Modelling Evolving Populations

Adam Prügel-Bennett
NORDITA, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

September 27, 1996

Abstract

A formalism is presented for modelling the evolutionary dynamics of a population of gene sequences. The formalism was originally developed for describing genetic algorithms. In this paper the formalism is elaborated by considering the evolution of an ensemble of populations. This allows the evolution to be modelled more accurately. To illustrate the formalism the problem of a population of gene sequences evolving in a multiplicative fitness landscape is considered. A comparison with simulations is made and shows very good agreement. More complicated problems have already been investigated including sexual recombination and evolution in a multi-valleyed fitness landscape. These results will be briefly reviewed.

1 Introduction

1.1 The formalism

In this paper we present a formalism for calculating the evolutionary dynamics of a population of gene sequences. The population evolves through a series of selection steps followed by modifications of the gene sequence. The fitness of each member of the population is assumed to be a function only of the gene sequence. Being able to understand this class of problems is a prerequisite for understanding many of the outstanding problems in population genetics. Although much insight can be gained from simulations a quantitative theory is clearly desirable. In this paper we present such a formalism. The formalism was first developed to describe the dynamics of a genetic algorithm (Prügel-Bennett & Shapiro, 1994; Shapiro *et al.*, 1994; Prügel-Bennett & Shapiro, 1995; Rattray, 1996). Here we develop the formalism further by considering the evolution of an *ensemble of populations* rather than just a single population. This provides a more coherent framework as well as being more accurate and giving additional information about the evolution. The formalism is similar in its structure to statistical mechanics. In this introduction we pursue this analogue. In the subsequent sections we show how the formalism can be used to derive quantitative equations for the evolution of a population evolving in a multiplicative fitness landscape.

The evolution of a single population can be regarded as a trajectory in the *phase space* of all possible sets of gene sequences. In order to simplify the discussion we will consider only populations of a fixed size. We denote this by P. We will also assume the genes are binary and have a fixed length N. In this case the phase space is a $P \times N$ dimensional hyper-cube. A population corresponds to a single point in the phase space. The *evolution operators*, selection, mutation, sexual recombination, etc. act on the population causing it to move in the phase space. This move will be stochastic, reflecting the stochastic nature of the evolution operators. Furthermore, selection and sexual recombination will cause a highly non-local move in the phase space. In principle the evolution could be described by a set of deterministic master equations for the probability density for the different sets of sequences. Although it is possible to write down the effects of the various evolution operators formally (Vose & Liepins, 1991; Nix & Vose 1991), the high dimensionality and the extremely complicated move-set makes this approach intractable for realistic problems.

As an alternative approach we have proposed a formalism whereby the dynamics is described in terms of a few *macroscopic variables*. For example, we have used the first few cumulants of the distribution of "fitnesses" in the population (Prügel-Bennett & Shapiro, 1994; Prügel-Bennett & Shapiro, 1995) with the possible addition of the average correlation within the population (Rattray, 1996; Rattray, 1996a). For each small region in the space of the macroscopic variables there will be a large number of possible points in the phase space of all sets of sequences. Thus, we are effectively integrating out all but a few degrees of freedom. In our previous description (Prügel-Bennett & Shapiro, 1994; Prügel-Bennett & Shapiro, 1995) we considered the evolution in terms of a single trajectory in the space of macroscopic variables. In this paper we describe the evolution in terms of the time development of an (non-interacting) ensemble of populations. The ensemble can be regarded as a probability density for a population to be at a particular point in the space defined by the macroscopic variables. This provides a more coherent framework as we can incorporate the stochastic nature of the evolution. It is more accurate because the stochastic fluctuations produce systematic corrections to the selection operator. Furthermore we obtain information about the expected spread of the final outcome. To describe the evolution of the ensemble we introduce a set of *ensemble variables*, for example, the ensemble average of the first few cumulants and their covariances. Unlike the macroscopic variables the ensemble variables are properties of the ensemble and not of a particular population.

To use this formalism it is necessary to calculate how the ensemble variables evolve under the various evolution operators. To do this we calculate how an arbitrary set of gene sequences is changed by the evolution operators and then average over all possible sets of gene sequences weighted by their probability of occurring. The drawback of this approach is that to know the correct weights for each set of gene sequences we may need an infinite set of ensemble variables. This is true, for example, in selection where we need to know the precise form of the distribution of fitnesses. To make the approach tractable we must use a finite set of ensemble variables. However, making this truncation may result in wrongly weighting some populations and thus we obtain only an approximate solution. This approximation may be improved either by modelling the unknown information more accurately, for example, by using maximum entropy techniques, or else by increasing the set of ensemble variables. A challenge presented by this approach is to find a compact set of ensemble variables which accurately describes the ensemble. Finding such a set may require trial and error, but, as a bonus, if we succeed, we have learned what are the important macroscopic variables in determining the evolution of a population. In practice we have found that we can achieve good agreement between our model and simulations with only a few ensemble variables.

1.2 An example: the multiplicative fitness landscape

As an illustration of this approach we consider a population of binary gene sequences which undergo selection and mutation only. It is not, however, the aim of this paper to provide a full analysis of this problem, which has been discussed at length in the literature (see, for example, Woodcock & Higgs, 1995). We denote the gene sequence by a vector $\vec{S}^\mu = (S_1^\mu, S_2^\mu, \ldots, S_N^\mu)$ of (Ising) variables $S_i^\mu \in \{-1, 1\}$ where $\mu = 1, \ldots, P$ labels the different members of the population. The genes can thus occur in only one of two alleles; a good one, $S_i^\mu = 1$, or a bad one, $S_i^\mu = -1$. We assign a *fitness measure*, F_μ, to each sequence $\vec{S}^\mu$ defined as

$$F_\mu = \sum_{i=1}^{N} S_i^\mu = N_+^\mu - N_-^\mu \tag{1.1}$$

where N_+^μ and N_-^μ are the number of good and bad genes respectively (of course, $N_+^\mu + N_-^\mu = N$). To perform selection we assign a weight, w_μ (which is the true fitness), to each member of the population and draw (with replacement) P new members from the old population with a probability $p_\mu = w_\mu / \sum_\nu w_\nu$. The weight is assumed to be equal to the Boltzmann factor[1]

$$w_\mu = \mathrm{e}^{\beta F_\mu}. \tag{1.2}$$

Selection is controlled by the parameter β: if $\beta = 0$ then each member of the population is drawn with the same probability; if β is infinite then only the fittest member of the population is selected. In mutation each gene S_i^μ is mutated with a probability γ.

Starting from a set of random gene sequences the selection operator will favour those sequences with more good alleles. Mutation will at first allow the population to explore the space of sequences, so that the average fitness of the population initially increases. Eventually, the number of good alleles will so outweigh the number of bad alleles that mutation is much more likely to be detrimental and cannot be sufficiently compensated for by selection for the mean fitness to continue to increase. When this happens the population settles into an equilibrium. Because the selection weight (or true fitness) is an exponential of an additive fitness measure it can be written as a product of weights, $\exp(\beta\, S_i^\mu)$, for each gene. Thus this model is just that for a population of gene sequences evolving in a multiplicative fitness landscape. In the genetic algorithm literature this model is called ones-counting or ones-max since the fitness measure is a function of N_+^μ the number of ones in the sequence.

We will also discuss the limit of long gene sequences $N \to \infty$. This is analogous to taking the thermodynamic limit in statistical mechanics. This limit includes Muller's ratchet as a special case.

Our formalism has been applied to populations evolving in environments described by more elaborate fitness measures (for example, where the genes have different efficacies or depend on each other) and including recombination (Prügel-Bennett & Shapiro, 1995; Rattray, 1996; Rattray, 1996a). To keep the discussion simple we will consider here only the simple model described above. However, in section 5, we will briefly review the results we have obtained previously.

[1] The use of a Boltzmann factor is just a convenient way to represent the problem: it is not used to establish the analogy between population dynamics and statistical mechanics. The more conventional way of writing this is $w_\mu = (1-s)^{N_-^\mu}$ which is equivalent to equation (1.2) with $\beta = -(\log(1-s))/2$.

1.3 Computing the evolution

For this problem the evolution depends only on the distribution of fitnesses in the populations. Although this is always true for selection it is not generally true for mutation. We will use the cumulants, κ_n, of the distribution of fitnesses, $\rho(F)$, as our macroscopic variables — here and elsewhere we will use the term *fitness distribution* to refer to the distribution of fitness measures F_μ rather than the true fitnesses $w_\mu = \exp(F_\mu)$. The first cumulant describes the mean fitness in the population, the second cumulant describes the variance in the fitnesses while the higher order cumulants measure how much the distribution differs from a Gaussian. The third and forth cumulants are related to the skewness and kurtosis. Possible ensemble variables to describe the ensemble of populations are the average of the first few cumulants $\bar{\kappa}_n$ and the covariances in the cumulants,

$$v_{n,m} = \langle\langle \kappa_n \kappa_m \rangle\rangle - \langle\langle \kappa_n \rangle\rangle \langle\langle \kappa_m \rangle\rangle \tag{1.3}$$

where $\langle\langle \cdots \rangle\rangle$ denotes the average over the ensemble of populations. However, it is more convenient to use a related set of ensemble variables namely the average cumulants and covariances for an ensemble of *infinite population distributions*.

