_P \left(1 - \frac{1}{\pi}\right) \bar{K}_2, \tag{15}$$

where r_P is the finite population correction given by

$$r_P = \frac{P - \langle n^2 \rangle}{P - 1}$$

and

$$\left\langle n^2 \right\rangle = \frac{1}{P} \sum_{\mu=1}^{P} n_\mu^2, \tag{16}$$

n_μ is the number of times we select the individual μ. For roulette wheel sampling and ranking selection

$$r_P = \frac{P - \langle n^2 \rangle}{P - 1} = \frac{3P^2 - 7P + 2}{3P(P-1)}. \tag{17}$$

Correlations under Selection

We need to calculate the effect of selection on the two string correlation, $\bar{Q}$, and the coefficients describing the linkage, $\bar{w}_0$, $\bar{w}_1$, etc. To do this we consider the effect of selection on the auto-correlation function $\bar{U}(k)$ and the cross-correlation function $\bar{V}(k)$. We can treat $\bar{Q}$ together with the cross-correlation since $Q = V(0)$. The constant component of linkage, $\bar{w}_0$, has a special status because of the identities (10) and (12)

$$\bar{w}_0 = \frac{\bar{\kappa}_2 + \bar{Q} - L}{\sqrt{L-1}}.$$

Thus, we can compute $\bar{w}_0^s$ from a knowledge of $\bar{\kappa}_2^s$ and $\bar{Q}^s$. Our task then is to compute the change in $\bar{Q}$ and in the higher components of $W(k)$. The calculation of the change in the auto-correlation and cross-correlation are quite distinct and will be treated separately. These are the most involved calculations in the paper.

Auto-correlation

The auto-correlation after selection is given by

$$U^s(k) = \frac{1}{P}\sum_{\alpha=1}^{P} U^s_\alpha(k) = \frac{1}{P}\sum_{\mu=1}^{P} n_\mu U_\mu(k),$$

where we have used the fact that members of the new population $\mathcal{P}^s = \{\mathbf{S}^\alpha\}_{\alpha=1}^P$ are all taken from the old population $\mathcal{P} = \{\mathbf{S}^\mu\}_{\mu=1}^P$, and n_μ is the number of times member μ is selected. The components of the auto-correlation after selection are, therefore, given by

$$u^s_n = \frac{1}{P}\sum_{\mu=1}^{P} n_\mu u^\mu_n, \tag{18}$$

where $u^\mu_n = \sum_k \psi_n(k) U_\mu(k)$. Unfortunately we cannot compute the right-hand side of (18) directly as we do not know how the components, u^μ_n, are related to the fitness and hence to n_μ. This is a common problem in a statistical mechanics approach—there is insufficient information to calculate the behaviour of the system. To make progress, an inference mechanism is required to recover the missing information. The most commonly used inference mechanism is maximum entropy and we will use it here to infer the distribution of u^μ_n's. A maximum entropy approach has been used previously in modelling populations to infer unknown quantities [10, 24], although it was slightly different information that was being estimated.

The approach we take is to estimate the distribution of the quantities u^μ_n and calculate how this distribution changes under selection. The first task is to compute the distribution of u^μ_n for a given fitness, F

$$P_0(u^\mu_n|F) \propto \operatorname*{Tr}_{\{\mathbf{S}\}} \delta\left(\sum_{k=1}^{P-1} \psi_n(k) \sum_{i=1}^{L} S_i\, S_{i+k} - u^\mu_n\right)\, \delta\left(\sum_{i=1}^{L} S_i - F\right),$$

where the trace is a shorthand for a sum over all possible strings and the delta's denote Dirac delta functions. This is not the distribution that occurs in a GA with two-point crossover, but rather the distribution of u^α_n obtained by summing over all strings with fitness F. The exact form of the distribution is hard to find, we can, however, calculate the first few moments. The calculation is given in Appendix B. The first moment is zero, the second moment is given by $\Delta^2(F) = 2L(1-(F/L)^2)^2$. We will use a Gaussian approximation, although the distribution is very far Gaussian—it is heavily negatively skewed. Despite this, the Gaussian assumption gives a good approximation, or, at least, including the third moment is found to give only a very small correction. The distribution of u^μ_n given a distribution $\rho(F)$ is thus

$$P_0(u^\mu_n) = \int_{-\infty}^{\infty} P_0(u^\mu_n|F)\, \rho(F)\, \mathrm{d}F \approx \frac{1}{\sqrt{2\pi}\,\Delta(K_1)} \mathrm{e}^{-(u^\mu_n)^2/(2\Delta^2(K_1))}$$

which we obtain by expanding $P_0(u^\mu_n|F)$ around $F = K_1$.

To infer the actual distribution $P(u^\mu_n)$ given our knowledge of the average of $\bar{u}_n$, we maximize the entropy

$$S = -\int P(u^\mu_n) \log\left(\frac{P(u^\mu_n)}{P_0(u^\mu_n)}\right)$$

subject to the constraints that the total probability sums to one and the average of the distribution is $\bar{u}_n$. The entropy measures the amount of information needed to specify the variation of $P(u_n^\mu)$ away from $P_0(u_n^\mu)$. By maximizing the entropy we chose the least informative distribution, which is also overwhelmingly the most likely distribution. Without any constraint on the mean we would find $P(u_n^\mu) = P_0(u_n^\mu)$, as we would expect. Maximizing the entropy gives

$$P(u_n^\mu) \propto \mathrm{e}^{\lambda\, u_n^\mu} P_0(u_n^\mu),$$

where λ is a Lagrange multiplier which must be chosen so that $P(u_n^\mu)$ has an average of $\bar{u}_n$. Using the Gaussian approximation we find

$$P(u_n^\mu) \approx \frac{1}{\sqrt{2\pi}\,\Delta(K_1)} \mathrm{e}^{-(u_n^\mu - \bar{u}_n)^2/(2\Delta^2(K_1))}. \tag{19}$$

We must now compute how $P(u_n^\mu)$ changes under selection. We consider an infinite population (because the quantities of interest are linear in n_μ this does not lead to any systematic error). The distribution of u_n^μ after selection is given by

$$P^s(u_n^\mu) \propto \int_{-\infty}^{\infty} \rho_s(F)\, P(u_n^\mu|F)\, \mathrm{d}F = \int_{-\infty}^{\infty} r(F)\rho(F)\, P(u_n^\mu|F)\, \mathrm{d}F, \tag{20}$$

where $\rho_s(F)$ is the distribution of fitnesses after selection and $r(F)$ is proportional to the rank

$$r(F) = 2\int_{-\infty}^{F} \rho(F')\, \mathrm{d}F'.$$

We expand $P(u_n^\mu|F)$ around $F = K_1$ giving

$$P(u_n^\mu|F) = P(u_n^\mu|K_1) + (F - K_1)\frac{\partial P(u_n^\mu|K_1)}{\partial K_1} + \cdots$$

substituting into (20) we find

$$P^s(u_n^\mu) = P(u_n^\mu|K_1) + (K_1^s - K_1)\frac{\partial P(u_n^\mu|K_1)}{\partial K_1} + \cdots \approx P(u_n^\mu|K_1^s).$$

