# MODELLING THE DYNAMICS OF GENETIC ALGORITHMS USING STATISTICAL MECHANICS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

October 1996

By
Lars Magnus Rattray
Department of Computer Science

# Contents

# List of Figures

# Abstract

A formalism for modelling the dynamics of Genetic Algorithms (GAs) using methods from statistical mechanics, originally due to Prügel-Bennett and Shapiro, is reviewed, generalized and improved upon. This formalism can be used to predict the averaged trajectory of macroscopic statistics describing the GA's population. These macroscopics are chosen to average well between runs, so that fluctuations from mean behaviour can often be neglected. Where necessary, non-trivial terms are determined by assuming maximum entropy with constraints on known macroscopics. Problems of realistic size are described in compact form and finite population effects are included, often proving to be of fundamental importance. The macroscopics used here are cumulants of an appropriate quantity within the population and the mean correlation (Hamming distance) within the population. Including the correlation as an explicit macroscopic provides a significant improvement over the original formulation.

The formalism is applied to a number of simple optimization problems in order to determine its predictive power and to gain insight into GA dynamics. Problems which are most amenable to analysis come from the class where alleles within the genotype contribute additively to the phenotype. This class can be treated with some generality, including problems with inhomogeneous contributions from each site, non-linear or noisy fitness measures, simple diploid representations and temporally varying fitness. The results can also be applied to a simple learning problem, generalization in a binary perceptron, and a limit is identified for which the optimal training batch size can be determined for this problem. The theory is compared to averaged results from a real GA in each case, showing excellent agreement if the maximum entropy principle holds. Some situations where this approximation brakes down are identified.

In order to fully test the formalism, an attempt is made on the strong NP-hard problem

of storing random patterns in a binary perceptron. Here, the relationship between the genotype and phenotype (training error) is strongly non-linear. Mutation is modelled under the assumption that perceptron configurations are typical of perceptrons with a given training error. Unfortunately, this assumption does not provide a good approximation in general. It is conjectured that perceptron configurations would have to be constrained by other statistics in order to accurately model mutation for this problem.

Issues arising from this study are discussed in conclusion and some possible areas of further research are outlined.

# Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

# Copyright Declaration

# Acknowledgements

I would like to thank Jon Shapiro for all his help and encouragement, and for being such an excellent supervisor. I would also like to thank Adam Prügel-Bennett for many helpful and enlightening discussions, and all the guys in my office who have had to endure so many of my stupid questions (especially George, whose patience I can never repay). Thanks also to the many others who have made my stay in Manchester over this period so enjoyable. I gratefully acknowledge a research studentship awarded by the Engineering and Physical Sciences Research Council (ref. 93315524).

This work is dedicated to Sarah for her love and patience, and to my family for their endless generosity and support.

# Chapter 1

# Introduction

## 1.1 The genetic algorithm

Genetic algorithms (GAs) are adaptive search techniques, which can be used to find good solutions for problems with poorly characterized and high-dimensional parameter spaces [18, 32]. They have already been successfully applied in a variety of problem domains [8] and a review of the literature shows that they are becoming increasingly popular. In the simple GA considered here, a genotype (or configuration) encodes the solution to a problem and a fitness function determines the merit of each solution by assigning a fitness value to each genotype. A population of solutions is created at random and evolves for a number of discrete generations under the action of genetic operators, analogous to the processes at work in biological populations. The most important operators are selection, where the population is improved through some form of preferential sampling, and crossover (or recombination), where genotypes are mixed, leading to non-local moves in the search space. Mutation is usually also included, producing random incremental changes to genotypes within the population. These operators are iterated sequentially until the GA is stopped, either because a solution with high enough fitness has been discovered, or because some threshold number of generations is exceeded (a more detailed description of the simple GA is provided in chapter 2, section 2.2).

This algorithm differs from traditional search heuristics, which typically make local moves around a single solution in order to sample the configuration space. For example, simulated annealing accepts moves from the current configuration to neighbouring configurations according to a probabilistic acceptance procedure such as the Metropolis algorithm [42]. Under this procedure, moves which increase fitness are always accepted, while moves which reduce fitness are accepted with some tunable probability which is reduced over time as the algorithm spends more time in configurations of higher fitness. This algorithm can be considered global if time-scales are sufficiently long for the process to equilibrate. However, time-scales of this order are often unachievable in practice and the search will become localized. In this case the usefulness of the method is determined by the local structure of the configuration space. If there are many local optima which are separated by regions of low fitness, then the algorithm will often become trapped at local optima which may be far from any global optimum.

The GA is different in two important respects. Firstly, the GA samples a population of configurations in order to determine the relative merit of each. For example, the probability of being chosen for the next generation under selection might be proportional to fitness, but would be normalized by the mean fitness over the whole population. Thus, moves between one population and the next under selection are not determined by a local sampling procedure (unless the population becomes localized around one configuration). The other important difference between the GA and more traditional search heuristics is the use of a crossover operator, which produces new configurations (offspring) by mixing existing configurations (parents). Crossover allows non-local moves within the population, because offspring may have very different configurations from either parent.

It not clear whether the non-local search taking place in the GA is an effective way to overcome the problems encountered by local search methods, although there is some empirical evidence for success [8]. It has been proposed that the GA finds good solutions to a problem through the recombination of mutually useful features from different population members. Indeed, this intuition lies behind the most influential theorem regarding GAs, Holland's Schema Theorem [32]. The Schema Theorem places emphasis on the preferential survival of building blocks which are already beneficial to solutions within the population. However, as will be seen in section 1.3.1, this theorem does not provide a sufficiently powerful formalism to explain the behaviour of GAs in general and can sometimes be misleading. In fact, there is no consensus on many theoretical and practical issues regarding GAs. For example, it is not known which problem domains are appropriate for GAs or how one should choose the search parameters in order to optimize performance. Answers to these questions are often sought through empiricism, yet this is an unsatisfactory approach as it lacks the generality required of a predictive theory.

## 1.2 Thesis goal

In order to better understand the GA and to answer quantitative questions, it is desirable to have a theoretical model. Such a model should be as simple as possible, without losing any essential features of the process under consideration. Of course, which features are essential depends on which questions are being asked. In this thesis, a theoretical formalism for modelling the dynamics of the GA using methods from statistical mechanics, originally due to Prügel-Bennett and Shapiro [53, 54], is generalized and improved upon. The formalism is used to solve the dynamics of the GA for a number of simple optimization problems, which are hopefully involved enough to provide some general insight. Problems of realistic size are described in a compact form and important finite population effects are included under the formalism. Most of the work in this thesis centres around the derivation of the equations of motion describing the dynamics of the GA, although there is also some analysis of these expressions. The aim is to review and improve upon this new theoretical formalism and to show its predictive power on a number of concrete examples.

Although this work is motivated by the wish to understand the GA as an optimization technique, it is also hoped that the formalism may be applied to related models from quantitative population genetics (see, for example, reference [12]). Where appropriate, parallels between the two fields are considered, although a thorough exposition of the quantitative genetics literature is not within the scope of this thesis.

Before describing the statistical mechanics formalism in greater detail, it is first instructive to describe some of the most influential theories from the literature on GAs.