For any distribution of fitnesses, $\rho_\infty(F)$ say, we can generate a finite population by randomly choosing P member from the distribution $\rho_\infty(F)$. Doing this many times we can generate an ensemble of distributions. The average cumulants for the ensemble are related to, but different from, those of $\rho_\infty(F)$. Denoting the cumulants of $\rho_\infty(F)$ by K_n then the average cumulants for the ensemble generated from $\rho_\infty(F)$ are $\bar{\kappa}_1 = K_1$, $\bar{\kappa}_2 = (1 - 1/P)\, K_2$, etc. We can use an ensemble of infinite population distributions as a representation for the true ensemble. The covariances for the true ensemble are again simply related to the cumulants and covariances of the ensemble for the infinite population distribution. Denoting the covariances for the infinite population distribution by $V_{n,m}$ we find, $v_{1,1} = V_{1,1} + K_2/P$, $v_{1,2} = (1-1/P)\,(V_{1,2} + K_3/P)$, etc. Details are given in appendix A. There are restriction on what ensembles can be modelled this way, for example $v_{1,1} \geq \kappa_2/P$, however, for the problem we consider here these restrictions are never violated. The advantage of using the "infinite population" ensemble is that the members of the population are chosen independently. This representation is analogous to using a canonical ensemble in statistical mechanics rather than a microcanonical ensemble. We will use the average cumulants $\bar{K}_n$ and covariances $V_{n,m}$ for the infinite population ensemble as ensemble variables. It should be stressed that using $\{\bar{K}_n, V_{n,m}\}$ as ensemble variables is entirely equivalent to using $\{\bar{\kappa}_n, v_{n,m}\}$ and is only done for calculational convenience. It does not imply that we are using an infinite population approximation.

To re-iterate our strategy we model the evolution of a population by considering how an ensemble of populations evolve. The ensemble at a particular generation t can be described by a probability density $P(\{\kappa_n\}, t)$ in the space of macroscopic variables. We illustrate the evolution schematically in figure 1. To parameterize the evolution we take the average of the first few cumulants for the infinite population ensemble and their covariances. The problem of describing the evolution is to calculate the (deterministic) evolution equations for the ensemble variables. Denoting the set of ensemble variables at generation t by $\Xi(t)$ then the change in the ensemble variables from one generation to the next can be calculated in stages as

$$\Xi(t) \xrightarrow{mut} \Xi^m(t) \xrightarrow{sel} \Xi^{m,s}(t) = \Xi(t+1) \tag{1.4}$$

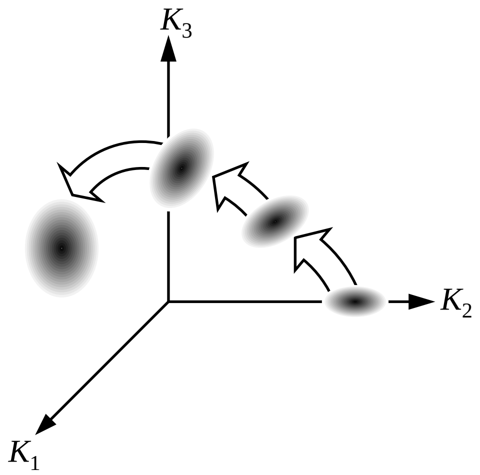

Figure 1: Schematic representation of the evolution of an ensemble of populations in the space of macroscopic variables, here cumulants of the distribution of fitnesses.

where the superscripts m and s denotes the ensemble variables after mutation and selection respectively.

To calculate how the ensemble variables evolve we consider what happens to an arbitrary set of gene sequences under the operation of selection or mutation. We then average over all sets of gene sequences weighted by their probability of being in the ensemble. In doing so we must parameterize the probability density for the ensemble, $P(\{K_n\}, t)$, in some way. For example we could assume that the ensemble is adequately described by a multi-variant Gaussian distribution. Averaging over the ensemble we obtain a set of non-linear equations relating the ensemble variables after applying an evolution operator to the ensemble variables before. Iterating these equation starting from an ensemble consisting of random sequences we obtain a model of the full evolution.

We will take as our ensemble variables the set $\Xi = \{\bar{K}_n, V_{l,m}\}$, where $n = 1, \ldots, n_c$ and $l+m \leq n_c$. In figure 2 we show an example of the predictions for the evolution of six of the ensemble variables using $n_c = 6$ together with simulation results. The parameters for the population were $P = 100$, $N = 100$, $\beta = 0.01$ and $\gamma = 4/N$. The theory is seen to agree very well with the simulations. Using less ensemble variables we obtain similar results except for a small discrepancy between the theory and simulations for the higher cumulants, particularly near to the onset of equilibrium at around 7 generations. These discrepancies become more pronounced for lower mutation rates and eventually the equations become unstable and do not correctly predict the long term evolution. This is a consequence of using a truncated cumulant expansion. We give a fuller discussion of the results to section 4.

1.4 Outline of paper

In the next section we will describe selection. In the first part of this section we discuss how selection changes the average cumulants in the ensemble — this has been described before in (Prügel-Bennett & Shapiro, 1994). In section 2.2 we discuss how selection changes the covariances of the cumulants. In section 3 we will describe the mutation

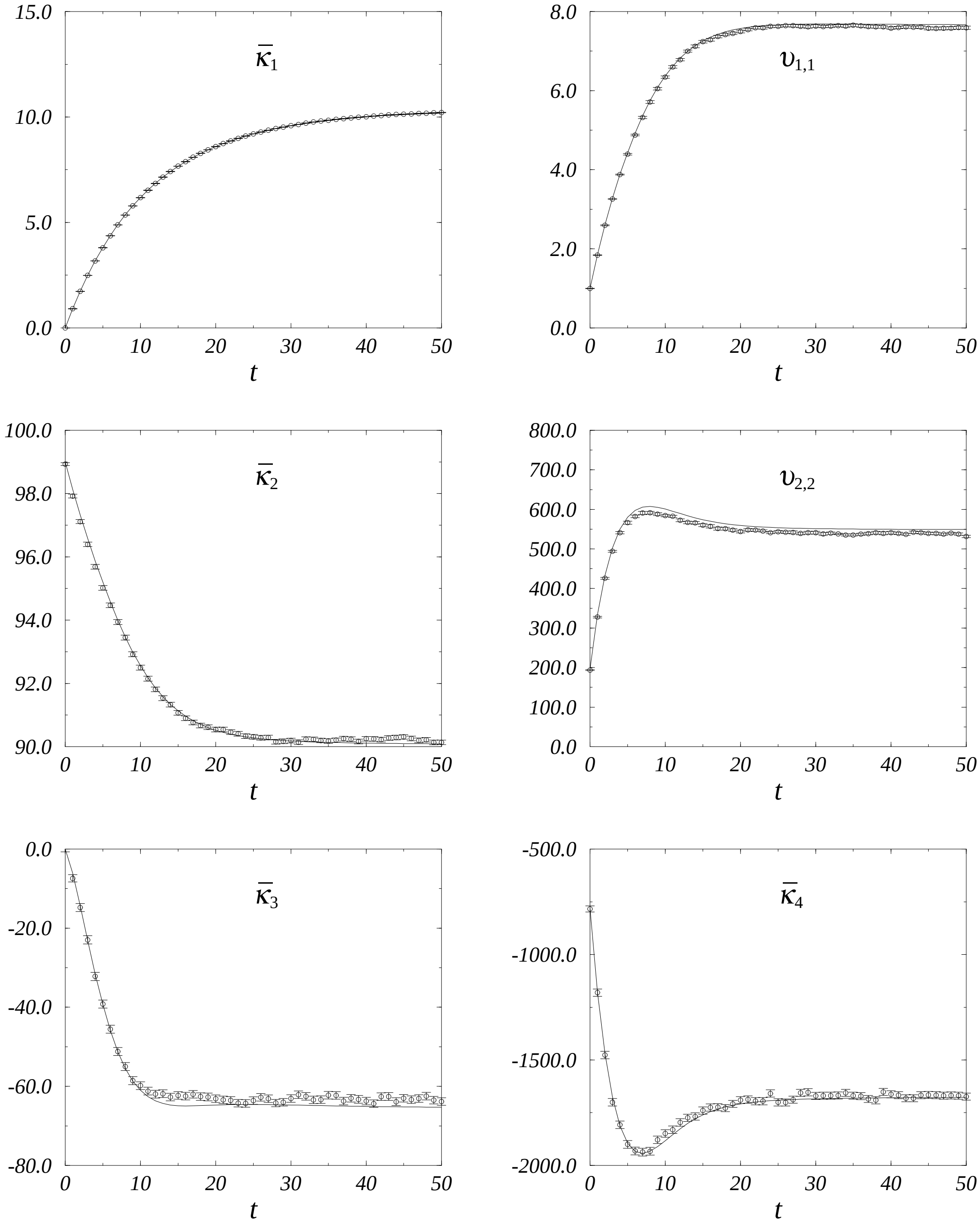

Figure 2: Examples of the predicted evolution of the ensemble variables $\bar{\kappa}_1$, $\bar{\kappa}_2$, $\bar{\kappa}_3$, $\bar{\kappa}_4$, $v_{1,1}$ and $v_{2,2}$ (solid lines) together with simulations (data points). In the example $P = 100$, $N = 100$, $\beta = 0.01$ and $\gamma = 4/N$. The simulations were averaged over 100 000 runs. The error bars show the predicted errors in the mean.

operator. In section 4.1 we describe how we use the calculations of the previous two sections to model the full evolution, as illustrated in figure 2. In section 4.2 we describe the limit of infinite long sequences. In section 5 we discuss the results of this paper and related them to the results obtained elsewhere using this approach. Finite population corrections and implementational details are given in the appendices.

2 Selection

2.1 The average cumulants after selection

We consider a selection mechanism where each member of the population is assigned a weight (the reproductive fitness) $w_\mu = w(F_\mu)$ depending on its fitness measure, F_μ. A new population is created by drawing members from the old population with a probability $p_\mu = w_\mu / \sum_\nu w_\nu$. The draws are performed independently and the chosen member is replaced so that it can be selected again. We will mainly consider the case when the members of the population are weighted by a Boltzmann factor $w_\mu = \exp(\beta F_\mu)$, although many other choices can be used (Rattray *et al.*, 1996).