Using (19) we find

$$P(u_n^\mu) \approx \frac{1}{\sqrt{2\pi}\,\Delta(K_1^s)} \exp\left(\frac{-\left(u_n^\mu - \bar{u}_n\,\Delta(K_1^s)/\Delta(K_1)\right)^2}{2\Delta^2(K_1^s)}\right).$$

From which we find

$$\bar{u}_n^s = \frac{\Delta(K_1^s)}{\Delta(K_1)}\,\bar{u}_n = \frac{L - R^s}{L - R}\,\bar{u}_n. \tag{21}$$

This is the result we were seeking, namely the change in $\bar{u}_n$ under selection. The final result is easy to understand intuitively. The average of the distribution of $\bar{u}_n$'s is reduced towards zero in proportion to the change in width of the distribution. Of course, this is only an approximation, we have made a maximum entropy assumption and ignored high-order corrections. We discuss the effect of including correction terms at the end of this section.

Cross-correlation

The cross-correlation, $V(k)$, and two string correlation, $Q = V(0)$, are more complicated than the auto-correlation as they involve pairs of string. As a consequence genetic drift (present in a finite population) will produce a systematic change in the cross-correlation. The cross-correlation after selection is given formally by

$$\begin{aligned}
V^s(k) &= \frac{1}{P(P-1)} \sum_{\alpha \neq \beta} V^s_{\alpha\beta}(k) \\
&= \frac{1}{P(P-1)} \sum_{\alpha,\beta} V^s_{\alpha\beta}(k) - \frac{1}{P-1} U^s_{\alpha}(k) \\
&= \frac{1}{P(P-1)} \sum_{\mu,\nu} n_\mu n_\nu V_{\nu\mu}(k) - \frac{1}{P-1} n_\mu U_\mu(k). \\
&= \frac{1}{P(P-1)} \sum_{\mu} (n_\mu^2 - n_\mu) U_\nu(k) + \frac{1}{P(P-1)} \sum_{\mu \neq \nu} n_\mu n_\nu V_{\nu\mu}(k). \qquad (22)
\end{aligned}$$

The first term on the right-hand side, involving the auto-correlation, is a finite population effect that arises from duplication of the fit individuals. The formulae (22) cannot be used directly as we would need to know how the correlation is related to the fitness (and thus to n_μ), which is information we do not have. We could try to infer this information as we did for the auto-correlation, however, this would be very complicated. Instead we take a different approach and consider the change in correlation to be the result of two effects—a natural correlation and a duplication term (this approach was first used in [11, 12]). The natural correlation arises because the population becomes fitter so that individuals are more alike (they contain more 1s). The duplication term arises in a finite population and is a consequence of certain members of the population being duplicated and are thus over represented. These two effects can be computed separately. We can compute the natural correlation by asking what happens in an infinite population. To calculate the duplication effect we consider the effect of a 'neutral selection' where each member of the population is assigned an arbitrary rank independent of its fitness. We then combine these two effects to obtain an estimate of the full effect of selection.

To model the natural change in correlation, we define the 'natural correlation'

$$R = \overline{\frac{1}{P^2} \sum_{\alpha=1}^{P} \sum_{\beta=1}^{P} \sum_{i=1}^{L} S_i^\alpha T_i^\beta}, \qquad (23)$$

where the bar denotes the ensemble average and $\mathbf{S}^\alpha$, and $\mathbf{T}^\beta$ are strings from different members of the ensemble. That is, R denotes the average correlation between string from independent runs of the GA. For the ones-counting problem

$$R = \frac{\bar{K}_1^2}{L}. \qquad (24)$$

Since we know how $\bar{K}_1$ changes under selection we can easily compute the natural correlation after selection. We can define the 'natural linkage' as

$$Z(k) = U(k) - R.$$

This would be the linkage we would expect in an infinite population. However, in a finite population, there is a discrepancy between the natural linkage and true linkage due to duplication effects. To model these we define the deviation, $d(k)$, between the true and natural linkage to be

$$W(k) = d(k)\, Z(k).$$

The linkage after selection is given by

$$W^s(k) = d^s(k)\, Z^s(k).$$

We know how to compute $Z^s(k)$ from the change in the natural selection and the change in $U(k)$. We need to compute $d^s(k)$. For this we use a trick: we assume that the change in $d^s(k)$ is the same if we were to perform a neutral selection. That is, a selection where we rank the members of the population in an arbitrary order. We can compute the change in the $d(k)$ exactly for neutral selection. Thus we are assuming the duplication effects and selection effects are independent (this is exactly the same assumption made in estimating the change in the variance [22]).

Under neutral selection

$$\text{NeutSel: } W(k) \to W^{ns}(k) = d^{ns}(k)\, Z^{ns}(k) = d^{ns}(k)\, Z(k)$$

since $Z^{ns}(k) = Z(k)$. From the definition of $W(k)$, (9), and (22) we find

$$W^s(k) = \frac{1}{P(P-1)} \sum_{\mu} (P\, n_\mu - n_\mu^2) U_\nu(k) - \frac{1}{P(P-1)} \sum_{\mu \neq \nu} n_\mu n_\nu V_{\nu\mu}(k).$$

We cannot evaluate this for normal selection as we do not know how the correlation is related to the fitness, but for neutral selection the number of times we choose an individual is independent of any correlation so that we can average over all possible ways of doing neutral selection. This gives

$$W^{ns}(k) = \left(\frac{P - \langle n^2 \rangle}{P-1} \right) \left(\frac{1}{P} \sum_{\mu} U_\nu(k) - \frac{1}{P(P-1)} \sum_{\mu \neq \nu} V_{\nu\mu}(k) \right) = r_P\, W(k),$$

where we have used the identity $\langle n_\mu n_\nu \rangle = (P - \langle n^2 \rangle)/(P-1)$. We make the approximation $d^{ns}(k) \approx d^s(k)$ since exactly the same number of strings will be duplicated under neutral selection as under normal selection (although this ignores any correlation between the fluctuations in the members being selected and their fitness). Thus,

$$d^s(k) \approx d^{ns}(k) = \frac{W^{ns}(k)}{Z(k)} = r_P\, \frac{W(k)}{Z(k)}$$

and

$$W^s(k) \approx r_P\, \frac{Z^s(k)}{Z(k)} W(k).$$

To calculate the change in the two string correlation function we consider the special case when $k = 0$, so that $W(0) = L - Q$ and $Z(0) = L - R$,

$$L - Q^s \approx r_P \left(\frac{L - R^s}{L - R} \right) (L - Q) = r_P \left(\frac{L^2 - (\bar{K}_1^s)^2}{L^2 - \bar{K}_1^2} \right) (L - Q).$$

Although this is an approximation it is found empirically to be remarkably accurate.