## 1.3 Genetic algorithm theory

There has only been limited success in developing a coherent theory for explaining how GAs work, although there is a large published literature (see, for example, proceedings of the International Conference on Genetic Algorithms). The theoretical analysis of GAs is a very difficult task for a number of reasons, some of which are listed below.

- The population resides in a very high dimensional space. For example, if each genotype is a binary string of length $N$ and the population is of fixed size $P$, then the population has approximately $2^{PN}/P!$ possible realizations (assuming $P \ll 2^N$).

- The mapping from genotype to fitness will often be complex and non-linear.

- The system is dynamic and has a complex transient behaviour. The population is therefore often far from any sort of equilibrium, or steady state.
- Crossover involves the interaction of population members through mixing, while selection involves the interaction of population members through competition. The population is therefore strongly interacting and must be considered as a whole.
- Because the actual population size is usually much smaller than the space of all genotypes, infinite population approximations are often misleading. Fluctuations lead to systematic effects in a finite population.
- GAs are used in many problem domains, leading to many different types of behaviour. It is unclear how general any GA theory could be, as many features of the search will be problem specific (this is also an important issue for other search heuristics).

Some of the most significant theoretical models of the GA are described below.

### 1.3.1 Schema theorem

The most influential theorem in the GA literature is Holland's schema theorem [32]. In general, a schema is a similarity template which specifies some features of a genotype. More specifically, consider a binary genotype (a string of binary alleles),

$$1\,0\,0\,1\,1\,0\,0\,1$$

In this case the relevant schemata are hyperplane partitions. A few examples of schemata which contain this genotype as an instance are,

$$*\,*\,0\,1\,1\,0\,0\,1 \qquad *\,*\,*\,1\,*\,0\,*\,* \qquad *\,0\,0\,1\,1\,0\,0\,*$$

where the $*$ denotes a 'don't care' symbol. If $N$ is the length of the string, then there are $2^N$ possible genotypes and $3^N$ possible schemata.

The schema theorem determines a lower bound on the expected number $m(H,t+1)$ of population members which are instances of schema $H$ at generation $t+1$. In the case where the probability of selection (with replacement) is proportional to fitness one finds,

$$E\{m(H,t+1)\} \geq m(H,t)\frac{F(H,t)}{\overline{F}(t)}\,(1-p_{\mathrm{d}}(H,t)) \tag{1.1}$$

where $F(H,t)$ is the mean fitness of genotypes which are instances of schema $H$ at generation $t$ and $\overline{F}(t)$ is the mean fitness of genotypes within the population. Here, $p_{\mathrm{d}}(H,t)$ is the probability that $H$ will be disrupted by genetic operators such as crossover or mutation. The inequality appears because this expression takes no account of new instances of schemata being generated by these operators and this significantly weakens the predictive power of the theory.

The key aspect of the above inequality lies in the interplay between the disruption term and the fitness term. Consider single point crossover, in which case a crossover point along the two parent's genotypes is randomly chosen and all the alleles on one side of this point are swapped between the parents. Clearly, this operator is more likely to disrupt a schema whose distance between outermost determined alleles (defining length) is large. Under uniform crossover and mutation it is the number of determined alleles within the schema (order) that matters. Holland concludes that instances of schemata which are unlikely to be disrupted by crossover or mutation and which consistently have above average fitness within the population will increase exponentially over time. This observation is the justification for the building block hypothesis, which was stated by Goldberg:

> A genetic algorithm seeks near optimal performance through the juxtaposition of short, low-order, high performance schemata, or building blocks [18, p 41].

Unfortunately, there are a number difficulties with this interpretation (see, for example, references [14, 25]). The fitness of schemata will often change dynamically during the search and the observed average fitness of schemata may differ greatly from their expected fitness in an unbiased sample. In this case it would be meaningless to view the search as a juxtaposition of building blocks. This is especially true of problems which exhibit symmetry breaking in their dynamics. Another problem with the building block hypothesis is that there is a great deal of fitness variance between instances of the same schema. Thus, the number of samples given to a schema within the population may be too small to provide any useful information about its expected fitness within the entire search space. Grefenstette shows how the building block hypothesis can give very misleading predictions regarding problem difficulty [25].

To exactly describe GA dynamics in terms of schemata would be very difficult in general, as schemata provide a non-orthogonal and highly redundant representation of the population. Of course, there might be specific examples where a subset of schemata provide an accurate characterization. For example, in simple population genetics models the allele frequency per site is often used, which corresponds to following the frequency of all order one schemata within the population [12]. Yet this representation is an approximation if the number of sites contributing to the fitness exceeds one, because the allele frequency at each site does not completely determine the state of the population. Assuming the random assortment of alleles at each site within the population leads to incorrect results in general, even when the alleles at each site contribute equally and independently to the fitness.

### 1.3.2 Vose-Liepins formalism

An alternative theoretical approach was developed by Vose and Liepins, who provide an exact method with which to describe the GA dynamics [73, 74]. Under their formalism, the genetic operators are described by transition matrices which act on a vector describing the precise state of the population. Nix and Vose extended this formalism to include finite population effects by incorporating a Markov Chain analysis, which was necessary to describe the stochastic nature of the dynamics in this case [47].

Because this formalism is exact, it suffers from the high dimensionality problem described at the beginning of section 1.3. It is very difficult to describe problems of realistic size because of the complexity of the transition matrices and it seems that the predictive scope of the formalism may be limited by its extreme generality. Although some effort has been made to reduce the state space for particular problems by lumping similar states together, the resulting models are still computationally heavy, even for very small problems [66].

### 1.3.3 Macroscopic models

Another approach is to describe the population by a small set a macroscopic parameters under the assumption that microscopic details are not of critical importance. This is the basis of the theoretical formalism used in this thesis, as introduced in the next section, but a number of other workers have also used this idea to develop dynamical models of the GA. By ignoring detail at the configuration level, the dimensionality of the state space may be reduced to a manageable number. For example, some results have recently been derived for the performance of the GA on a class of additive problems (related to those discussed in chapter 4) [45, 67, 72]. However, these models assume a particular form of distribution which is only applicable in large populations and for very specific problems.

Often, authors do not choose appropriate quantities to average. In particular, averages are sometimes taken over a probability distribution and this is insensitive to finite population fluctuations, only giving accurate results in the infinite population limit. For example, Srinivas and Patnaik produce equations of motion for the moments of the fitness distribution in terms of the moments of the initial distribution [68]. These are moments of the average distribution and consequently the equations do not describe a finite population. Their treatment of mutation and crossover was also rather dubious, as a parameter which described the degree of disruption under each operator was chosen empirically in order to get the best fit between theory and experiment. No satisfactory explanation was given for how this parameter might be selected in general.

Macroscopic descriptions of population dynamics are also used in quantitative population genetics. Here, the importance of finite population effects are more widely appreciated and the infinite population limit is usually taken explicitly. When finite population effects are quantified for models with a large number of sites, the results are generally only exact in the limit of very weak selection [7].