Selection is non-trivial for finite populations as it causes the population to correlate. This reduces the width of the population faster than otherwise expected. Although this correlation will be small in large populations it accumulates over generations and significantly reduces the efficiency with which the population explores the phase space of possible gene sequences.

To understand selection we will first consider what happens to the first couple of cumulants. Suppose we have a population consisting of P members and we select a new population also consisting of P members. If we denote the number of times an individual μ is selected by n_μ then the first two cumulants after selection will be given by

$$\begin{aligned} \kappa_1' &= \frac{1}{P} \sum_{\mu=1}^{P} n_\mu F_\mu \\ \kappa_2' &= \frac{1}{P} \sum_{\mu=1}^{P} \left(n_\mu - \frac{n_\mu^2}{P} \right) F_\mu^2 - \frac{1}{P^2} \sum_{\mu \neq \nu} n_\mu n_\nu F_\mu F_\nu . \end{aligned} \tag{2.1}$$

The probability of a particular draw, $\vec{n} = (n_1, n_2, \ldots, n_P)$, is given by the multinomial probability distribution

$$P(\vec{n}) = P! \prod_{\mu=1}^{P} \frac{p_\mu^{n_\mu}}{n_\mu!} \; \delta\left(\sum_{\nu=1}^{P} n_\nu - P \right). \tag{2.2}$$

Averaging equations (2.1) over all possible outcomes we find

$$\begin{aligned} \kappa_1^s &= \sum_{\mu=1}^{P} p_\mu F_\mu \\ \kappa_2^s &= \left(1 - \frac{1}{P}\right) \left(\sum_{\mu=1}^{P} \left(p_\mu - p_\mu^2\right) F_\mu^2 - \sum_{\mu \neq \nu} p_\mu p_\nu F_\mu F_\nu \right) \end{aligned} \tag{2.3}$$

where we have used

$$\langle n_\mu \rangle = P\, p_\mu, \qquad \langle n_\mu\, n_\nu \rangle = P\,(P-1)\, p_\mu\, p_\nu, \qquad \left\langle n_\mu^2 \right\rangle = P\,(P-1)\, p_\mu^2 + P\, p_\mu.$$

The factor $\mathcal{P}_2 = (1 - 1/P)$ in the expression for κ_2^s arises because there are P members in the *new* population. To find the cumulants for the infinite population distribution we use the relationship $K_1 = \kappa_1$ and $K_2 = \kappa_2/\mathcal{P}_2$, see appendix A. Assuming that the F_μ's are drawn from an infinite population ensemble we can treat them as independent. This allows us to average them independently over a distribution $\rho_\infty(F_\mu)$ with cumulants K_n. Finally we can average over an ensemble of distributions with average cumulants $\bar{K}_n$ and covariance $V_{n,m}$. Thus

$$\begin{aligned}
\bar{K}_1^s &= \left\langle\!\!\left\langle \left\langle \sum_{\mu=1}^{P} p_\mu\, F_\mu \right\rangle_\rho \right\rangle\!\!\right\rangle \\
\bar{K}_2^s &= \left\langle\!\!\left\langle \left\langle \left(\sum_{\mu=1}^{P} \left(p_\mu - p_\mu^2\right) F_\mu^2 - \sum_{\mu\neq\nu} p_\mu\, p_\nu F_\mu F_\nu \right) \right\rangle_\rho \right\rangle\!\!\right\rangle
\end{aligned} \tag{2.4}$$

where $\langle \cdots \rangle_\rho$ denotes the average

$$\left\langle f(F_\mu) \right\rangle_\rho = \left(\prod_{\nu=1}^{P} \int_{-\infty}^{\infty} \rho_\infty(F_\nu)\, \mathrm{d}F_\nu \right) f(F_\mu) \tag{2.5}$$

and $\langle\!\langle \cdots \rangle\!\rangle$ denotes the average over the ensemble.

To calculate all the cumulants after selection we can use the generating function

$$G_1(z) = \left\langle\!\!\left\langle \left\langle \operatorname*{Tr}_{\{\vec{n}\}} P(\vec{n}) \log \left(\sum_{\mu=1}^{P} n_\mu \mathrm{e}^{z\, F_\mu} \right) \right\rangle_\rho \right\rangle\!\!\right\rangle \tag{2.6}$$

where the trace denotes the sum over all possible ways of doing the selection. The average cumulants after selection are given by the derivatives of the generating function, $\bar{\kappa}_n = G_1^{(n)}(0)$. This formula simplifies if we wish to find the cumulants for the infinite population after selection since, when we select an infinite population, $n_\mu = P\, p_\mu = P\, w_\mu \sum_\nu w_\nu$. Expanding the logarithm

$$\log \left(\frac{\sum_\mu w_\mu \mathrm{e}^{z\, F_\mu}}{P \sum_\nu w_\nu} \right) = \log \left(\sum_\mu w_\mu \mathrm{e}^{z\, F_\mu} \right) - \log \left(\frac{1}{P} \sum_\nu w_\nu \right) \tag{2.7}$$

and noting that the second term does not depend on z so can be dropped, we obtain the generating function for the average cumulants of the infinite population

$$G_1^\infty(z) = \left\langle\!\!\left\langle \left\langle \log \left(\sum_{\mu=1}^{P} w_\mu \mathrm{e}^{z\, F_\mu} \right) \right\rangle_\rho \right\rangle\!\!\right\rangle \tag{2.8}$$

where $\bar{K}_n = G_1^{\infty\,(n)}(0)$. We are effectively breaking up the selection operation into two pieces; we first select an infinite population with the weighting factor $w(F_\mu)$, and then we select P members at random from the infinite population. It is easy to check the first two derivatives of $G_1^\infty(z)$ give the same results as equation (2.4).

When we select in proportion to the Boltzmann factor, equation (1.2), we can use the somewhat simpler generating function

$$G_1^\infty(\beta) = \left\langle\!\!\left\langle \left\langle \log\left(\sum_{\mu=1}^{P} \mathrm{e}^{\beta F_\mu}\right) \right\rangle_\rho \right\rangle\!\!\right\rangle \tag{2.9}$$

where the average cumulants after selection are now given by $\bar{K}_n = G_1^{\infty\,(n)}(\beta)$. For simplicity in the following discussion we will use the generating function given in equation (2.9) although the generalization to equation (2.8) is straightforward.

To calculate the effect of selection we can use a trick due to Derrida (Derrida, 1981) of writing the logarithm in terms of the integral representation

$$\log(Z) \equiv \int_0^\infty \frac{\mathrm{e}^{-t} - \mathrm{e}^{-tZ}}{t}\,\mathrm{d}t. \tag{2.10}$$

The generating function then becomes

$$G_1^\infty(\beta) = \left\langle\!\!\left\langle \int_0^\infty \frac{\mathrm{e}^{-t} - g^P(t,\beta)}{t}\,\mathrm{d}t \right\rangle\!\!\right\rangle \tag{2.11}$$

where

$$g(t,\beta) = \int \rho_\infty(F)\,\mathrm{e}^{-t\mathrm{e}^{\beta F}}\,\mathrm{d}\,F. \tag{2.12}$$

These two integrals can be performed numerically to obtain the cumulants after selection. This is described in appendix B.

We can also find an expansion in $1/P$ for cumulants after selection provided $\beta\sqrt{K_2}$ remains sufficiently small. For the system studied in this paper this expansion is valid. To perform this expansion we first expand the exponential in equation (2.12) in a power series in $t\exp(\beta F)$. This allows the integral over F to be performed. Re-exponentiating the leading term in t and performing the integral over t we obtain

$$G_1^\infty(\beta) \approx \hat{G}^\infty(\beta) = \left\langle\!\!\left\langle \log(\hat{\rho}_\infty(\beta)) - \frac{1}{2\,P}\frac{\hat{\rho}_\infty(2\beta)}{\hat{\rho}_\infty^2(\beta)} \right\rangle\!\!\right\rangle \tag{2.13}$$

where we have dropped constant terms and terms of order $1/P^2$, and where

$$\hat{\rho}_\infty(\beta) = \int_{-\infty}^{\infty} \rho_\infty(F)\,\mathrm{e}^{\beta F}\,\mathrm{d}F = \exp\left(\sum_{n=1} \frac{K_n \beta^n}{n!}\right). \tag{2.14}$$

Putting this into equation (2.13) we obtain

$$\hat{G}_1^\infty(\beta) = \left\langle\!\!\left\langle \sum_{n=1}^{n_c} \frac{\beta^n K_n}{n!} - \frac{1}{2P}\exp\left(\sum_{n=2}^{n_c} \frac{(2^n-2)\,\beta^n K_n}{n!}\right) \right\rangle\!\!\right\rangle. \tag{2.15}$$

We have obtained a generating function for the average cumulants after selection starting from a given set of cumulants. To calculate how the set of cumulants for an ensemble evolve we need to average over all members in the ensemble. To do this we must choose some parameterization of the probability density for the ensemble. Here we will model the ensemble as a multi-variant Gaussian distribution

$$P(\{K_n\}) = \frac{1}{\sqrt{(2\,\pi)^{n_c}\,\det[V]}}\,\mathrm{e}^{-\frac{1}{2}\sum_{n,m}(K_n-\bar{K}_n)[V^{-1}]_{n,m}(K_m-\bar{K}_m)} \tag{2.16}$$

where n_c is the number of cumulants we choose to keep. In using this form of distribution we are making assumptions about the higher cumulants. This may lead to systematic errors in which case we may need to include other ensemble variables to describe the ensemble distribution more precisely — the cost of this is more work on our part. Alternatively we may try to model the ensemble distribution using physical arguments. Thus, for example, the distribution of K_2 could be modelled by a gamma distribution to reflect the fact that K_2 is a sum of squares of residuals (see appendix B). However, in practice we have found that using the distribution (2.16) works very well.