To calculate the change in $W(k)$ we would need to keep $U(k)$ and $V(k)$ as macroscopic variables, however, numerically this is difficult because $W(k)$, being the difference of two large quantities, $W(k) = U(k) - V(k)$, is very sensitive to any small errors. To avoid this we make the assumption that $V(k) = R$, so that $W(k) = Z(k)$ and $W(k)\, Z^s(k)/Z(k) = Z^s(k) = U^s(k) - R^s(k)$. Under this assumption

$$\bar{w}_n^s = r_P \left(\frac{L - R^s}{L - R} \right) \bar{w}_n. \tag{25}$$

This is the selection formula for $\bar{w}_n$ (with $n > 0$). The spatial variations in linkage (i.e., the value of $\bar{w}_n$) are reduced due to two effects. The first r_P is due to duplications. The second comes from a selection for smoother functions, which tend to have higher fitness. Initially, both these effects are very small, however, as the population becomes fitter the selection for smoother functions (smaller u_n^α) becomes significant.

3.2 Mutation

Mutation affects each site of a string independently

$$\text{Mut: } S_i \to S_i^m = \begin{cases} -S_i & \text{with probability } u \\ S_i & \text{with probability } 1-u \end{cases}$$

The effect of mutation is simple to calculate and has been described in many places, for example, reference [10, 13]. Denoting $\Upsilon = (1 - 2u)$, we can write the effect of selection as

$$\begin{array}{llcll} \text{Mut:} & \bar{K}_1 & \to & \bar{K}_1^m & = \Upsilon \bar{K}_1 \\ \text{Mut:} & \bar{K}_2 & \to & \bar{K}_2^m & = \Upsilon^2 \bar{K}_2 + (1 - \Upsilon^2) L \\ \text{Mut:} & \bar{Q} & \to & \bar{Q}^m & = \Upsilon^2 \bar{Q} \\ \text{Mut:} & \bar{W}(k) & \to & \bar{W}^m(k) & = \Upsilon^2 \bar{W}(k). \end{array} \tag{26}$$

The last equality implies

$$\text{Mut: } w_n \to w_n^m = \Upsilon^2\, w_n. \tag{27}$$

These expressions are exact.

3.3 Recombination

In recombination we randomly pair the members of the population and apply the crossover operator

$$\text{TPC: } (\mathbf{S}^\mu, \mathbf{S}^\nu) \to (\mathbf{S}^\alpha, \mathbf{S}^{\alpha'}),$$

where the string variables in the two offspring are given by

$$S_i^\alpha = \chi_i^\alpha S_i^\mu + (1 - \chi_i^\alpha) S_i^\nu$$

$$S_i^{\alpha'} = (1 - \chi_i^\alpha) S_i^\mu + \chi_i^\alpha S_i^\nu,$$

with $\chi_i \in \{0, 1\}$. By performing crossover in complementary pairs we leave the string variables (alleles) conserved, thus removing any genetic drift from crossover. As a consequence the average fitness, $\bar{K}_1$, and correlation, $\bar{Q}$, are unchanged by crossover. How we determine χ_i^α depends on the type of crossover.

Two-point crossover

For two-point crossover, we choose cutting points A and B (with $A < B$) so that

$$\chi_i^\alpha = \begin{cases} 1 & \text{if } A \geq i > B \\ 0 & \text{if } i < A \text{ or } i \geq B. \end{cases}$$

We will consider the case where $A \in \{1, 2, \ldots, L/2\}$ is a randomly chosen cutting point and $B = A + L/2$, so that half the string variables are exchanged at each crossover.

The second cumulant can be calculated from the identity (10),

$$\bar{K}_2^c = L - \bar{Q}^c + \sqrt{L-1}\,\bar{w}_0^c, \tag{28}$$

although to use this formula we must calculate how $W(k)$ is changed by crossover (the correlation is unchanged by crossover $\bar{Q}^c = \bar{Q}$). To calculate the change in the auto-correlation function $U_\alpha(k)$ we note that

$$\begin{aligned} U_\alpha^c(k) &= \sum_{i=1}^{L} S_i^\alpha S_{i+k}^\alpha \\ &= \sum_{i=1}^{L} \left(\chi_i^\alpha S_i^\mu + (1-\chi_i^\alpha) S_i^\nu\right) \left(\chi_{i+k}^\alpha S_{i+k}^\mu + (1-\chi_{i+k}^\alpha) S_{i+k}^\nu\right). \end{aligned}$$

We now average over all ways of all ways of performing crossover. For two-point crossover

$$\begin{aligned} \left\langle \chi_i^\alpha \chi_{i+k}^\alpha \right\rangle_A &= \frac{[L-2k]_L}{2L} \\ \left\langle \chi_i^\alpha (1-\chi_{i+k}^\alpha) \right\rangle_A &= \frac{1}{2}\left(1 - \frac{[L-2k]_L}{L}\right), \end{aligned}$$

from which we find

$$\langle U_\alpha^c(k) \rangle_A = \frac{[L-2k]_L}{2L}\left(U_\mu(k) + U_\nu(k)\right) + \frac{1}{2}\left(1 - \frac{[L-2k]_L}{L}\right)\left(V_{\mu\nu}(k) + V_{\nu\mu}(k)\right).$$

Averaging over all members of the population we find

$$U^c(k) = U(k) + \left(\frac{[L-2k]_L}{L} - 1\right) W(k).$$

The cross-correlation between two arbitrary strings

$$V_{\alpha,\beta}(k) = \sum_i S_i^\alpha S_{i+k}^\beta$$

is unchanged on average by crossover. However, in the case when the strings correspond to complementary offspring produced from the same pair of parent strings, the cross-correlation will change. In this case

$$V_{\alpha,\alpha'}^c(k) = \left(1 - \frac{[L-2k]_L}{L}\right) \frac{U_\mu(k) + U_\nu(k) - V_{\mu,\nu}(k) - V_{\nu,\mu}(k)}{2} + \frac{V_{\mu,\nu}(k) + V_{\nu,\mu}(k)}{2}.$$