## 1.4 The statistical mechanics formalism

The formalism used in this thesis was originally introduced by Prügel-Bennett and Shapiro [53, 54], and provides a theoretical model for GA dynamics using methods from statistical mechanics. This formalism falls into the class of macroscopic models described above. The population is described by a relatively small set of macroscopic order parameters and deterministic expressions are derived for the averaged trajectory of each macroscopic under the action of the genetic operators. The macroscopics are chosen so that they average well between different realizations of the dynamics and, where possible, any non-trivial terms are averaged out by maximizing entropy with constraints on known macroscopics. The macroscopics might be, for example, statistics describing the distribution of fitness or the similarity of genotypes within the population.

This approach allows an accurate description of the dynamics for a number of simple optimization problems, which are hopefully involved enough to provide some insight into how the GA searches in more general situations [54, 56, 57, 58]. The problems are of realistic size (in terms of genotype length) and finite population effects are incorporated into the model, often proving to be an essential ingredient in accurately characterizing the dynamics. This formalism requires problem independent information and is therefore less general than the Vose-Liepens formalism, yet by losing this generality it is possible to accurately predict the dynamical trajectory of the GA in interesting and non-trivial situations. The expressions describing the dynamics are compact and simple enough to analyse, leading to some novel insight into how each operator works and how to set parameters of the search.

## 1.5 Thesis outline

In this thesis the statistical mechanics formalism is extended beyond the original results due to Prügel-Bennett and Shapiro [53, 54] in order to encompass a greater range of optimization problems and describe more involved dynamical behaviour. Most of this work involves the derivation of the discrete time equations which are required to describe the dynamics of the GA for these simple, although sometimes non-trivial, optimization problems. These equations and their derivation also provides insight into the processes at work within the GA and how one might choose search parameters in order to optimize performance. However, this formalism is still being developed and the first task is to determine which problem classes can accurately be modelled. Where possible, theoretical results are compared to results from a real GA in order to justify the assumptions and approximations required by the method. A short summary of each chapter is provided below.

Chapter 2 – The statistical mechanics formalism

The statistical mechanics formalism is introduced, along with relevant definitions and notation. The mapping between genotype and fitness is divided into two stages for convenience: a mapping from genotype to phenotype and from phenotype to fitness. The macroscopics which describe the population are cumulants of the phenotype distribution and the mean correlation (a measure of genotype similarity) within the population.

Chapter 3 – Selection

The effect of selection on the distribution of phenotypes is problem independent and is therefore discussed in isolation. The selection procedure is described and a result due to Prügel-Bennett and Shapiro [52, 53] for calculating cumulants of the population after Boltzmann selection is generalized to a broader class of selection schemes.

Chapter 4 – Functions of an additive genotype

A class of problems which are particularly amenable to analysis are functions in which alleles of the genotype contribute additively to the phenotype. Results are reproduced from Prügel-Bennett and Shapiro [54] which describe the effects of crossover and mutation on phenotype cumulants, along with a maximum entropy calculation for determining non-trivial terms. The validity of the maximum entropy ansatz is tested and some limitations are identified.

As well as evolving phenotype cumulants, expressions for the change in mean correlation under each operator are derived and this provides a significant improvement over the original formulation. The theory is compared to averaged results for directional selection (one-max and the random-field paramagnet) and stabilizing selection (the subset sum problem), showing excellent predictive power as long as the maximum entropy ansatz provides a good approximation.

Chapter 5 – Noise corrupted fitness and a simple learning problem

The selection calculation is generalized to include a stochastic fitness measure. The theory is applied to a simple learning problem, generalization in a perceptron with binary weights, where there is in the fitness evaluation due to the finite size of each training batch. The dynamics is solved for this problem and the theory is compared to averaged results from a real GA, showing excellent predictive power. A limit is identified where the effects of noise can be removed by increasing the population size appropriately and this allows the optimal training batch size to be determined.

Chapter 6 – Attempting a strong NP-hard problem

The formalism is applied to the problem of storing random patterns in a perceptron with binary weights. This problem is NP-hard in the strong sense and differs from the other problems considered in this thesis because of the strongly non-linear relationship between genotype and phenotype (in this case, the training error). Mutation is modelled under the assumption that perceptron configurations within the population are typical of configurations with a given training error. Unfortunately, this assumption proves to be false in most cases and the theory does not accurately describe mutation in general. It is conjectured that perceptron configurations should be constrained by extra statistics in order to ensure more representative averaging.

Chapter 7 – Increasing biological realism: diploidy and temporally varying fitness

Diploid genotypes have previously been used in GAs for maintaining diversity within the population under a temporally varying fitness measure. In this chapter the statistical mechanics formalism is generalized to deal with a simple diploid system. The dynamics is solved for one-max with zero dominance and with a random binary dominance map (using a limiting form of crossover which completely decouples the alleles at every site). A very simple temporally varying fitness measure is also considered and the dynamics of a haploid GA are solved for this problem. This work is incomplete and a number of possible generalizations are discussed, such as diploidy with an adaptive dominance map and simple models of co-evolution.

Chapter 8 – Conclusion and outlook

In the final chapter, results and conclusions from the preceding chapters are reviewed and some promising areas of further research are considered.

# Chapter 2

# The statistical mechanics formalism

## 2.1 Introduction

Modelling the dynamics of an evolving population is made difficult by the high dimension of the space in which the population resides. Although it is possible to write down an exact master equation describing the dynamics as a Markov process, with each genetic operator included as a transition matrix [47, 73], it is difficult to make progress towards a more compact description of the dynamics without some form of simplification. A useful ansatz, often used in statistical mechanics, is to assume microscopic disorder with constraints on a small number of macroscopic quantities. A familiar example of this principle is the ideal gas, which accurately models a system of order $10^{23}$ molecules under certain conditions, yet requires the knowledge of only two macroscopic quantities (for example, the temperature and pressure) in order to fully determine a macrostate.

In its most general form, the statistical mechanics formalism models the GA as an ensemble of populations, each described by a small number of macroscopics [52]. The evolution of this ensemble provides a probabilistic description for the many possible trajectories which a single realization of the dynamics could take. The macroscopics which have proved most appropriate in the problems considered here are cumulants of some appropriate quantity within the population and the mean correlation within the population. The order parameters which describe the ensemble of populations in this case might be the mean values and covariances of these macroscopics over different realizations of the dynamics. Of course, for an exact description of the ensemble it may be that an infinite set of order parameters are required, yet in practice a truncated set often provides sufficient accuracy. This is a controlled approximation in principle, as extra order parameters may be introduced to improve accuracy.

Many of the problems considered so far under this formalism are well described by mean behaviour alone, so that the covariances of each macroscopic may be neglected. In this case the dynamics are said to self-average and this is found to be an accurate approximation for the problems under consideration in this thesis. This is typical of statistical mechanics approaches, which often focus on self-averaging quantities, but may not be a reliable assumption in general (see, for example, reference [52]), so the results presented here will always be justified by comparison with results from a real GA. Under this self-averaging assumption, the ensemble converges onto a point in the space of macroscopics which describes the mean population member. The dynamics then describes the deterministic trajectory of this point over time.