Performing the average in equation (2.15) using equation (2.16) we find

$$\hat{G}_1^\infty(\beta) = \sum_{n=1}^{n_c}\frac{\beta^n\,\bar{K}_n}{n!} - \frac{1}{2P}\exp\left(\sum_{n=2}^{n_c} b_n(\beta)\,\bar{K}_n + \frac{1}{2}\sum_{n,m=2}^{n_c} b_n(\beta)\,b_m(\beta)\,V_{n,m}\right) \tag{2.17}$$

where $b_n(\beta) = (2^n-2)\,\beta^n/n!$. The average cumulants after selection can be computed by take derivative of $\hat{G}_1^\infty(\beta)$. The first cumulant changes as

$$\bar{K}_1^s = \bar{K}_1 + \beta\bar{K}_2 + \cdots - \frac{1}{P}\beta\bar{K}_2\mathrm{e}^{\beta^2\bar{K}_2^2+\cdots} + \cdots. \tag{2.18}$$

To leading order, $\bar{K}_1$ changes in proportion to β times the variance, $\bar{K}_2$, of the fitness distribution (this is only a good approximation if $\beta \sim 1/\sqrt{\bar{K}_2}$). The first cumulant does not influence the higher cumulants because selection is invariant under a rescaling of the fitness measures (i. e. $F_\mu \to F_\mu + c$). The second cumulant changes according to

$$\bar{K}_2^s = \bar{K}_2 + \beta\bar{K}_3 - \frac{1}{P}\left(\bar{K}_2 + \beta^2\bar{K}_2^2\right)\mathrm{e}^{\beta^2\bar{K}_2^2+\cdots} + \cdots. \tag{2.19}$$

Even when there is no selection (i. e. $\beta = 0$) the variance is reduced by $1-1/P$ because, by chance, some members of the population are selected more than others. As the selection strength β increases the variance is reduced. To obtain the full dependence of $\bar{K}_2^s$ on β one needs to use the exact equations for $G_1^\infty(\beta)$ and perform the integrals numerically. This is discussed in section 2 of Prügel-Bennett & Shapiro, 1994. The variances is also effected by the higher cumulants. The third cumulant becomes negative ($\bar{K}_2^s$ will be a monotonically decreasing function of β so its derivative, $\bar{K}_3^s$, will always be negative). This will then reduce $\bar{K}_2^s$ even more. For small β the direct effect of the ensemble covariances are small for the first few cumulants, however, even for $\beta = 0$ the fourth cumulant evolves as

$$\bar{K}_4^s\Big|_{\beta=0} = \left(1-\frac{7}{P}\right)\bar{K}_4 - \frac{6}{P}\left(V_{2,2}+\bar{K}_2^2\right) \tag{2.20}$$

thus fluctuations (here $V_{2,2}$) caused by the stochastic nature of the operators produce important corrections in the evolution of the higher cumulants for finite populations. These corrections eventually effect the estimates for the evolution of the lower cumulants because of the coupling between the cumulants. The covariances of the ensemble effect the average cumulants because selection depends non-linearly on the cumulants in the population. This is not true of mutation and sexual recombination where we obtain linear relationships between the cumulants before and after the applying the operators.

2.2 The covariances after selection

In the last section we saw that the effects of selection depends on the covariances, $V_{n,m}$, of the ensemble. These covariances arise because the evolution operators are stochastic; that is, there will be fluctuations from population to population. To calculate the covariances of the ensemble after selection we use the generating function

$$G_2^\infty(z_1,z_2) = \left\langle\!\!\left\langle \left\langle \log\left(Z(z_1)\right) \log\left(Z(z_2)\right)\right\rangle_\rho \right\rangle\!\!\right\rangle - \left\langle\!\!\left\langle \left\langle \log\left(Z(z_1)\right)\right\rangle_\rho \right\rangle\!\!\right\rangle \left\langle\!\!\left\langle \left\langle \log\left(Z(z_2)\right)\right\rangle_\rho \right\rangle\!\!\right\rangle \quad (2.21)$$

where

$$Z(z) = \sum_{\mu=1}^{P} w(F_\mu) \mathrm{e}^{zF_\mu}. \quad (2.22)$$

The covariances are given by

$$V_{n,m}^s = \left.\frac{\partial^{n+m}}{\partial z_1^n z_2^m} G_2^\infty(z_1,z_2)\right|_{z_1=z_2=0} = G_2^{\infty(n,m)}(0,0). \quad (2.23)$$

Using equation (2.10) we can derive a new identity

$$\left\langle \log(Z_1)\log(Z_2)\right\rangle_c \equiv \int_0^\infty \frac{\mathrm{d}t_1}{t_1} \int_0^\infty \frac{\mathrm{d}t_2}{t_2} \left\langle \mathrm{e}^{-t_1 Z_1} \mathrm{e}^{-t_2 Z_2}\right\rangle_c \quad (2.24)$$

(we use the subscript c to denote the "connected" correlation functions $\langle A\,B\rangle_c = \langle A\,B\rangle - \langle A\rangle\,\langle B\rangle$).

For the case of Boltzmann weights we can also use a slightly simpler generating function $G_2^\infty(\beta_1,\beta_2)$, which is again given by equation (2.21) but now with

$$Z(\beta) = \sum_{\mu=1}^{P} \mathrm{e}^{\beta F_\mu} \quad (2.25)$$

and $V_{n,m}^s = G_2^{\infty(n,m)}(\beta,\beta)$. The covariances after selection can also be computed by numerical integration. In this case it is more efficient to use a slightly modified representation. The implementational details are again discussed in appendix B

We can also obtain an expansion in $1/P$ for $V_{n,m}^s$ provided $\beta\sqrt{K_2}$ is sufficiently small. The calculation is similar to that for G_1^∞. We find

$$\begin{aligned}\hat{G}_2^\infty(\beta_1,\beta_2) \;=\;& \left\langle\!\!\left\langle \log\left(\hat{\rho}_\infty(\beta_1)\right)\log\left(\hat{\rho}_\infty(\beta_2)\right)\right\rangle\!\!\right\rangle_c + \frac{1}{P}\left\langle\!\!\left\langle \frac{\hat{\rho}_\infty(\beta_1+\beta_2)}{\hat{\rho}_\infty(\beta_1)\,\hat{\rho}_\infty(\beta_2)}\right\rangle\!\!\right\rangle \\ & -\frac{1}{2P}\left(\left\langle\!\!\left\langle \log\left(\hat{\rho}_\infty(\beta_1)\right)\frac{\hat{\rho}_\infty(2\beta_2)}{\hat{\rho}^2_\infty(\beta_2)}\right\rangle\!\!\right\rangle_c + \left\langle\!\!\left\langle \log\left(\hat{\rho}_\infty(\beta_2)\right)\frac{\hat{\rho}_\infty(2\beta_1)}{\hat{\rho}^2_\infty(\beta_1)}\right\rangle\!\!\right\rangle_c\right).\end{aligned} \quad (2.26)$$

(Where we use the notation $\hat{G}_2^\infty$ to denote the approximation to order $1/P$ of G_2^∞.) Assuming the fluctuations are of the form given in equation (2.16) we find

$$\left\langle\!\!\left\langle \log\left(\hat{\rho}_\infty(\beta_1)\right)\log\left(\hat{\rho}_\infty(\beta_2)\right)\right\rangle\!\!\right\rangle_c = \sum_{n,m=1}^{n_c} \frac{\beta_1^n\,\beta_2^m}{n!\,m!} V_{n,m} \quad (2.27)$$

while

$$\left\langle\!\!\left\langle \frac{\hat{\rho}_\infty(\beta_1+\beta_2)}{\hat{\rho}_\infty(\beta_1)\hat{\rho}_\infty(\beta_2)} \right\rangle\!\!\right\rangle = \exp\left(\sum_{n=2}^{n_c} c_n(\beta_1,\beta_2)\bar{K}_n + \frac{1}{2}\sum_{m,n=2}^{n_c} c_m(\beta_1,\beta_2)c_n(\beta_1,\beta_2)V_{n,m}\right) \quad (2.28)$$

where $c_n(\beta_1,\beta_2) = ((\beta_1+\beta_2)^n - \beta_1^n - \beta_2^n)/n!$ and finally

$$\left\langle\!\!\left\langle \log\left(\hat{\rho}_\infty(\beta_1)\right)\frac{\hat{\rho}_\infty(2\beta_2)}{\hat{\rho}^2_\infty(\beta_2)} \right\rangle\!\!\right\rangle_c = \quad (2.29)$$

$$\sum_{n=1}^{n_c}\sum_{m=2}^{n_c}\frac{\beta_1^n}{n!}b_m(\beta_2)V_{n,m}\exp\left(\sum_{m=2}^{n_c} b_m(\beta_2)\bar{K}_m + \frac{1}{2}\sum_{n,m=2}^{n_c} b_m(\beta_2)\,b_n(\beta_2)\,V_{n,m}\right).$$

The first three covariances are

$$V^s_{1,1} = V_{1,1} + \beta V_{1,2} + \cdots + \frac{1}{P}\left(\bar{K}_2 + \cdots\right) \quad (2.30)$$

$$V^s_{1,2} = V_{1,2} + \beta V_{2,2} + \cdots + \frac{1}{P}\left(\bar{K}_3 + \cdots\right) \quad (2.31)$$

$$V^s_{2,2} = V_{2,2} + \cdots + \frac{1}{P}\left(\bar{K}_4 + 2\bar{K}_2^2 + \cdots\right). \quad (2.32)$$

These together with the equations for the average cumulants after selection provide a set of equations relating the ensemble variables before and after selection. Note that the covariances $V_{1,n}$ effect only themselves but do not influence the evolution of the other ensemble variables. Nevertheless as they are quantities of interest we keep them as ensemble variables.