Thus, on average

$$V^c(k) = V(k) + \left(\frac{1}{P-1}\right)\left(1 - \frac{[L-2k]_L}{L}\right) W(k)$$

and

$$\text{TPC: } \bar{W}(k) \to \bar{W}^c(k) = \left(\frac{P}{P-1}\right)\frac{[L-2k]_L}{L}\bar{W}(k) - \frac{\bar{W}(k)}{P-1}.$$

From the definition of the orthogonal polynomials the coefficients w_n are changed by

$$\begin{aligned} \bar{w}_0^c &= \frac{P}{P-1}\left(a_0\bar{w}_0 + \sqrt{b_1}\bar{w}_1\right) - \frac{\bar{w}_0}{P-1} \\ \bar{w}_n^c &= \frac{P}{P-1}\left(\sqrt{b_n}\bar{w}_{n-1} + a_n\bar{w}_n + \sqrt{b_{n+1}}\bar{w}_{n+1}\right) - \frac{\bar{w}_n}{P-1}, \qquad n > 0, \end{aligned} \tag{29}$$

where a_n and b_n are known variables (see Appendix A). The correction arising from the complementary offspring is usually negligible, however, we included it because the equations describing crossover are then exact.

Although in this paper we have stressed the calculation of the effect of selection, the choice of macroscopic variables was determined by the crossover calculation. We can compute the change in the second cumulant under two-point crossover exactly if we know how $W(k)$ changes, which we have just seen we can calculate exactly. The price of using $W(k)$ as a macroscopic variable is that it is difficult to calculate how it changes under selection, which is why that calculation has been stressed in this paper. Fortunately, mutation was very simple to calculate for this quantity.

Uniform crossover

For comparison, in uniform crossover the string variable are shuffled at random between the two parents

$$\chi_i = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

The only difference between modelling uniform and two-point crossover is in computing the effect of crossover on $\bar{W}(k)$. It is straightforward to show that

$$\text{UC: } \bar{W}(k) \to \bar{W}^c(k) = \frac{\bar{W}(k)}{2}$$

for uniform crossover (e.g. [10]). In uniform crossover $\bar{W}(k)$ remains flat throughout the evolution and so

$$\bar{K}_2^c = \frac{\bar{K}_2 + L - Q}{2}.$$

3.4 Corrections

The error in our calculation comes from computing the effect of selection. The change in the cumulants depend on higher cumulants, which we do not keep. To include higher cumulants, such as the skewness, we would have to compute how they change under crossover, which we could only do exactly by including three string correlation functions. It may be possible to estimate the three string correlations from the two string correlations using maximum entropy arguments, but we have not attempted this. We can get a feel for the size of these corrections by putting in the empirically measured third cumulant, this does not change the qualitative behaviour, but it does affect the quantitative behaviour.

The second major source of error comes from estimating the change to the linkage function $W(k)$ caused by selection. The accuracy of the constant term $\bar{w}_0$ depends on the quality of the estimate for the change in Q and $\bar{K}_2$, which are discussed above. Our prediction of the change in the non-constant (spatially varying) terms $\bar{w}_n$ is dependent on a number of approximations. A relatively innocuous approximation is the truncation of the series $\sum_n \bar{w}_n \psi_n(k)$. The $\bar{w}_n$'s fall off rapidly with n and including any more than five terms in the series gives rise to no appreciable differences in the results. A more significant approximation was in using a Gaussian to approximate $P(u_n^\alpha|F)$. This is a very poor approximation of the true distribution $P(u_n^\alpha|F)$, and yet it (together with the rest of the maximum entropy calculation) seems to give a reasonable estimate of the change in $\bar{u}_n$ caused by selection. We tried incorporating a correction due to the skewness of $P(u_n^\alpha|F)$, but found that it gave no appreciable correction to the Gaussian approximation.

As we will see, our estimate of $\bar{w}_n$ is poor for large n in intermediate times (10–30 generations). This may be a consequence of the approximations we invoked in the maximum entropy calculation. Another source of error might come from our perfunctory treatment of estimating the change in $\bar{w}_n$ from $\bar{u}_n$. This prevented us from having to treat the $\bar{u}_n$'s and $\bar{v}_n$'s as macroscopic variables. These are harder to treat as they are related to $\langle F_\alpha^2 \rangle$ and $\langle F_\alpha F_\beta \rangle$ which are given in terms of nonlinear functions of the cumulants and thus depend on the ensemble fluctuations. We was unable to estimate them satisfactorily using the other macroscopic variables, so was forced into a somewhat crude approximation. It would be difficult to improve this situation without including ensemble variables describing the ensemble fluctuations, although these have never been calculated for ranking selection.

4 Model Dynamics

Cumulants and Correlation

Using the results from the previous section we can construct a simple model of the (ensemble) dynamics by iterating the update equations for the macroscopic parameters. Figure 3 shows a comparison between simulations and our theoretical model for the first two cumulants, and the two string correlation. It shows the behaviour expected. The initial (finite population) variance is $\bar{\kappa}_2 = (1 - 1/P)L$, but this decreases due to selection, although it is increased again by mutation and crossover. (The finite population cumulants we denote by κ_n. They are related to the unbiased estimates of the cumulants for the infinite population, K_n—also known as Fisher's k-statistics—according to $\bar{\kappa}_1 = \bar{K}_1$ and $\bar{\kappa}_2 = (1 - 1/P)\bar{K}_2$.) The two string correlation increases from zero. Eventually an equilibrium is reached where the decrease in variance caused by selection is balanced by the increase caused by mutation and crossover. The average fitness is increased by $\sqrt{\bar{K}_2/\pi}$ during selection and decreased by a factor $1 - 2u$ by mutation. It is unaffected by crossover. Thus, the average fitness is strongly dependent on the variance.