## 2.2 The simple genetic algorithm

The work in this thesis is restricted to the simple GA. The population will usually be of fixed size and evolves over a number of discrete and non-overlapping generations. The genotypes are fixed length binary strings which are randomly generated in the initial population (this is a haploid representation – diploids are considered in chapter 7). The binary variable at each site within the genotype is called an allele. This representation is convenient for the problems considered in this thesis, although it is not always appropriate. An objective function determines the fitness associated with each genotype. Each generation a number of genetic operators are applied sequentially, as described below.

**Selection**

Under selection, the population is improved by some form of preferential sampling. This can be carried out in a number of ways. In this thesis, each population member is assigned some probability of selection and a new population is selected from the old with replacement. The probability of selection will generally be some non-decreasing function of the fitness. A number of specific schemes are considered in chapter 3.

**Crossover**

Under standard crossover, the population is paired off at random and the genotypes in each pair are mixed to produce two children. The genotypes can be mixed in a number of ways and which form of crossover is most appropriate depends on the problem under consideration. If there is no spatial ordering within the genotype then it may be appropriate to use uniform crossover, in which case alleles are swapped at each site within the parents with some fixed probability. If there is spatial ordering then it may be costly to disrupt the genotype and single-point crossover might be more appropriate, in which case a crossover site is chosen at random and the string portions on one side of this site are swapped between the parents. For the problems considered in this thesis there is typically no spatial ordering and uniform crossover is used in most cases.

**Mutation**

Under mutation, alleles are randomly flipped throughout the population with some low probability. The mutation rate is sometimes reduced (annealed) over time in order to improve performance, but in this work it will remain fixed.

## 2.3 Modelling the dynamics: an overview

It is assumed that the dynamics averages sufficiently well so that only mean behaviour of the macroscopics which describe the population is required. Each genetic operator will be modelled by a set of difference equations describing the expected change in each macroscopic under that operator. This provides insight into the action of each operator and the full dynamics can be simulated by iterating the difference equations in sequence. Any terms which cannot be determined explicitly from known macroscopics may be determined by invoking a maximum entropy ansatz.

Finite population effects are found to be of great importance when characterizing GA dynamics. To model a finite population, it is assumed the the population is a finite sample taken from an infinite parent population [52]. It is most natural to follow macroscopics associated with the parent distribution from which the finite population is sampled. Selection is the only operator which involves significant finite population effects, since the other two operators do not involve sampling. It is therefore reasonable to split the dynamics into two phases: a finite population phase and an infinite population phase.

1. A finite population is randomly sampled from an infinite population.

2. Selection acts on the finite population and creates an infinite population. The proportion of each population member represented in the infinite population after selection is equal to its selection probability. Mutation and crossover are then applied to this infinite population.

These steps are iterated until the GA is stopped. This process is statistically equivalent to a standard GA acting on a finite population. Vose and Wright introduce a similar sampling procedure in reference [75], but they follow an exact microscopic description of the population rather than a small number of macroscopic statistics.

## 2.4 Definitions and conventions

### 2.4.1 Genotype → phenotype → fitness

In the problems considered here the genotype is a string of binary alleles $\{S_1, S_2 \cdots, S_N\}$ where $S_i \in \{-1, 1\}$ are Ising spins. Each population member is assigned a phenotypic value which is calculated through some deterministic function of the genotype (although the phenotype is a single number here, in general it could take a much more general form). Population member $\alpha$ has alleles $\{S_i^\alpha\}$ and phenotype $R_\alpha$. A fitness measure $F_\alpha$ will be some function of the phenotype (stochastic or deterministic),

$$F_\alpha = \mathcal{F}(R_\alpha) \qquad R_\alpha = \mathcal{R}(\{S_i^\alpha\}) \tag{2.1}$$

Fitness is not calculated directly from the genotype because it is often more convenient to follow the distribution of phenotypes within the population. For example, the phenotype might be the mean allele within the genotype in a function of unitation (phenotypes of this sort are a special case of those considered in chapter 4).

The fitness distribution is denoted $\mathcal{P}(F)$ and can be obtained from the distribution of phenotypes within the population $p(R)$ through the transformation,

$$\mathcal{P}(F) = \int \mathrm{d}R\, p(R)\, \delta(F - \mathcal{F}(R)) \tag{2.2}$$

where $\delta(x)$ is the Dirac delta function and $\mathcal{F}(R)$ is the function which assigns fitness to each phenotype. These distributions are usually only used when referring to an infinite population, which is often approximated by a continuous distribution.

### 2.4.2 Cumulants

Cumulants of the phenotype distribution within the population determine the shape of the distribution. These are very natural statistics for describing distributions which are close to Gaussian, since the higher cumulants are a measure of deviation from a Gaussian distribution. The first two cumulants are the mean and variance, while the third and fourth cumulants are related to the skewness and kurtosis respectively.

The natural logarithm of a partition function $Z$ is the generating function for each cumulant of a finite population [1],

$$\kappa_n = \lim_{\gamma\to 0} \frac{\partial^n}{\partial\gamma^n} \log Z \qquad Z = \sum_{\alpha=1}^{P} \mathrm{e}^{\gamma R_\alpha} \tag{2.3}$$

where $P$ is population size and $\kappa_n$ is the $n$th cumulant. The first two cumulants of a finite population are,

$$\begin{aligned} \kappa_1 &= \frac{1}{P}\sum_{\alpha=1}^{P} R_\alpha \; = \; \langle R_\alpha\rangle_\alpha && (2.4) \\ \kappa_2 &= \frac{1}{P}\sum_{\alpha=1}^{P}(R_\alpha)^2 - \left(\frac{1}{P}\sum_{\alpha=1}^{P} R_\alpha\right)^2 \\ &= \left(1-\frac{1}{P}\right)\left(\langle R_\alpha^2\rangle_\alpha - \langle R_\alpha R_\beta\rangle_{\alpha\neq\beta}\right) && (2.5) \end{aligned}$$

where the brackets denote population averages,

$$\langle R_\alpha\rangle_\alpha = \frac{1}{P}\sum_{\alpha=1}^{P} R_\alpha \qquad \langle R_\alpha R_\beta\rangle_{\alpha\neq\beta} = \frac{1}{P(P-1)}\sum_{\alpha=1}^{P}\sum_{\beta\neq\alpha} R_\alpha R_\beta \tag{2.6}$$

Although a finite population is being modelled, it is often more natural to describe the dynamics in terms of an infinite population from which the finite population is a random sample. Let $K_n$ be the $n$th cumulant of an infinite population. The cumulants of the infinite population phenotype distribution $p(R)$ are generated from the logarithm of a characteristic function $\rho(\gamma)$[1],

$$K_n = \lim_{\gamma\to 0} \frac{\gamma^n}{\partial\gamma^n} \log\rho(\gamma) \qquad \rho(\gamma) = \int \mathrm{d}R\, p(R)\, \mathrm{e}^{\gamma R} \tag{2.7}$$

[1] This is usually written with an explicitly imaginary argument to ensure convergence of the integral, in which case it is a Fourier transform.