3 Mutation

We consider a mutation mechanism in which each gene is mutated with a probability γ

$$S_i^\mu \to S_i'^\mu = \begin{cases} -S_i^\mu & \text{with probability } \gamma \\ S_i^\mu & \text{with probability } 1-\gamma. \end{cases} \quad (3.1)$$

Averaging over the two possible outcomes we find

$$\left\langle S_i'^\mu \right\rangle_{mut} = \Gamma S_i^\mu, \qquad \Gamma = 1 - 2\gamma. \quad (3.2)$$

Using this result it is straightforward to calculate what happens to the average cumulants due to the mutation operator. The fitness after mutation of a sequences $\vec{S}^\mu$ is on average

$$\left\langle F'_\mu \right\rangle_{mut} = \left\langle \sum_{i=1}^N S_i'^\mu \right\rangle_{mut} = \Gamma\sum_{i=1}^N S_i^\mu = \Gamma F_\mu. \quad (3.3)$$

The first cumulant of an infinite population distribution is just the first cumulant itself,

$$K_1 = \left\langle F_\mu \right\rangle_\mu$$

where the average is over every member in the population. Thus

$$\bar{K}_1^m = \left\langle\!\!\left\langle \left\langle \left\langle F'_\mu \right\rangle_{mut} \right\rangle_\mu \right\rangle\!\!\right\rangle = \Gamma \left\langle\!\left\langle K_1 \right\rangle\!\right\rangle = \Gamma \bar{K}_1 \tag{3.4}$$

where we have commuted the average over the members of the population and the average over all possible mutations. The higher cumulants are only a little more complicated

$$\begin{aligned} \left\langle (F'_\mu)^2 \right\rangle_{mut} &= \left\langle \sum_{i\neq j} S_i'^\mu S_j'^\mu + \sum_i (S_i'^\mu)^2 \right\rangle_{mut} \\ &= \Gamma^2 \sum_{i\neq j} S_i^\mu\, S_j^\mu + \sum_i 1 = \Gamma^2\, F_\mu^2 + (1-\Gamma^2)\, N \end{aligned} \tag{3.5}$$

where we have used $(S_i'^\mu)^2 = 1$. The second cumulant for an infinite population is

$$K_2 = \left\langle F_\mu^2 \right\rangle_\mu - \left\langle F_\mu \right\rangle_\mu^2 .$$

Using this result together with equations (3.3) and (3.5) we find

$$\bar{K}_2^m = \Gamma^2\, \bar{K}_2 + (1-\Gamma^2)\, N \tag{3.6}$$

The higher cumulants are

$$\bar{K}_3^m = \Gamma^3\, \bar{K}_3 - 2\,\Gamma\,(1-\Gamma^2)\, \bar{K}_1 \tag{3.7}$$

$$\bar{K}_4^m = \Gamma^4\, \bar{K}_4 - 8\,\Gamma^2\,(1-\Gamma^2)\, \bar{K}_2 - 2\,(1-\Gamma^2)\,(1-3\,\Gamma^2)\, N \tag{3.8}$$

$$\bar{K}_5^m = \Gamma^5\, \bar{K}_5 - 20\,\Gamma^3\,(1-\Gamma^2)\, \bar{K}_3 + 8\Gamma\,(1-\Gamma^2)(2-3\,\Gamma^2)\, \bar{K}_1 \tag{3.9}$$

$$\begin{aligned} \bar{K}_6^m = {} & \Gamma^6\, \bar{K}_6 - 40\,\Gamma^4\,(1-\Gamma^2)\, \bar{K}_4 + 8\Gamma\,(1-\Gamma^2)(17-23\,\Gamma^2)\, \bar{K}_2 \\ & +8\,(1-\Gamma^2)(2-15\,\Gamma^2+15\,\Gamma^4) N \end{aligned} \tag{3.10}$$

The covariances can be found similarly. $V_{1,1}$ is defined as

$$V_{1,1} = \left\langle\!\!\left\langle K_1^2 \right\rangle\!\!\right\rangle - \left\langle\!\!\left\langle K_1 \right\rangle\!\!\right\rangle^2 = \left\langle\!\!\left\langle \frac{1}{P} \sum_{\mu=1}^{P} F_\mu^2 \right\rangle\!\!\right\rangle - \bar{K}_1^2 \tag{3.11}$$

so that

$$V_{1,1}^m = \Gamma^2 V_{1,1} \tag{3.12}$$

and similarly $V_{n,m}^m = \Gamma^{n+m}\, V_{n,m}$. This simple result arises because for mutation on an infinite population there are no fluctuations. If we had used the cumulants κ_n for the finite population distribution as ensemble variables then we would have found additional terms contributing to $v_{n,m}$ as we would expect.

These equations are exact for all mutation rates. We see that, for this problem, selection only depends on the cumulants and not on the details of the sequence. This is not typical, but we might have anticipated that this would be the case here as the change in fitness due to mutation depends only on the fitness of the gene sequence and not on the sequence itself.

4 Modelling the Full Evolution

4.1 Solving the full dynamics

We assume that our initial population consists of random sequences such that each gene is ± 1 with equal probability. Averaging over the two possible alleles we could choose we find that $\langle S_i^\mu \rangle = 0$ (but, of course, $(S_i^\mu)^2 = 1$) from which it is straightforward to calculate the average cumulants for the initial population. All the odd cumulants are zero while

$$\bar{K}_2(0) = N, \qquad \bar{K}_4(0) = -2N, \qquad \bar{K}_6(0) = 16N, \tag{4.1}$$

etc. The initial covariances, $V_{n,m}(0)$, are all zero. (The initial cumulants and covariance can also be obtained by taking $\gamma = 1/2$ ($\Gamma = 0$) in the mutation equations.)

To model the full dynamics we apply the evolution operators to the current set of ensemble variables to obtain a new set. Applying first the selection operator and then the mutation operator we can find how the ensemble variables change from one generation to the next. Starting from the initial set of ensemble variables we can calculate the full evolution. In the calculation we evolve the ensemble variables $\bar{K}_n$ and $V_{n,m}$ and convert them to $\bar{\kappa}_n$ and $v_{n,m}$, using the formulae given in appendix A, to compare with the measured results form simulations. The details of the simulation are discussed in appendix C. An example of the full dynamics was given in figure 2 in the introduction.

For small mutation rates the equations we obtain are actually unstable and do not correctly predict the long time evolution of the population. This results from using a truncated set of ensemble variables. Mutation suppresses the higher cumulants and covariances by a power of $\Gamma = 1 - 2\gamma$. However, if Γ becomes very close to one this suppression is reduced and the higher cumulants and covariances become important in determining the long time dynamics. An example of the breakdown in the predictions of the evolution is given below (figures 3). This breakdown is a consequence of the structure of the equations for selection. Each cumulant is strongly coupled to the cumulant of one higher order (see equations (2.18) and (2.19)). The effect of this higher cumulant is to reduce the cumulant below it. Thus by truncating the cumulant expansion we introduce an instability into the equations. Sexual recombination significantly reduces the higher cumulants so ironically populations which undergo sexual recombination are often simpler to model.

4.2 The $N \to \infty$ limit

In statistical mechanics it is often convenient to take the limit of an infinite size system (the thermodynamic limit) as the behaviour of large systems is usually similar up to a trivial rescaling and small finite size corrections. Often the problem also simplifies in this limit. For the case of evolving population of gene sequences we can also consider the limit of long sequences, $N \to \infty$. The meaningful ensemble variable is then the rescaled first cumulant $k_1 = \bar{K}_1/N$. However, we must use a mutation rate of order 1 (rather than N) otherwise in the limit $N \to \infty$ there could be no beneficial fluctuations (mutation would reduce all the fitnesses in the population by $\Gamma = 1 - 2\gamma$). We will therefore consider mutation rates $\gamma = g/N$, where g is the average number of genes that are mutated per sequence. The change in the fitness due to mutation will be given by

$$F'_\mu = F_\mu - 2\pi_{g(1+k_1)/2} + 2\pi_{g(1-k_1)/2}. \tag{4.2}$$

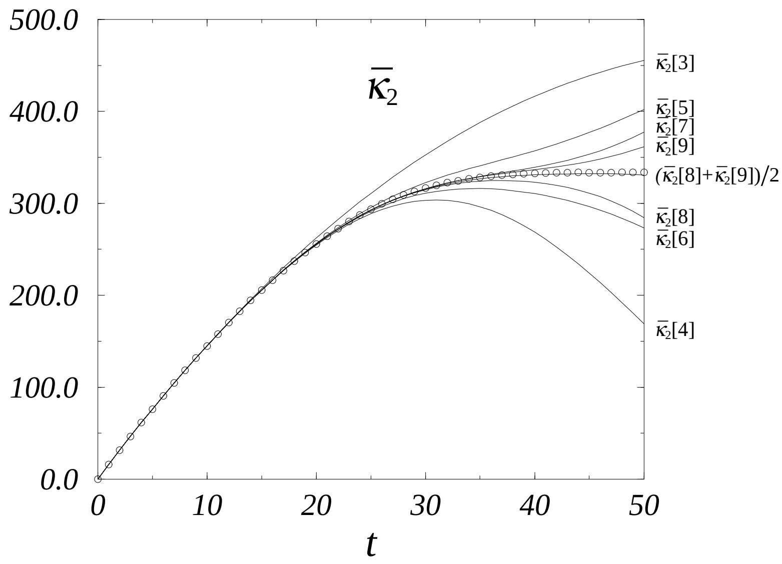

Figure 3: The curves show the predictions for $\bar{\kappa}_2$ versus the number of generations t using sets of ensemble variables with $n_c = 3$ to 9. The predictions are for Muller's ratchet ($k_1 = 1$) with $P = 100$, $\beta = 0.005$ and $\gamma = 4/N$. Also plotted is simulation data averaged over 100 000 runs ($\circ$). For clarity the error bars are not shown but are negligible on the scale of the figure.