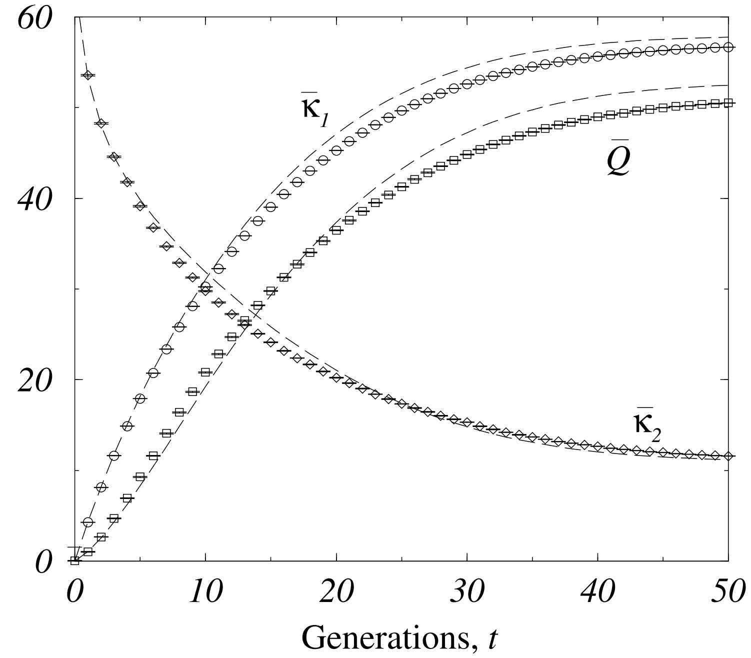

Figure 3: Theoretical (dashed lines) and simulation (data points) are shown for the first cumulant (○), second cumulant (◇), and two string correlation (□). The results are for the ones-counting problem with $L = 64$, a population size of 100, a mutation rate of $1/64$ and using two-point crossover. The theory uses two cumulants. The simulations are averaged over 10 000 runs.

The agreement between theory and simulation is not outstanding, particularly for long times. The most significant contribution to this discrepancy comes from ignoring higher cumulants. As the population becomes fitter the population tends to become negatively skewed. This has the effect of reducing both the first and second cumulants. The higher cumulants only affect selection—the mutation and crossover results are exact and depend only on the first two cumulants and the correlation. We did not include higher cumulants into our model as we would need to introduce many more macroscopic variables to calculate how they change under crossover. However, we can use an empirical measure of the third cumulant to judge the error we can expect from neglecting it. Empirically it is found that

$$\bar{K}_3 \approx \frac{-16}{1 + \mathrm{e}^{0.3(10-\bar{K}_1)}}.$$

This is an entirely *ad hoc* formula, but it gives a good fit to the third cumulant. Using this third cumulant we obtain a correction to the change in the first two cumulants caused by selection—the corrections due to skewness are given in Appendix C, which are taken from [23]. Figure 4 shows the three cumulant theory against simulation results. We show the third cumulant, although it should be stressed that the agreement between the theory curve and simulations is a result of fitting by hand and not a prediction of the theory. Adding in the third cumulant substantially improves the prediction of the first two cumulants and the correlation. There remains a small discrepancy, but it is now much harder to attribute this to any one approximation. There will be corrections from higher cumulants although empirically the fourth cumulant is found to be relatively small and is unlikely to make a large difference. More substantially we have ignored corrections due to fluctuations in the ensemble. Such corrections have been considered in a simpler model and were found to be a significant source of error [13]. Another, important source of error might come from the linkage corrections. We turn to these next.

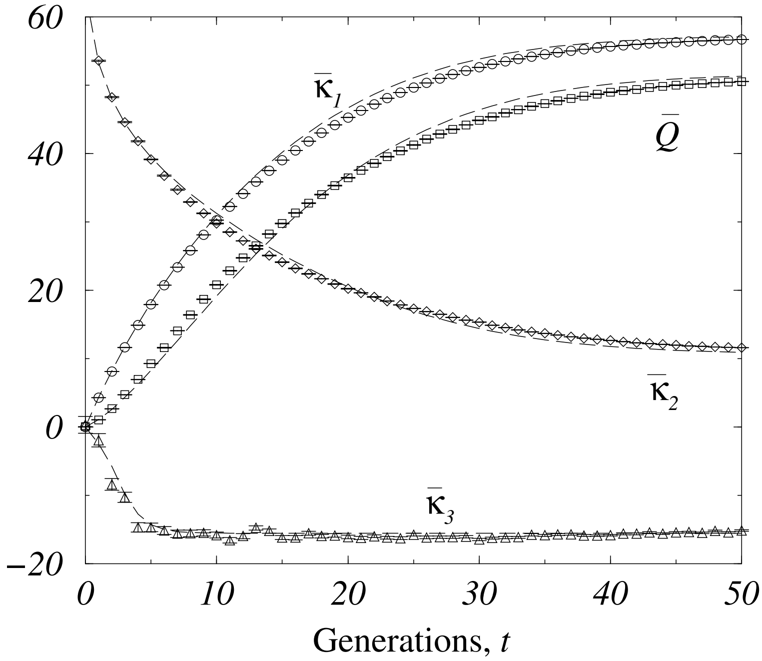

Figure 4: Theoretical (dashed lines) and simulation (data points) are shown for the first cumulant (○), second cumulant (◇), third cumulant (△), and two string correlation (□). The results are for the same problem as Figure 3, except we have included an empirical estimate for the third cumulant in the theory.

Linkage

The main focus of this paper has been on the spatial correlations induced by linkage. Figure 5 shows the dynamics of the linkage coefficients versus the number of generations. The theoretical curves are computed using the three cumulant model (the three cumulant model more accurately predicts the linkage coefficients than the two cumulant model, although the difference in not that marked). We obtain the correct qualitative behaviour although the quantitative agreement is poor during the middle of the run. The theory predicts a faster fall off of the high-linkage coefficients than actually occurs—this is also true of higher order coefficients which we have modelled, but not shown. The cause of the discrepancy is most likely to be a consequence of the errors in estimating the selection effects on the auto-correlation function, although the evolution equations are sufficiently coupled that it is difficult to pin-point the precise source of the discrepancy.