The characteristic function is analogous to the finite population partition function and will often be written in terms of a cumulant expansion,

$$\rho(\gamma) = \exp\left(\sum_{n=1}^{\infty} \frac{K_n \gamma^n}{n!}\right) \tag{2.8}$$

It is well known that the variance of a finite sample is expected to be lower by a factor of $1 - 1/P$ than that of the parent distribution from which it is sampled (see equation (2.5)). Similar corrections can also be calculated for the higher cumulants. Expectation values for the first four cumulants of a finite population sampled from an infinite population were derived by Prügel-Bennett and Shapiro [54],

$$\kappa_1 = K_1 \tag{2.9a}$$

$$\kappa_2 = P_2 K_2 \tag{2.9b}$$

$$\kappa_3 = P_3 K_3 \tag{2.9c}$$

$$\kappa_4 = P_4 K_4 - 6P_2 (K_2)^2/P \tag{2.9d}$$

Here, $P_2$, $P_3$ and $P_4$ give the finite population corrections,

$$P_2 = 1 - \frac{1}{P} \qquad P_3 = 1 - \frac{3}{P} + \frac{2}{P^2} \qquad P_4 = 1 - \frac{7}{P} + \frac{12}{P^2} - \frac{6}{P^3} \tag{2.10}$$

### 2.4.3 Expanding around a Gaussian

Given a finite number of cumulants, it is sometimes necessary to construct a consistent and appropriate distribution. A convenient approximation is to expand around a Gaussian distribution using a Gram-Charlier expansion [70].

$$p(R) = \frac{1}{\sqrt{2\pi K_2}} \exp\left(\frac{-(R-K_1)^2}{2K_2}\right) \left[1 + \sum_{n=3}^{n_c} \frac{K_n}{n! K_2^{n/2}} H_n\left(\frac{R-K_1}{\sqrt{K_2}}\right)\right] \tag{2.11}$$

where $H_n(x)$ are Hermite polynomials and $n_c$ is the number of cumulants used. The Hermite polynomials are defined by,

$$H_n(x) = (-1)^n \mathrm{e}^{\frac{x^2}{2}} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left(\mathrm{e}^{-\frac{x^2}{2}}\right) \tag{2.12}$$

Four cumulants are sufficient for the problems considered in this thesis and the third and fourth Hermite polynomials are $H_3(x) = x^3 - 3x$ and $H_4(x) = x^4 - 6x^2 + 3$. The Gram-Charlier function is not a well defined probability distribution since it is not necessarily positive, but it has the correct cumulants and provides a very good approximation in many cases.

### 2.4.4 Correlation

The correlation is a measure of the microscopic similarity of genotypes and is important because selection correlates a finite population, sometimes leading to premature convergence to a poor solution. It is also important when calculating the effect of crossover, which involves the interaction of different population members. The simplest correlation measure between two population members, $\alpha$ and $\beta$, is defined as,

$$q_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N} S_i^\alpha S_i^\beta \tag{2.13}$$

Recall that $S_i \in \{-1, 1\}$, so that this quantity is positive when strings are more similar than two random strings and is negative otherwise (this is closely related to the Hamming distance). The mean correlation within the population is $q$, defined as,

$$q = \langle q_{\alpha\beta}\rangle_{\alpha\neq\beta} \;=\; \frac{1}{P(P-1)}\sum_{\alpha=1}^{P}\sum_{\beta\neq\alpha} q_{\alpha\beta} \tag{2.14}$$

## 2.5 Best population member

Although the population will be described by the mean correlation and phenotype cumulants, the aim is usually to predict the evolution of the best population member. The fitness of the best individual within the population can be formally written as (assuming it is unique),

$$F_{\text{best}} = \sum_{\alpha=1}^{P}\left(F_\alpha \prod_{\beta\neq\alpha}\Theta(F_\alpha - F_\beta)\right) \qquad \Theta(x) \;=\; \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2.15}$$

where $\Theta(x)$ is the Heaviside function. The expectation value for this quantity can be calculated if it is assumed that population members are independently sampled from an infinite parent population with phenotype distribution $p(R)$ [54]. Let $\mathcal{P}(F)$ be the fitness distribution, which is related to the phenotype distribution through equation (2.2). Then,

$$\begin{aligned} \langle F_{\text{best}}\rangle \;&=\; \int \prod_{\alpha}\big(\mathrm{d}F_\alpha \mathcal{P}(F_\alpha)\big)\sum_{\alpha=1}^{P}\left(F_\alpha \prod_{\beta\neq\alpha}\Theta(F_\alpha - F_\beta)\right) \\ &=\; P\int \mathrm{d}F\,\mathcal{P}(F)F\left(\int_{-\infty}^{F}\mathrm{d}F'\,\mathcal{P}(F')\right)^{P-1} \end{aligned} \tag{2.16}$$

Often, the best population member lies at the edges of the phenotype distribution and may not be accurately calculated when using a truncated set of cumulants to describe the distribution. Fluctuations in mean behaviour may also be large, because the higher cumulants vary substantially between different realizations of the dynamics. However, for the problems considered in this thesis, reasonable accuracy was achieved with the above expression.

In writing equation (2.16) it is assumed that population members can be considered statistically independent and can take any value of fitness from a continuum. Both these assumptions may break down under certain circumstances.

- If the population becomes highly correlated, then population members are no longer statistically independent to a good approximation. Indeed, there may be a significant probability that duplicates exist within the population. This reduces the effective size of the population and will reduce the fitness of the best population member on average. In some situations it may be possible to estimate the probability of duplicates appearing in the population, in order to amend the estimated best fitness. This is carried through in the context of a maximum entropy distribution in chapter 4, section 4.5.3.

- The discrete nature of the phenotype space may become important; for example, if the population's variance becomes comparable to the typical distance between phenotypes in state space. In this case the population can no longer be described by a small number of macroscopics and it would be necessary to characterize fine-grain features of the population. This is probably most important in problems with large numbers of degenerate genotypes, since this increases the granularity of the phenotype space.

Although the first of these issues can be corrected for in certain circumstances, in general these considerations go beyond the basic formalism presented here. In practice, the assumption that the population is accurately modelled by selecting independently from a continuous parent distribution works well until the GA is very close to convergence. If mutation is included, then this assumption is often still accurate for the whole dynamics, including the final equilibrium or steady state.

## Chapter 3

# Selection

## 3.1 Introduction

The effect of selection on the distribution of phenotypes within the population is independent of the genotype to phenotype mapping for a particular problem. This is a consequence of the fitness being a function of the phenotype only. It is therefore possible to model selection without reference to a specific problem. One might also wish to evolve the mean correlation within the population under selection, in which case one does require problem specific information. The discussion presented in this chapter is restricted to problem independent results for selection.

After introducing expressions for a general selection procedure, a number of specific schemes are considered. The first method discussed is Boltzmann selection, which is a scaled form of fitness proportional selection. This is a very natural method of selection when the fitness distribution is close to Gaussian, because it preserves the shape of a Gaussian distribution [9, 53]. Boltzmann selection is the scheme used in this thesis and is therefore considered in greatest detail. Ranking, truncation and tournament selection are also considered here, as they are often the most popular selection procedures [2, 5, 19]. By including a number of selection schemes, it is hoped that the generality of the formalism will become apparent. These methods are often preferred over the various forms of fitness proportional selection because they are rank based and are therefore insensitive to the particular choice of fitness function. This makes them less susceptible to over-selecting on highly fit individuals, which might otherwise lead to rapid and premature convergence. However, as long as the population remains relatively close to Gaussian this is not a problem for Boltzmann selection.