where π_x denotes a Poisson deviate with average x. The changes in the cumulants due to mutation are thus

$$\bar{K}_n^m - \bar{K}_n = \begin{cases} -2^n g k_1, & n \text{ odd} \\ 2^n g, & n \text{ even.} \end{cases} \tag{4.3}$$

The case when $k_1 = 1$ (i. e. starting from a population of perfect gene sequences) is known as Muller's ratchet (Muller, 1964) because the average fitness of the population decreases in spite of selection. It has received considerable attention as it provides a possible explanation for the emergence of sexual reproduction and the demise of asexual mutants (Kondrashov 1988, Maynard Smith & Nee 1990 [16]). It has also inspired considerable theoretical interested (see, for example, Nowak & Schuster (1989), Charlesworth *et al* 1993 Higgs & Woodcock, 1995) although its relevance to real populations is questionable as it seems highly unlikely that the fitness of a population would ever significantly exceed its equilibrium value. The large N limit is a case where the predictions from the formalism break down (since the mutation rate per gene, γ, goes to zero). An illustration of this is given in figure 3 where we show the predicted evolution of $\bar{\kappa}_2$ for a population of size $P = 100$, with $g = 4$, $\beta = 0.005$ and $k_1 = 1$ (Muller's ratchet) for different sets of ensemble variables. We use the notation $\bar{\kappa}_2[n_c]$ to denote the prediction for $\bar{\kappa}_2$ using a set of ensemble variables which includes the first n_c cumulants and the covariances $V_{n,m}$ with $n + m \leq n_c$. The initial prediction is very good but after some time the predictions diverge away from the actual behaviour. Increasing the number of ensemble variables prolongs the time before this divergence. Taking the average of the predictions using

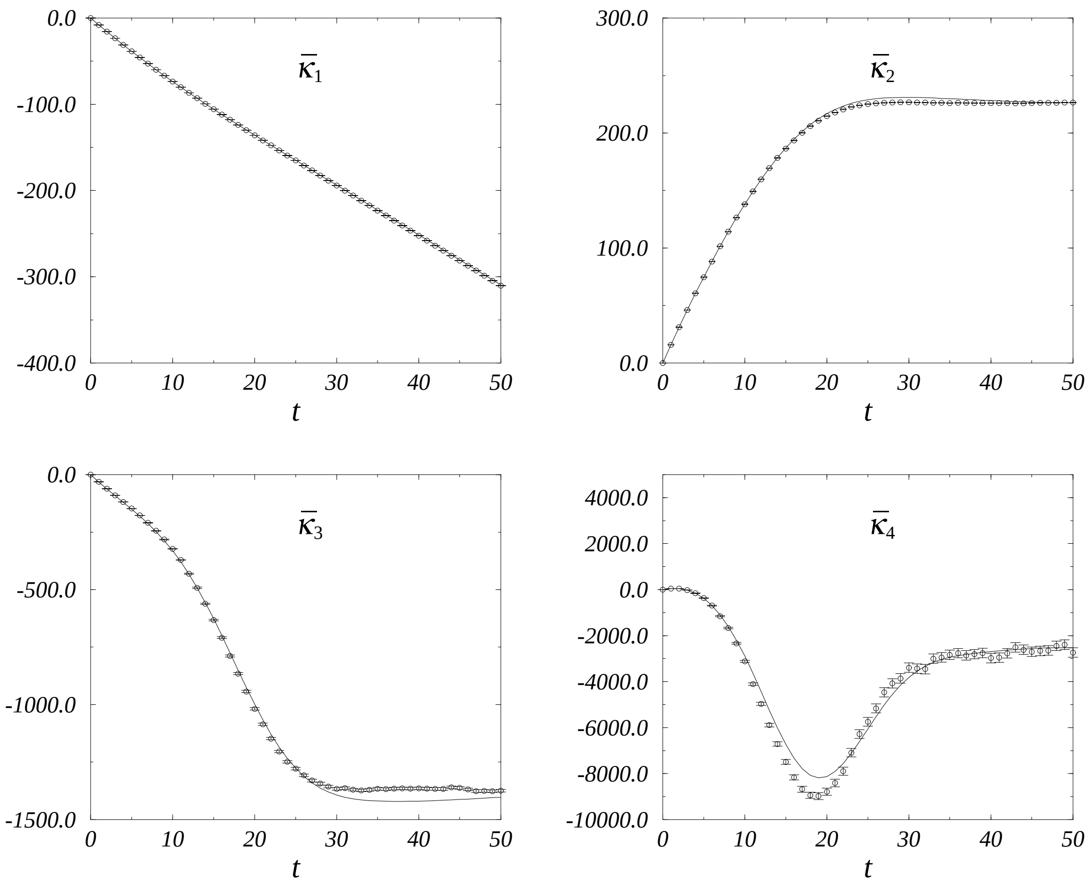

Figure 4: Examples of the predicted evolution of the ensemble variables $\bar{\kappa}_1$, $\bar{\kappa}_2$, $\bar{\kappa}_3$ and $\bar{\kappa}_4$ (solid lines) together with simulations (data points) for Muller's ratchet. In the theory fluctuations were ignored (i. e. $V_{n,m} = 0$). In the example $P = 100$ $\beta = 0.01$ and $g = 4$. The simulations were averaged over 100 000 runs. The error bars show the predicted errors in the mean.

the two largest number of ensemble variables we obtain a reasonable approximation to the true dynamics. The diverging solution is a result of using a truncated set of ensemble variables. If we ignore fluctuations altogether (i. e. by setting $V_{n,m} = 0$) we can obtain a set of dynamical equations for the evolution of the average cumulants which provide a reasonable approximation to the true dynamics. This is illustrated in figure 4 which shows the predicted evolution of the first four cumulants for Muller's ratchet with $P = 100$, $g = 4$ and $\beta = 0.01$. The first seven cumulants where used as ensemble variables. Using fewer ensemble variables gave rise to significant errors in the estimate of the dynamics. The discrepancy between the theory and simulations for the higher cumulants presumably arises because of ignoring fluctuations. Including fluctuations gives a more accurate short term prediction of the behaviour before the diverging behaviour sets in. Clearly it is desirable to find a better method for dealing with this instability.

The examples above illustrate the typical behaviour of a population of infinitely long gene sequences. After some transient period the average fitness of the population changes

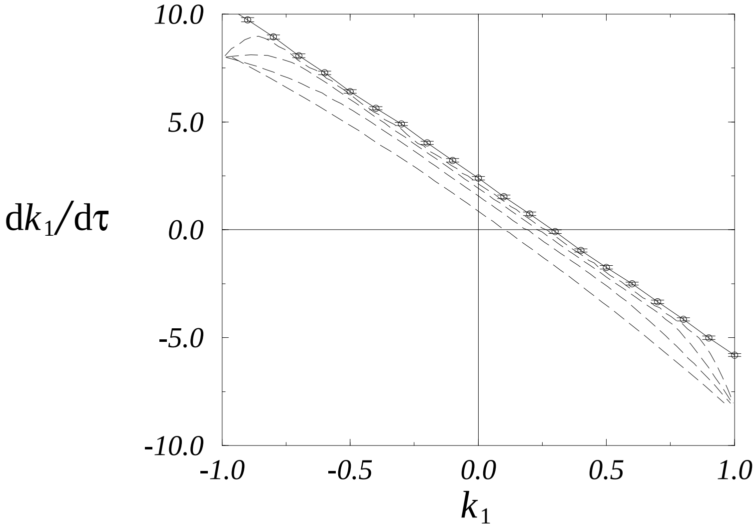

Figure 5: Plot of $\mathrm{d}k_1/\mathrm{d}\tau = 0$ versus k_1 for $P = 100$, $\beta = 0.01$ and $g = 4$ for an infinite long gene sequences (solid line). The same quantity is shown for gene sequences of length $N = 100$, 250, 500 and 1 000 (dashed lines). The results for finite gene sequences converges to the $N \to \infty$ result as N is increased. The data was obtained from simulations averaged over 10 000 runs. Error bars are shown only for the case of an infinitely long gene sequences, but are of a similar magnitude for the finite length sequences.

at some constant average rate depending on k_1 and the mutation rate. As the mutation rate is of order 1 the change in k_1 will be infinitesimal for finite time in the limit $N \to \infty$. Thus we must also rescale the time (i. e. $\tau = t/N$) in order to obtain meaningful results. A typical quantity we might be interested in is the rescaled rate of change in fitness $\mathrm{d}k_1/\mathrm{d}\tau$ which will be equal to the average rate of change of the mean fitness $\mathrm{d}\bar{K}_1/\mathrm{d}t$ after any transience (assuming the transience does not scale like N) — there are fluctuations in $\bar{K}_1$ but these only grow as $\sqrt{t}$ so $\mathrm{d}k_1/\mathrm{d}\tau$ is well defined. We illustrate this in figure 5 where we show $\mathrm{d}k_1/\mathrm{d}\tau$ versus k_1 for $P = 100$, $\beta = 0.01$ and $g = 4$, measured from simulations. The equilibrium fitness for the distribution occurs when $\mathrm{d}k_1/\mathrm{d}\tau = 0$. For the example shown this occurs around $k_1 = 0.29$. We also show $\mathrm{d}k_1/\mathrm{d}\tau$ measured for finite systems starting from an initial population with $k_1 = \pm 1$. For gene sequences consisting of around 1 000 or more genes the finite size corrections are almost negligible.

The behaviour discussed above does not describe the initial transient behaviour. A random population will typically have a variance proportional to N and the increase in the average fitness due to selection will be of order $\sqrt{N}$ (i. e. the width of the distribution of fitnesses). Thus to describe the transient behaviour in the limit of infinitely long genes would require a different scaling of the time.