Figure 6 shows the linkage function $W(k)$ at different generations. The solid lines show simulation results while the dotted lines show the theoretical predictions. The theory gives good predictions in the initial evolution and as the evolution approaches the

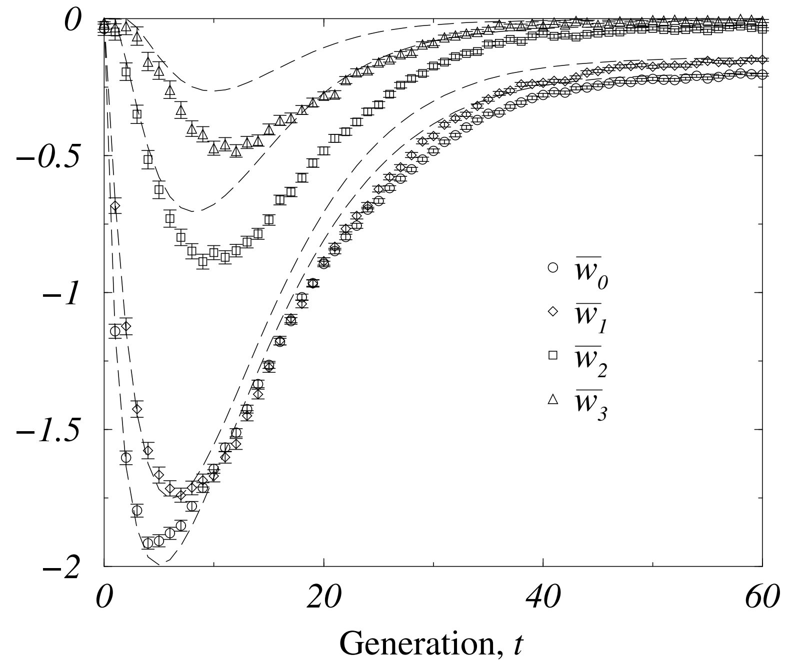

Figure 5: Theoretical (dashed lines) and simulation (data points) are shown for the first four linkage coefficients, w_0 (○), w_1 (◇), w_2 (△), and w_3 (□). The results are for the same system as shown in the previous figures. The theoretical predictions initially follow the simulations. The simulation results show averages over 10 000 runs.

equilibrium state, but there is a discrepancy for intermediate times which underestimates the degree of nonlinearity in $W(k)$. From Figure 6 we see that neighbouring sites are less correlated than sites on the opposite side of the strings. This is an indication that the bad alleles ($S_i^\alpha = -1$) are less likely to occur close together. This is what we would intuitively expect in this model. If bad alleles lie on nearby sites then it is likely that one offspring will inherit this pair of bad alleles. This will give the offspring a selective disadvantage. Eventually solutions with neighbouring bad alleles are eliminated. After many generations there are so few bad alleles that there is little room for any structure, thus $W(k)$ once again becomes linear (the linkage function will never be flat because of the action of two-point crossover).

5 Discussion

We have derived a model that captures the qualitative features of a GA with two-point crossover and, for the most part, is in reasonably good quantitative agreement. The ones-counting problem is particularly simple to model as it contains no epistasis (i.e., no interactions between sites). As a result the process has a limited 'memory' (where the population ends up does not depend on its past history). The modelling process is therefore forgiving of small errors—they tend to dissipate rather than grow. The difficulty was, therefore, always in modelling the intermediate time scales where memory effects are important. The GA remembers its past through the distribution of fitnesses, the string

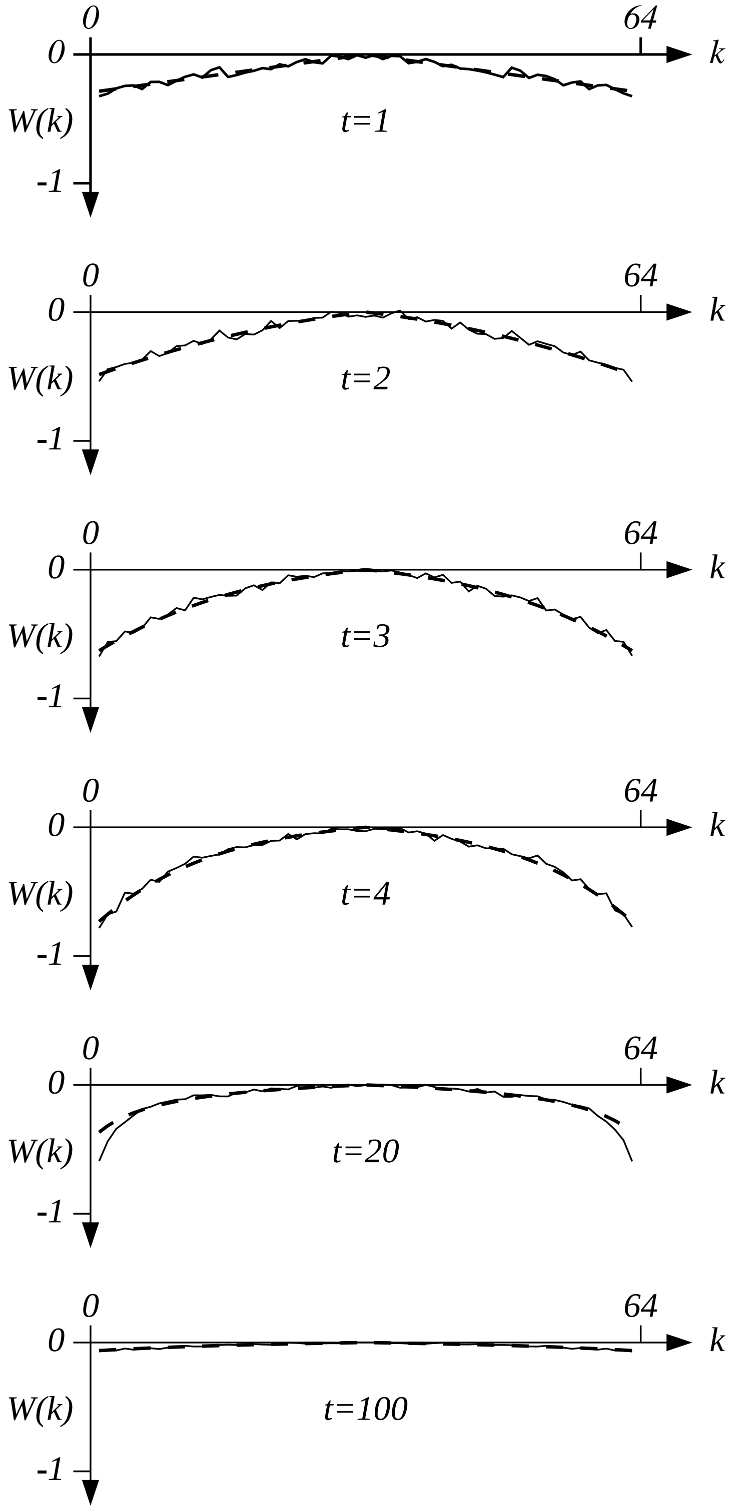

Figure 6: The linkage function $W(k)$ is shown at generation 1, 2, 3, 4, 20, and 100. The simulation results are shown as a solid curve, while the theoretical predications are shown as dotted curves. The simulations are average over $10\,000$ runs. We do not show error bars, but the error in the means can by gauged from the size of the fluctuations.

correlation and, with multipoint crossover, the auto-correlation. In contrast, a GA using uniform crossover has no structure in its auto-correlation making it considerable easier to model.