Many previous studies of selection model the population as a continuous and smooth distribution of phenotypes [2, 5, 19]. This is clearly an approximation in a finite population, where there are a finite number of discrete phenotypes within the population. As described in the previous chapter, it is more appropriate to consider a finite population as a random sample from an infinite parent population. The distribution of phenotypes within the infinite population will be described by a small number of cumulants, which provides a good approximation for distributions which are close to Gaussian. The number of cumulants required will depend on how much the distribution deviates from a Gaussian. Of course, when the population becomes highly correlated the assumption that population members are independently sampled from a continuous parent distribution will break down. In practice, it seems that assuming independence gives accurate results even when the population has almost fully converged.

## 3.2 A general selection procedure

There are many selection schemes available for use in GAs (see reference [5] for a recent review). Here, a general procedure is considered which can be used to describe a number of specific selection schemes. Each population member is assigned some probability of selection and a new population is selected from the old with replacement. In the case of fitness proportional selection, where each population member is selected in proportion to its fitness, this form of sampling is known as roulette wheel selection [18]. Each population member is assigned a slot in the roulette wheel whose size is proportional to the probability of selection and new population members are chosen by spinning the wheel. Less noisy forms of sampling are often used in order to try and choose as close to the desired amount of each population member as possible. Under one such method, known as universal stochastic sampling, the roulette wheel described above is divided into $P$ equal sectors and the population member whose slot lies at the edge of each sector is chosen for the next generation [3]. Such methods are more difficult to model exactly, because selection events are no longer independent.

Each population member is assigned a weight, $w_\alpha = w(F_\alpha, \{F_1, F_2, \dots, F_P\})$, which may be a function of the fitness value assigned to itself and other population members. The probability of selecting population member $\alpha$ is $p_\alpha$ and is given by,

$$p_\alpha = \frac{w_\alpha}{\sum_\alpha w_\alpha} \tag{3.1}$$

This probability is exactly the definition of fitness in biology and should not be confused with the fitness measure $F_\alpha$, which is an arbitrary function of the phenotype.

Following the discussion given by Prügel-Bennett [52], selection will be split into two stages. Firstly, $P$ population members are chosen from an infinite population at random. Secondly, an infinite population is selected from this finite population with the probability of selecting each population member given by equation (3.1). This probability is exactly the proportion of population member $\alpha$ represented in the infinite population after selection. The expected properties of a finite population after selection can be determined by again selecting $P$ population members at random. The relationship between the first four cumulants of an infinite and a finite population are given in equations (2.9a) to (2.9d).

### 3.2.1 Generating the cumulants after selection

The cumulants of the phenotype distribution within the infinite population after selection can be generated from the logarithm of a partition function,

$$K_n^{\mathrm{s}} = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \log Z_{\mathrm{s}} \qquad Z_{\mathrm{s}} = \sum_{\alpha=1}^{P} w_\alpha \mathrm{e}^{\gamma R_\alpha} \tag{3.2}$$

In order to calculate the expectation values of the cumulants after selection, one can average over the population before selection, which is randomly sampled from the infinite population phenotype distribution $p(R)$.

$$\langle \log Z_{\mathrm{s}} \rangle = \int \prod_\alpha \big( \mathrm{d}R_\alpha \, p(R_\alpha) \big) \log Z_{\mathrm{s}} \tag{3.3}$$

Following Prügel-Bennett and Shapiro, one can average over the logarithm using Derrida's trick of representing the logarithm by an integral[1] [10, 53].

$$\langle \log Z_{\mathrm{s}} \rangle = \int_0^\infty \mathrm{d}t \, \frac{\mathrm{e}^{-t} - \langle \mathrm{e}^{-tZ_{\mathrm{s}}} \rangle}{t} \tag{3.4}$$

If $w_\alpha$ is a function of $R_\alpha$ alone (through $F_\alpha$), then the average in equation (3.4) decouples and the cumulants after selection for $n > 0$ are given by,

$$K_n^{\mathrm{s}} = - \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \int_0^\infty \mathrm{d}t \, \frac{f^P(t, \gamma)}{t} \tag{3.5}$$

where,

$$f(t, \gamma) = \int \mathrm{d}R \, p(R) \exp \left( -t w(R) \, \mathrm{e}^{\gamma R} \right) \tag{3.6}$$

Here, $w(R)$ is the selection weight written as a function of the phenotype. In most cases it is necessary to compute the integrals in this expression numerically, although the integral in $t$ can be computed in closed form for binary tournament selection, discussed in section 3.4.3.

[1] To see this, notice that $\frac{1}{Z} = \int_0^\infty \mathrm{d}t \, \mathrm{e}^{-Zt}$ and integrate both sides with respect to $Z$ (as long as $Z > 0$).

### 3.2.2 Expanding around the infinite population result

It is possible to expand the cumulants after selection in $1/P$ by expanding around the infinite population result. This was done for Boltzmann selection in reference [54]. Here, the method is generalized to any selection scheme as long as the higher central moments of $w(R)$ do not diverge too rapidly (this is usually the case for relatively weak forms of selection). In this case it is possible to expand $f(t,\gamma)$, defined in equation (3.6), for small $\gamma$ (the $\gamma \to 0$ limit is relevant here) and one finds,

$$f^P(t,\gamma) \simeq \exp\bigl(-tP\psi_1(\gamma)\bigr)\left(1 + \frac{Pt^2}{2}\left(\psi_2(\gamma) - \psi_1^2(\gamma)\right)\right) \tag{3.7}$$

where,

$$\psi_n(\gamma) = \int \mathrm{d}R\, p(R)\left(w(R)\,\mathrm{e}^{\gamma R}\right)^n \tag{3.8}$$

Completing the integral in equation (3.5), one finds that the cumulants after selection up to $O(1/P)$ are given by,

$$K_n^{\mathrm{s}} = \lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\left[\log(\psi_1(\gamma)) - \frac{1}{2P}\left(\frac{\psi_2(\gamma)}{\psi_1^2(\gamma)}\right) + O\left(\frac{1}{P^2}\right)\right] \tag{3.9}$$

The leading term here is the infinite population result, which corresponds to averaging directly (an annealed average) over the partition function in equation (3.2).

### 3.2.3 Mean correlation after selection

It may be necessary to find the mean correlation after selection (see equation (2.14)). The mean correlation in an infinite population after selection is given by,

$$\begin{aligned} q_{\mathrm{s}} &= \sum_{\alpha=1}^{P} p_\alpha^2 + \sum_{\alpha=1}^{P}\sum_{\beta\neq\alpha} p_\alpha p_\beta q_{\alpha\beta} \\ &= \Delta q_{\mathrm{d}} + q_{\mathrm{nat}} \end{aligned} \tag{3.10}$$

This is also the expectation value for the correlation of a finite population after selection. The first term is due to the duplication of population members when selecting from a finite population, since the correlation of duplicates is unity. The second term is due to the natural change in correlation as the population increases in fitness. The second term depends on the relationship between genotype and phenotype and is therefore problem specific. The first term is more general and can be averaged over the distribution of phenotypes within the population as in the calculation for the cumulants after selection.