A second possible limit to consider is the infinite population limit. However, the effect of selection on a finite population is to introduce additional correlations in the population. These correlations build up over time so even when the finite population corrections are small from one generation to the next they are very important in determining the long time evolution. Whether there exists a non-trivial scaling which allows this limit to be taken meaningfully remains an open question.

5 Discussion

We have presented a new formalism for modelling the dynamics of evolving populations. As an illustration of its use we have considered the classic problem of a population evolving in a multiplicative fitness landscape. We have also considered the limiting case of infinitely long gene sequences. This includes as a special case Muller's ratchet. The formalism provides, for the first time, a satisfactory model of the full dynamics in these problems. However, the formalism is not restricted to this class of problems. In the context of genetic algorithms we have already studied evolution in more complicated fitness landscapes and involving sexual recombination. Below we briefly review some of this work which is relevant to biology.

In this paper we have studied only selection and mutation operators acting on the population. It is also possible to study sexual recombination within this formalism (Prügel-Bennett & Shapiro, 1995; Rattray, 1996; Rattray, 1996a). For example, we can consider the case when we generate a population of $2P$ members and combine them in pairs to produce P "children". The simplest form of recombination we can consider is to choose, with a probability a, an allele from one parent and otherwise choose the allele from the other parent. Repeating this for each gene in the sequence. The effect of sexual recombination can be calculated using a similar calculation to mutation (section 5, Prügel-Bennett & Shapiro, 1995). The average first cumulant is unchanged by recombination while the average second cumulant becomes

$$\bar{K}_2^c = \left(a^2 + (1-a)^2\right)\bar{K}_2 + 2a\,(1-a)\,N\,(1-q) \tag{5.1}$$

where q is the average correlation between the gene sequences

$$q = \frac{1}{P(P-1)}\sum_{\mu\neq\nu} q_{\mu,\nu}, \qquad\qquad q_{\mu,\nu} = \frac{1}{N}\sum_{i=1}^{N} S_i^\mu S_i^\nu. \tag{5.2}$$

The higher cumulants depend in a more complicated way on the correlations between the gene sequences and the fitnesses of the gene sequences. The calculation for the evolution with recombination is consequently more complicated than that for just selection and mutation as we need to know additional information about the correlation of the gene sequences. This information is not available from a knowledge of the cumulants alone. We can try to estimate the correlation from a maximum entropy argument (section 4, Prügel-Bennett & Shapiro, 1995). The validity of such arguments can be checked directly by comparing against simulations. In the problems considered in Prügel-Bennett & Shapiro, 1995 the second cumulant was found to be close to $\mathcal{P}_2\,N\,(1-q)$ so the second cumulant was not significantly altered by sexual recombination, however, the higher cumulants were substantially reduced. A more direct approach is to keep the average correlation as an ensemble variable. This was carried through in Rattray, 1996 and has subsequently been improved upon (Rattray, 1996a).

In the example we discussed in this paper each gene contributed equally to the fitness. A closely related but slightly more sophisticated model is one in which each gene has a different importance

$$F_\mu = \sum_{i=1}^{N} J_i\,S_i. \tag{5.3}$$

The J_i's measure the contribution of each gene to the fitness. This model has been studied in (Prügel-Bennett & Shapiro, 1995) where it was called the random-field paramagnet. Even with this simple modification the analysis becomes more complicated because mutation depends on the configuration of the gene sequences. To solve this problem we again used a maximum entropy approach. Another related but richer model is the "spin-glass chain"

$$F_\mu = \sum_{i=1}^{N} J_i\, S_i\, S_{i+1}. \tag{5.4}$$

This can be thought of as a model in which the genes depend on one another. The efficacy of a gene now depends on its two neighbouring genes in the sequence. In this problem there are typically many local minima to single mutations. That is there are many gene sequences for which mutation of any single gene is deleterious and yet the gene sequence is not optimal. For this model the sexual recombination operator described above typically causes a large loss of fitness. A more effective form of sexual recombination is to swap all the genes after some randomly chosen point in the sequence. This causes considerably less disruption although it produces a less efficient mixing of the genes in the population and consequently reduces the higher cumulants less. This model has also been studied in (Prügel-Bennett & Shapiro, 1995), although the treatment required many approximations. We have also shown that in some classes of problems with local-minima using sexual recombination can allow the population to explore the space of gene sequences much more efficiently than mutation alone (Shapiro & Prügel-Bennett, 1996).

We believe the approach to modelling evolving populations outlined in this paper can provide a useful tool for understanding subtle biological effects. A few of these we are beginning to explore. An example of this is the role of noise in selection (Rattray, 1996b). In real populations there will be some random element determining which individuals survive. These effects are quite subtle to see in simulations, however, they are relatively easily modelled within this approach and give a new understanding of their role in evolution.

To understand qualitatively the dynamics of evolving populations we have also studied a model dynamics obtained by simplifying the equations for the full evolution (Prügel-Bennett & Shapiro, 1995). This allows the evolution to be solved in closed form and provides some further insight into how the various evolution operators work together.

The formalism presented in this paper can provide a very accurate description of evolving populations. By working in the space of macroscopic variables rather than the full phase space we can describe the ensemble using a compact probability measure. This was a crucial step in making this approach tractable. This formalism is still being developed and there remains many open questions. For example, what are the best set of ensemble variables to use. This is particularly pertinent as we have seen that using a truncated cumulant expansion leads to unstable evolution equations for low mutation rates. We have only started to apply this approach to questions in biology. There remains a large number of open questions and problems which we believe this formalism can help to address.

Acknowledgement

The formalism discussed in this paper was developed in collaboration with Jonathan Shapiro and Magnus Rattray. I would like to thank them for their continued support, suggestions and encouragement. I would also like to thank Paul Higgs and Glenn Woodcock for introducing me to the multiplicative fitness landscape and Muller's ratchet as models of biological systems.

A The Infinite Population Distribution

Given a distribution $\rho_\infty(F)$ with cumulants K_n we can ask what are the average cumulants, κ_n, and covariances $v_{n,m}$ for the ensemble of populations generated by drawing P members from the distribution $\rho_\infty(F)$ in all possible ways. These are simple to compute. Denoting the average over all possible draws by $\langle\cdots\rangle_\infty$ then the first two cumulants are

$$\kappa_1 = \left\langle \frac{1}{P}\sum_\mu F_\mu \right\rangle_\infty = \frac{1}{P}\sum_\mu \langle F_\mu\rangle_\infty = \frac{1}{P}\sum_\mu K_1 = K_1 \tag{A.1}$$

$$\begin{aligned}\kappa_2 &= \left\langle \frac{1}{P}\sum_\mu F_\mu^2 - \left(\frac{1}{P}\sum_\mu F_\mu\right)^2 \right\rangle_\infty = \left\langle \left(\frac{1}{P} - \frac{1}{P^2}\right)\sum_\mu F_\mu^2 - \frac{1}{P^2}\sum_\mu \sum_{\nu\neq\mu} F_\mu F_\nu \right\rangle_\infty \\ &= \left(1-\frac{1}{P}\right)\left(\left\langle F_\mu^2\right\rangle_\infty - \left\langle F_\mu\right\rangle_\infty^2\right) = \mathcal{P}_2 K_2\end{aligned} \tag{A.2}$$

where $\mathcal{P}_2 = (1 - 1/P)$. Note that K_2 is the familiar estimator for the variance of a distribution given P samples

$$K_2 = \frac{1}{\mathcal{P}_2}\kappa_2 = \frac{1}{P-1}\sum_{\mu=1}^{P}\left(F_\mu - \kappa_1\right)^2 \tag{A.3}$$

Following a similar procedure we find for the higher cumulants

$$\kappa_3 = \mathcal{P}_3 K_3, \qquad \kappa_4 = \mathcal{P}_4 K_4 - \frac{6\mathcal{P}_2}{P}K_2^2 \tag{A.4}$$

where $\mathcal{P}_3 = \mathcal{P}_2(1-2/P)$ and $\mathcal{P}_4 = \mathcal{P}_2(1-6/P+6/P^2)$. The covariances for the ensemble generated from $\rho_\infty(F)$ are found similarly, for example, $v_{1,1} = K_2/P$, $v_{1,2} = \mathcal{P}_2 K_3/P$ and $v_{2,2} = (2\mathcal{P}_2 K_2^2 + \mathcal{P}_2^2 K_4)/P$.

We now consider an ensemble, $P(\{K_n\})$, of distribution $\rho_\infty(F)$ with average cumulants $\bar{K}_n$ and covariances $V_{n,m}$. We wish to find the average cumulants, $\bar{\kappa}_n$, and covariances, $v_{n,m}$, for the ensemble of true distributions — these are what we would measure by averaging over many simulations. They can be found by averaging over the expressions we obtained for the single distribution. The covariances pick up a term proportional to the covariance of the infinite population ensemble. The only subtlety arises in averaging κ_4 and $v_{2,2}$ because of the K_2^2 term

$$\left\langle\!\left\langle K_2^2 \right\rangle\!\right\rangle = \bar{K}_2^2 + V_{2,2}.$$

The equations we finally arrive at are

$$\bar{\kappa}_1 = \bar{K}_1 \tag{A.5}$$

$$\bar{\kappa}_2 = \mathcal{P}_2 \bar{K}_2 \tag{A.6}$$

$$\bar{\kappa}_3 = \mathcal{P}_3 \bar{K}_3 \tag{A.7}$$

$$\bar{\kappa}_3 = \mathcal{P}_4 \bar{K}_4 - \frac{6\mathcal{P}_2}{P}\left(\bar{K}_2^2 + V_{2,2}\right) \tag{A.8}$$

$$v_{1,1} = V_{1,1} + \frac{1}{P}\bar{K}_2 \tag{A.9}$$

$$v_{1,2} = \mathcal{P}_2 V_{1,2} + \frac{\mathcal{P}_2}{P}\bar{K}_3 \tag{A.10}$$

$$v_{2,2} = \left(1 - \frac{1}{P^2}\right) V_{2,2} + \frac{1}{P}\left(2\,\mathcal{P}_2\,\bar{K}_2^2 + \mathcal{P}_2^2 \bar{K}_4\right). \tag{A.11}$$

B Calculating Selection By Numerical Integration

For large selection rates the $1/P$ expansion breaks down and we must resort to calculating the integrals numerically. In this section we outline how this can be achieved. As we must numerically average over the fluctuations the calculation becomes more labour intensive as we increase the number of ensemble variables. We will restrict our attention to the case where we keep the first four cumulants and average over fluctuations in K_1 and K_2 only.