We can now describe in more detail the impact of different crossovers on the dynamics of a GA. Some of these features will be peculiar to the ones-counting problem, although we shall try to focus on the generic aspects of the dynamics. The increase in the average fitness through selection is proportional to the width of the distribution of fitnesses. Ranking selection strongly reduces the fitness variance. Crossover uses the genetic variation to restore the fitness variance towards the 'linkage equilibrium' value of $L - \bar{Q}$. The difference between the actual variance and the linkage equilibrium variance is given by $\sqrt{L-1}\,\bar{w}_0$. Crossover, reduces $\bar{w}_0$ and thus takes the variance towards the $L - \bar{Q}$. In uniform crossover, $\bar{w}_0$ is halved by each crossover, while, in two-point crossover, $\bar{w}_0$ is change by

$$\text{TPC: } w_0 \to w_0^c \approx \frac{w_0}{2} + \frac{w_1}{2\sqrt{3}}$$

(where we have neglected corrections of order $1/P$ and $1/L$). Thus two-point crossover drives the system towards linkage equilibria slower than uniform crossover because of the spatial variation in the linkage measured by w_1. As the population converges the deviation from linkage equilibrium w_0 becomes very small both for uniform and two-point crossover. As a consequence, the final equilibrium populations are very similar for both types of crossover. This asymptotic behaviour arises because of the simplicity of the ones-counting problem. It will not necessarily be the case for problems with a high degree of epistasis.

In deriving our model, we have been forced into making a number of approximations, some of which have been uncontrolled. This is the nature of this type of calculation. There exist no non-trivial limits where the complete analysis becomes exact. One exact limit is the neutral selection limit—which we exploited in finding the finite population corrections—however, this is not of much interest. A limit where the problem does simplify is the infinite population limit. For ranking selection the behaviour in the infinite population limit is not too different to that for finite populations. (This is not the case for Boltzmann selection where linkage disequilibria only arises from finite population corrections.) However, even in the infinite population limit we are faced with having to model the distribution of fitnesses and we are forced to infer properties of the distribution of auto-correlation coefficients. Thus, although we are spared complications to do with fluctuations (which we have largely ignored anyway) and some finite population corrections we are still cannot make any exact calculation. In addition, if we had taken the infinite population limit we would have difficulty comparing our results with simulations. Taking the limit of infinitely long strings leads to a further set of complications [20]. A final limit that is often used and significantly simplifies the analysis is the diffusion limit where linkage equilibrium is assumed [25], but this is obviously inappropriate for studying linkage. Thus, at present, the approximations adopted in the paper appear to be the best way forward if we wish to understand crossover-induced linkage.

The problem we have studied here is highly unrepresentative of hard optimization problems. The fact that it has no epistasis and consequently no local minima means that the final state of the GA does not depend on the past history. A far more interesting model would be one with many local minima. One model of this type that has been superficially studied using a statistical mechanics approach is the spin-glass chain [9, 10]. However, in these previous studies, the problem of spatial correlations was ignored. One of the contributions of this present paper is to provide some of the analytic tools necessary for these more complex types of problem to be studied.

Acknowledgement

I would like to thank Jon Shapiro and Alex Rogers for providing encouragement.

Appendix A Orthogonal Polynomials

The first two orthogonal polynomials are

$$\begin{aligned}
\psi_0(k) &= \frac{1}{\sqrt{L-1}} \\
\psi_1(k) &= \sqrt{\frac{3L}{L^3-4L^2+8L-8}}\Big((L-2)+2(L-1)[L-2k]_L\Big).
\end{aligned}$$

We can easily generate the orthogonal polynomials recursively. We define the unnormalized polynomials $\phi_n(k)$ through the recursion

$$\phi_{n+1}(k) = \left([L-2k]_L - a_n\right)\phi_n(k) + b_n\phi(k), \tag{30}$$

where

$$a_n = \frac{1}{c_n}\sum_{n=1}^{L-1}[L-2k]_L\phi_n^2(k), \qquad b_n = \frac{c_n}{c_{n-1}}, \qquad c_n = \sum_{n=1}^{L-1}\phi_n^2(k)$$

with $\phi_{-1}(k) = 0$, $\phi_0(k) = 1$ and $c_{-1} = 1$. (The coefficients a_n and b_n are those used in (29).) The orthogonal polynomials are given by

$$\psi_n(k) = \frac{1}{\sqrt{c_n}}\phi_n(k).$$

From the recurrence relation (30) we find

$$[L-2k]_L\psi_n(k) = \sqrt{b_{n+1}}\psi_{n+1}(k) + a_n\psi_n(k) + \sqrt{b_n}\psi_{n-1}(k)$$

from which we obtain the recursion (29).

Appendix B Moments of distribution of u_n^α

We are interested in find the moments of

$$u_n^\alpha = \sum_{k=1}^{L-1}\psi_n(k)\sum_{i=1}^{L}S_i^\alpha S_{i+k}^\alpha,$$

where $\langle S_i^\alpha\rangle = F_\alpha/L$ and the variables S_i^α are independent of each other. (Actually, this is an approximation as it only ensures that the string fitness is equal to F_α on average. For finite strings, this canonical approximation and the correct microcanonical average will give slightly different results. However, at the level of accuracy that we are working this does not give rise to a significant error.) Given the independence assumption we find

$$\langle u_n \rangle = \left(\frac{F_\alpha}{L}\right)^2 \sum_{k=1}^{L-1} \psi_n(k) = 0.$$

To calculate the second moment we have to separate terms where the variables S_i^α sit on the same site. The evaluation of $\langle u_n^\alpha \rangle$ reduces to a combinatoric problem of counting all configurations weighted by the number of variables siting at unique site. The calculation simplifies because

$$\sum_{k=1}^{L-1} \psi_n^2(k) = \sum_{k=1}^{L-1} \psi_n(k)\, \psi_n(L-k) = 1.$$

and

$$\sum_{k=1}^{L-1} \psi_n(k) = 0.$$

Using these we find

$$\left\langle u_n^2 \right\rangle = 2L \left(1 - \left(\frac{F_\alpha}{L}\right)^2\right)^2 \equiv \Delta^2(F).$$