Using the definition of $p_\alpha$ in equation (3.1) one finds,

$$\begin{aligned}\langle\Delta q_{\rm d}\rangle &= \int\prod_\alpha\big(\mathrm{d}R_\alpha\, p(R_\alpha)\big)\frac{\sum_\alpha w_\alpha^2}{(\sum_\alpha w_\alpha)^2}\\ &= P\int\prod_\alpha\big(\mathrm{d}R_\alpha\, p(R_\alpha)\big)w_\alpha^2\int_0^\infty\mathrm{d}t\, t\exp\left(-t\sum_\alpha w_\alpha\right)\end{aligned} \tag{3.11}$$

The integral in $t$ provides a useful way to decouple the average for the case where $w_\alpha$ depends only on $R_\alpha$. In this case one finds,

$$\langle\Delta q_{\rm d}\rangle = P\int_0^\infty \mathrm{d}t\, t f(t) g^{P-1}(t) \tag{3.12}$$

where,

$$f(t) = \int \mathrm{d}R\, p(R) w^2(R)\exp\left(-tw(R)\right) \tag{3.13a}$$

$$g(t) = \int \mathrm{d}R\, p(R)\exp\left(-tw(R)\right) \tag{3.13b}$$

As in the cumulant calculation, the integrals in this expression often require numerical enumeration. However, as shown in section 3.2.2 it is possible to expand in $1/P$ as long as fluctuations of $w(R)$ around mean behaviour are not too large. In this case, up to $O(1/P)$ one finds,

$$\langle\Delta q_{\rm d}\rangle \simeq \frac{\psi_2(0)}{P\psi_1^2(0)} + O\left(\frac{1}{P^2}\right) \tag{3.14}$$

where $\psi_n(\gamma)$ is defined in equation (3.8).

## 3.3 Boltzmann selection

Boltzmann selection will be used in this thesis and this scheme is therefore considered in greatest detail. This is a very natural selection scheme for fitness distributions which are close to Gaussian, since it preserves the shape of a Gaussian distribution [9, 53], and it is easy to choose the selection strength so that selective pressure is invariant under addition or multiplication of a constant to the fitness. For the simple additive problems considered in chapter 4 this form of selection is also equivalent to the multiplicative landscapes often considered relevant in population genetics [11]. All the results presented in this section (except the correlation result) were originally derived by Prügel-Bennett and Shapiro in references [52, 53].

Under Boltzmann selection, the selection weight for each population member is,

$$w_\alpha = \exp(\beta F_\alpha) \tag{3.15}$$

where $\beta$ is the selection strength and determines the relative probability of selecting different population members. For zero $\beta$ each population member is selected with equal probability, while for very high $\beta$ only the fittest population member will be selected.

A variety of fitness functions are considered in this thesis, including a quadratic function of the phenotype in chapter 4 and stochastic functions of the phenotype in chapter 5. Specific expressions describing Boltzmann selection will be derived for these problems as they are introduced. In this chapter the simplest situation is considered, where fitness equals the value of the phenotype ($F_\alpha = R_\alpha$), so that selection acts directly on the phenotype distribution. Borrowing the population genetics terminology, this will be called directional selection [7].

### 3.3.1 Directional selection

Under directional selection, fitness is equal to the phenotype and the partition function for Boltzmann selection simplifies to,

$$Z_{\mathrm{s}} = \sum_{\alpha=1}^{P} \exp[(\beta+\gamma)R_\alpha] \tag{3.16}$$

Substituting this partition function into equation (3.5), one finds that the cumulants after selection are given by [53],

$$K_n^{\mathrm{s}} = -\frac{\partial^n}{\partial\beta^n}\int_0^\infty \mathrm{d}t\,\frac{f^P(t,\beta)}{t} \tag{3.17}$$

where,

$$f(t,\beta) = \int \mathrm{d}R\, p(R)\exp\left(-t\mathrm{e}^{\beta R}\right) \tag{3.18}$$

In general, the integrals in equations (3.17) and (3.18) can be computed numerically, using the Gram-Charlier expansion described in equation (2.11) to parameterize the phenotype distribution. For the simulation results presented in this work the inner integral was computed by Gauss-Hermite quadrature and the outer integral was computed by Gauss-Laguerre quadrature [51].

### 3.3.2 Weak selection expansion

The expansion described in section 3.2.2 is accurate for sufficiently small $\beta\sqrt{K_2}$ [53]. For directional Boltzmann selection $\psi_n(\gamma)$, defined in equation (3.8), is very naturally expressed in terms of phenotype cumulants (see equation (2.8)),

$$\begin{aligned}\psi_n(\gamma) &= \int \mathrm{d}R\, p(R)\, \mathrm{e}^{n(\beta+\gamma)R} \\ &= \exp\left(\sum_{i=1}^{\infty} \frac{n^i(\beta+\gamma)^i K_i}{i!}\right) \end{aligned} \tag{3.19}$$

Substituting this expression into equation (3.9), Prügel-Bennett and Shapiro show that the cumulants after selection in this limit are given by [53],

$$K_n^{\mathrm{s}} = \frac{\partial^n}{\partial\beta^n}\left[\sum_{i=1}^{\infty}\frac{\beta^i K_i}{i!} - \frac{1}{2P}\exp\left(\sum_{i=1}^{\infty}\frac{(2^i-2)\beta^i K_i}{i!}\right)\right] \tag{3.20}$$

Expanding in $\beta$ for the first few cumulants gives,

$$K_1^{\mathrm{s}} = K_1 + \beta\left(1-\frac{1}{P}\right)K_2 + \frac{\beta^2}{2}\left(1-\frac{3}{P}\right)K_3 + \cdots \tag{3.21a}$$

$$K_2^{\mathrm{s}} = \left(1-\frac{1}{P}\right)K_2 + \beta\left(1-\frac{3}{P}\right)K_3 + \frac{\beta^2}{2}\left[\left(1-\frac{7}{P}\right)K_4 - \frac{6}{P}(K_2)^2\right] \tag{3.21b}$$

$$K_3^{\mathrm{s}} = \left(1-\frac{3}{P}\right)K_3 + \beta\left[\left(1-\frac{7}{P}\right)K_4 - \frac{6}{P}(K_2)^2\right] + \cdots \tag{3.21c}$$

Notice that the variance and higher cumulants change even for zero selection strength, due to random sampling effects. In an infinite population, Boltzmann selection preserves the shape of a Gaussian distribution and higher cumulants are never introduced into the population. These expressions show that higher cumulants are introduced into a finite population sampled from a Gaussian, most noticeably the third cumulant becomes negative leading to a skewed population. This is a consequence of the fact that a finite population has sparsely populated tails, so that there is a limit to the progress which can be made by selection alone. As the skewness of the population becomes negative, equation (3.21b) shows how this accelerates the reduction in variance under further selection, which in turn slows down the increase in mean fitness.

The other genetic operators are required to reduce the magnitude of the higher cumulants and repopulate the tails of the distribution, in order that the population may make further progress under selection.