We need to evaluate terms such as

$$\bar{K}_1^s = \left\langle\!\!\left\langle \left\langle \frac{\sum_\mu w_\mu F_\mu}{\sum_\nu w_\nu} \right\rangle_\rho \right\rangle\!\!\right\rangle. \tag{B.1}$$

Using the identity

$$\frac{1}{A^n} = \frac{1}{\Gamma(n)} \int_0^\infty t^{n-1}\, \mathrm{e}^{-t\,A}\, \mathrm{d}t. \tag{B.2}$$

(here with $n = 1$) we can write this as

$$\begin{aligned}
\left\langle \frac{\sum_\mu w_\mu F_\mu}{\sum_\nu w_\nu} \right\rangle_\rho &= \left\langle \int_0^\infty \sum_\mu w(F_\mu)\, F_\mu\, \mathrm{e}^{-t \sum_\nu w(F_\nu)}\, \mathrm{d}t \right\rangle_\rho \\
&= P \int_0^\infty \left\langle w(F)\, F\, \mathrm{e}^{-t \sum_\nu w(F)} \right\rangle_\rho \left\langle \mathrm{e}^{-t \sum_\nu w(F)} \right\rangle_\rho^{P-1} \mathrm{d}t
\end{aligned} \tag{B.3}$$

where we have used the fact that the sums over μ decouple. In the last line $\langle \cdots \rangle_\rho$ signifies the average over a single distribution $\rho_\infty(F)$, not a product of distributions. To calculate $\bar{K}_1^s$ we first perform the average over $\rho_\infty(F)$ for a given set of cumulants, $\{\bar{K}_n\}$, and a given t. We then average over t and finally average over the ensemble $P(\{\bar{K}_n\})$. To perform the average over the distribution we need to approximate it in some way. A convenient approximation to the distribution is the function

$$\rho_\infty(F) = \frac{\mathrm{e}^{-(F-K_1)^2/(2K_2)}}{\sqrt{2\pi K_2}} \left(1 + \sum_{n=3}^{n_c} \frac{K_n}{K_2^{n/2}}\, u_n\!\left(\frac{F-K_1}{\sqrt{K_2}}\right)\right) \tag{B.4}$$

where $u_n(x)$ are proportional to the Hermite polynomials. They are defined by $u_0(x) = 1$, $u_1(x) = x$ and

$$u_n(x) = \frac{1}{n}(x\,u_{n-1}(x) - u_{n-2}(x)). \tag{B.5}$$

The first few polynomials are $u_2(x) = (x^2-1)/2$, $u_3(x) = (x^3-3x)/3!$, $u_4(x) = (x^4 - 6x^2+3)/4!$. We can then perform the integral numerically. We have also tried more sophisticated forms for the distribution in which we use the distribution which maximizes the entropy

$$-\int \rho_\infty(F) \log\left(\frac{\rho_\infty(F)}{m(F)}\right) \mathrm{d}F$$

subject to the constraints on the first few cumulants and where $m(F)$ is the prior coming from the density of states. However, this distribution does not significantly improve the predictions.

Since the selection weight is invariant under a shift of the fitness origin we can make the change of variables $F \to F' = F - K_1$ to eliminate the dependence of $\bar{K}_n^s$ on K_1. The average cumulant after selection is then $K_1^s = K_1 + K_1'^s$, where $K_1'^s$ is the change in K_1 due to selection and does not depend on K_1 but only the higher cumulants. We can thus trivially average over the fluctuations in K_1. To average over fluctuations in K_2 we assume that it is distributed according to a Gamma distribution

$$P(K_2) = \frac{\mathrm{e}^{-K_2/b} K_2^{a-1}}{\Gamma(a) b^a} \tag{B.6}$$

where $a = \bar{K}_2^2/V_{2,2}$ and $b = V_{2,2}/\bar{K}_2$. This ensures that K_2 always remains positive.

The higher cumulants and covariances can be found by a similar procedure. For example,

$$V_{1,1} = \left\langle\!\!\left\langle \left\langle \left(\frac{\sum_\mu w_\mu F_\mu}{\sum_\nu w_\nu} \right)^2 \right\rangle_\rho \right\rangle\!\!\right\rangle \tag{B.7}$$

In this case we obtain a term $\left(\sum_\nu w_\nu\right)^2$ in the denominator. We can again use the identity (B.2) with $n = 2$ now to decouple the sums over different members of the population. Again we can make the change of variables $F' = F - K_1$, but now we obtain a term $\langle\!\langle K_1 K_1'^s \rangle\!\rangle$. $K_1'^s$ depends on K_2 (and the higher cumulants) while K_1 is correlated with K_2 through $V_{1,2}$. Thus we must be careful to average over the fluctuations properly. To do this we use

$$\int P(K_1|K_2)\, K_1 \; \mathrm{d}K_1 = \bar{K}_1 + \frac{V_{1,2}}{V_{2,2}} \left(K_2 - \bar{K}_2\right). \tag{B.8}$$

To calculate the effect of selection for the first few cumulants many similar integrals need to be performed. These can be done efficiently in parallel. This can be achieved simply either by using Gaussian quadrature or by treating the integrals as solutions to a set of (uncoupled) differential equations.

C Simulations

To perform numerical simulations of this problem we need only know the fitnesses of the members of the population and not the gene sequences. We initialize the fitnesses to

$N - 2\beta_{N,1/2}$ where $\beta_{n,p}$ is a binomial deviate distributed according to

$$P(\beta_{n,p} = m) = \binom{n}{m} p^m (1-p)^{n-m}. \tag{C.1}$$

Mutation is accomplished using two binomial deviates

$$F_\mu^m = F_\mu - 2\,\beta_{(N+F)/2,\gamma} + 2\,\beta_{(N-F)/2,\gamma}. \tag{C.2}$$

In the limit $N \to \infty$ the binomial deviates $\beta_{xN,\gamma}$ go over to Poisson deviates π_{gx}, where $g = N\gamma$ and

$$P(\pi_\mu = m) = \frac{\mu^m}{m!}\mathrm{e}^{-\mu}. \tag{C.3}$$

Simulations for any length of sequence is consequently very fast. The error bars given on the simulation data always show the estimated errors in the mean due to fluctuations.

References

[1] D. Charlesworth, M. T. Morgan, and B. Charlesworth. Mutation Accumulation in Finite Populations. *J. Hered.*, 84:321–325, 1993.

[2] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev.*, B24:2613–2626, 1984.

[3] P. G. Higgs and G. Woodcock. The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection. *J. Math. Biol.*, 33(7):677–702, 1995.

[4] A. S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336:435–440, 1988.

[5] Shapiro J. L., A. Prügel-Bennett, and M. Rattray. A Statistical Mechanical Formulation of the Dynamics of Genetic Algorithms. In T. C. Fogarty, editor, *Lecture notes in Computer Science* **865**, pages 17–27, Berlin, 1994. Springer-Verlag.

[6] H. J. Muller. The relation of recombination to mutational advance. *Mutation Res.*, 1:2–9, 1964.

[7] A. Nix and M. D. Vose. Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5:79–88, 1991.

[8] M. Nowak and P. Schuster. Error Thresholds of Replication in Finite Populations, Mutation Frequencies and the Onset of Muller's Ratchet. *J. theor. Biol.*, 137(4):375–395, 1989.

[9] A. Prügel-Bennett and J. L. Shapiro. An Analysis of Genetic Algorithms Using Statistical Mechanics. *Phys. Rev. Lett.*, 72(9):1305–1309, 1994.

[10] A. Prügel-Bennett and J. L. Shapiro. The Dynamics of a Genetic Algorithm for Simple Random Ising Systems. *Physica D*, 1462, 1996.

[11] M. Rattray. The dynamics of a genetic algorithm under stabilizing selection. *Complex Systems*, 9(3):213–234, 1995.

[12] M. Rattray. Noisy fitness evaluations in genetic algorithms and the dynamics of learning. Paper submitted to Foundations of Genetic Algorithms (Foga) 96, 1996.

[13] M. Rattray, A. Prügel-Bennett, and J. L. Shapiro. Selection in Genetic Algorithms. In preparation, 1996.

[14] M. Rattray and J. Shapiro. The dynamics of Genetic Algorithms for a simple learning problem problem. *J. Phys. A.*, 1996. Submitted for publication.

[15] J. L. Shapiro and A. Prügel-Bennett. When Genetic Algorithms Beats Stochastic Hill Climbing: The Basin with a Barrier. In preparation., 1996.

[16] J. Maynard Smith and S. Nee. Clicking into Decline? *Nature*, 348:391–392, 1990.

[17] M. D. Vose and G. E. Liepins. Punctuated equilibria in genetic search. *Complex Systems*, 5:31–44, 1991.

[18] G. Woodcock and P. G. Higgs. Population evolution in a single peak fitness landscape. *J. Theor. Biol.*, 176:61–73, 1996.