The higher order moments can be calculated similarly, although the calculation rapidly becomes very complicated, with many more terms appearing. The leading term at third order is

$$\left\langle u_n^2 \right\rangle = \left(1 - \left(\frac{F_\alpha}{L}\right)^2\right)^3 S_n,$$

where

$$\begin{aligned} S_n &= \sum_{k=2}^{L-1} \sum_{l=1}^{k} \psi_n(k)\psi_n(l)\psi_n(k-l) + \sum_{k=1}^{L-1} \sum_{l=1}^{L-k-1} \psi_n(k)\psi_n(l)\psi_n(L-k-l) \\ &\quad + \sum_{k=1}^{L-1} \sum_{l=L-1-l}^{L-1} \psi_n(k)\psi_n(l)\psi_n(2L-k-l). \end{aligned}$$

Appendix C Selection of a Skewed Distribution

We give the corrections to selection caused by a skewness $s = \bar{K}_3/\bar{K}_2^{3/2}$, taken from [23],

$$\begin{aligned} \bar{K}_1^s &= \bar{K}_1 + \sqrt{\frac{\bar{K}_2}{\pi}} \left(1 - \frac{s^2}{96}\right) \\ \bar{K}_2^s &= r(P)\bar{K}_2 \left(1 - \frac{1}{\pi} + \frac{s}{2\sqrt{\pi}} + \frac{s^2}{48\pi} - \frac{s^4}{9216\pi}\right). \end{aligned}$$

These are not exact. They rely on the approximation we have used throughout this paper of separating the finite size effects from the selection expected in an infinite population—this error, however, is small as demonstrated in [23]. More significantly the skewness is highly nonlinear in the cumulants and we have ignored corrections due to fluctuations.

References

[1] R. Y., R. Y., and S. A., "A computational view of population genetics," *Random Structures & Algorithms*, vol. 12, no. 4, pp. 313–334, 1998.

[2] A. Prügel-Bennett, "The mixing rate of different crossover operators," in *Foundations of Genetic Algorithms 6*, San Francisco: Morgan Kaufmann, 2001. In press.

[3] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press (Ann Arbor), 1975.

[4] K. A. De Jong, *An Analysis of the Behaviour of a Class of Genetic Adaptive Systems*. PhD thesis, University of Michigan, 1975.

[5] G. Syswerda, "Uniform crossover in genetic algorithms," in *Proceedings of the Third International Conference on Genetic Algorithms*, pp. 2–9, Morgan Kaufmann (San Mateo), 1989.

[6] W. M. Spears and K. A. De Jong, "On the virtues of parameterised uniform crossover," in *Proceedings of the Fourth International Conference on Genetic Algorithms*, pp. 230–236, Morgan Kaufmann (San Mateo), 1991.

[7] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[8] P. Galinier and J. K. Hao, "Hybrid evolutionary algorithms for graph coloring," *Journal of Combinatorial Optimization*, vol. 3, no. 4, pp. 379–397, 1999.

[9] A. Prügel-Bennett and J. L. Shapiro, "An analysis of genetic algorithms using statistical mechanics," *Physical Review Letters*, vol. 72, no. 9, pp. 1305–1309, 1994.

[10] A. Prügel-Bennett and J. L. Shapiro, "The dynamics of a genetic algorithm for simple random Ising systems," *Physica D*, vol. 104, pp. 75–114, 1997.

[11] M. Rattray, "The dynamics of a genetic algorithm under stabilizing selection," *Complex Systems*, vol. 9, no. 3, pp. 213–234, 1995.

[12] M. Rattray and J. L. Shapiro, "The dynamics of genetic algorithms for a simple learning problem," *Journal of Physics: A.*, vol. 29, pp. 7451–7473, 1996.

[13] A. Prügel-Bennett, "Modelling evolving populations," *Journal of Theoretical Biology*, vol. 185, pp. 81–95, 1997.

[14] J. L. Shapiro and A. Prügel-Bennett, "Genetic algorithms dynamics in two-well potentials with basins and barriers," in *Foundations of Genetic Algorithms 4* (R. K. Belew and M. D. Vose, eds.), (San Francisco), pp. 101–116, Morgan Kaufmann, 1997.

[15] M. Rattray and J. L. Shapiro, "Noisy fitness evaluations in genetic algorithms and the dynamics of learning," in *Foundations of Genetic Algorithms 4* (R. K. Belew and M. D. Vose, eds.), (San Francisco), pp. 117–139, Morgan Kaufmann, 1997.

[16] S. Bornholdt, “Probing genetic algorithm performance of fitness landscapes,” in *Foundation of Genetic Algorithms 4* (R. K. Belew and M. D. Vose, eds.), (San Francisco), pp. 141–154, Morgan Kaufmann, 1997.

[17] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, “Finite populations induce metastability in evolutionary search,” *Physics Letters A*, vol. 229, pp. 144–150, 1997.

[18] A. Rogers and A. Prügel-Bennett, “Modelling the dynamics of steady-state genetic algorithms,” in *Foundations of Genetic Algorithms 5* (W. Banzhaf and C. Reeves, eds.), (San Francisco), pp. 57–68, Morgan Kaufmann, 1999.

[19] A. Rogers and A. Prügel-Bennett, “Genetic drift in genetic algorithm selection schemes,” *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 298–303, 1999.

[20] A. Prügel-Bennett, “On the long string limit,” in *Foundations of Genetic Algorithms 5* (W. Banzhaf and C. Reeves, eds.), (San Francisco), pp. 45–56, Morgan Kaufmann, 1999.

[21] A. Rogers and A. Prügel-Bennett, “Evolving populations with overlapping generations,” *Theoretical Population Biology*, vol. 57, no. 2, pp. 121–129, 2000.

[22] A. Rogers and A. Prügel-Bennett, “The dynamics of a genetic algorithm on a model hard optimization problem,” *Complex Systems*, vol. 11, no. 6, pp. 437–464, 2000.

[23] A. Prügel-Bennett, “Finite population effects for ranking and tournament selection,” *Complex Systems*, vol. 12, no. 2, pp. 183–205, 2000.

[24] M. Rattray, *Modelling the Dynamics of Genetic Algorithms using Statistical Mechanics*. PhD thesis, University of Manchester, UK, 1996.

[25] M. Rattray and J. L. Shapiro, “Cumulant dynamics of a population under multiplicative selection, mutation and drift.” LANL eprint archive, 2000.

[26] T. Blickle and L. Theile, “A mathematical analysis of tournament selection,” in *Proceedings of the Sixth International Conference on Genetic Algorithms* (L. J. Eshelman, ed.), (Hillsdale, New Jersey), pp. 9–16, Lawrence Erlbaum Associates, 1995.

[27] J. E. Baker, “Reducing bias and inefficiency in the selection algorithm,” in *Proceedings of the Second International Conference on Genetic Algorithms*, Lawrence Erlbaum Associates (Hillsdale), 1987.