Using an appropriately rescaled selection strength $\beta = \beta'/\sqrt{\kappa_2/2\log P}$, Prügel-Bennett and Shapiro show that the reduced variance under selection from a Gaussian distribution has a shoulder at a point in the region of $\beta' \sim 0.5$ [53][2]. After this point the variance after selection drops sharply, indicating rapid convergence of the population. They suggest that the selection strength should be chosen in this region, as this achieves a large increase in mean fitness for a relatively small cost in terms of lost variance. In this work the selection strength is scaled inversely to the population's standard deviation $\beta = \beta_{\rm s}/\sqrt{\kappa_2}$ for directional selection and the finite population factor is not included explicitly, as population size is usually taken to be constant.

### 3.3.3 Increased correlation due to duplication

The increased correlation due to duplication can be calculated for directional Boltzmann selection from equation (3.12). For small $\beta$ one can again use the $1/P$ expansion. Substituting the expression for $\psi_n(\gamma)$ given in equation (3.19) into equation (3.14) one finds,

$$\Delta q_d \simeq \frac{1}{P}\left(1 + K_2\beta^2 - K_3\beta^3 + O(\beta^4)\right) \tag{3.22}$$

This shows explicitly how the negative third cumulant introduced by selection increases the correlation within the population under further selection, which results in increased convergence and reduced performance in most cases. For a full description of the effects of selection on the correlation within the population it is also necessary to consider the natural increase term in equation (3.10), which will depend on the specific problem under consideration.

### 3.3.4 Beyond mean behaviour

While following the mean behaviour of each macroscopic has proved sufficiently accurate for modelling the problems discussed in this thesis, a more general approach is to also include

[2] The weak selection approximation seems to break down in the neighbourhood of this point.

fluctuations from mean behaviour. In this way it is possible to model the GA as an ensemble of non-interacting populations, each weighted appropriately. This was carried through by Prügel-Bennett for an asexual population (no crossover) on a simple additive problem known as one-max, which is introduced in chapter 4 [52].

Fluctuations from mean behaviour were introduced by following covariances of the cumulants which described each population within the ensemble. The order parameters which described the ensemble were the mean values and covariances of each cumulant. Although the effect of fluctuations was found to be rather small, they proved to be important in asexual dynamics where the higher cumulants become important because they are not suppressed by crossover (as discussed in chapter 4, section 4.4). It seems that it is most important to include fluctuations from mean behaviour when accurate modelling of the dynamics requires the inclusion of many cumulants. In this thesis the population is usually adequately described by four cumulants and fluctuations are assumed to have a negligible effect in this case.

## 3.4 Other selection schemes

In the following three sections some popular alternative selection schemes are discussed; truncation selection, ranking selection and tournament selection. These schemes are all based on fitness rank rather than fitness value and are often preferred over fitness proportional selection schemes because they are less sensitive to the shape of distribution or particular choice of fitness measure. The phenotype distributions under consideration in this thesis are typically close to Gaussian, so this is not an important issue here and Boltzmann selection is an appropriate method.

These selection schemes have previously been described in terms of their effects on the moments or cumulants of a continuous Gaussian fitness distribution, which is effectively an infinite population approximation [5, 7]. Each method fits naturally into the finite population selection procedure outlined in section 3.2, showing the generality of this approach. In the following sections, generating functions are derived for the cumulants after directional selection. As in Boltzmann selection, the approximate expansion derived in section 3.2.2 is required to obtain a closed form result for truncation and ranking selection, while an exact closed form result is possible for binary tournament selection. The aim here is mainly to demonstrate the flexibility of the present approach and this study is by no means exhaustive or complete.

### 3.4.1 Truncation selection

Under truncation selection, the population is ranked according to fitness and a threshold rank is chosen above which all population members are equally likely to be selected and below which population members are discarded. This form of selection is also used by breeders in artificial selection and is well understood in terms of its effect on the moments of an infinite Gaussian distribution [7].

Prügel-Bennett and Shapiro consider a simplification where every population member above some threshold fitness $F_{\mathrm{t}}$ is given equal probability of selection (although they do not consider finite population corrections) [54]. This differs from a threshold rank because the fitness at a particular rank may fluctuate. Under this simplification the number of individuals which are discarded may fluctuate around some mean value. The selection weight in this case is simply,

$$w_\alpha = \Theta(F_\alpha - F_{\mathrm{t}}) \qquad \Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.23}$$

Consider directional selection ($F_\alpha = R_\alpha$). In this case the cumulants of an infinite population after selection are given by equation (3.5) with,

$$f(t,\gamma) = 1 - \int_{F_{\mathrm{t}}}^{\infty} \mathrm{d}R\, p(R) \left(1 - \exp\left(-t\mathrm{e}^{\gamma R}\right)\right) \tag{3.24}$$

It is possible to apply the expansion described in section 3.2.2 for typical population sizes, as long as $F_{\mathrm{t}}$ is not too large. The cumulants up to $O(1/P)$ after truncation selection are then given by equation (3.9),

$$K_n^{\mathrm{s}} = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \left[\log(\psi_1(\gamma)) - \frac{1}{2P}\left(\frac{\psi_2(\gamma)}{\psi_1^2(\gamma)}\right)\right]$$

where,

$$\psi_n(\gamma) = \int_{F_{\mathrm{t}}}^{\infty} \mathrm{d}R\, p(R)\, \mathrm{e}^{n\gamma R} \tag{3.25}$$

A common choice of truncation threshold is at, or close, to the mean of the distribution. For example, with $F_{\mathrm{t}} = K_1$ the mean and variance of the distribution after selection from a Gaussian distribution are,

$$\begin{aligned} K_1^{\mathrm{s}} &= K_1 + \sqrt{\frac{2K_2}{\pi}} && \text{(3.26a)} \\ K_2^{\mathrm{s}} &= \left(1 - \frac{2}{P}\right)\left(1 - \frac{2}{\pi}\right) K_2 && \text{(3.26b)} \end{aligned}$$

Notice that the finite population factor in the second cumulant is equivalent to the effect of unbiased sampling from a population of size $P/2$. This is what one might expect here, because $P/2$ is exactly the expected number of population members whose fitness is greater than the mean fitness. Selection can then be considered as unbiased sampling from these population members. As the threshold increases, the expected number of population members beyond the threshold will decrease and the $O(1/P)$ term will increase until the expansion breaks down.

Unless a very low threshold fitness is used, a truncated cumulant expansion might not describe the population after this form of selection accurately, since it may be far from Gaussian. For this reason, truncation selection is probably the least appropriate selection scheme to model using a cumulant expansion, unless crossover is disruptive enough to return the population close to Gaussian each generation.

### 3.4.2 Ranking selection

Under ranking selection, the population is ordered according to fitness and each individual is weighted according to its rank within the population. There are a number of variants of this form of selection and to simplify matters only linear ranking selection is considered here, in which case the selection weight is simply the rank of an individual added to some constant which determines the strength of selection,

$$w_\alpha = \sum_{\beta=1}^{P} \Theta(F_\alpha - F_\beta) + C \qquad \text{(3.27)}$$

The strongest possible linear ranking scheme has $C = 0$. Increasing $C$ leads to a reduced selection strength, in terms of the ratio of the weight assigned to the best and worst population members respectively. For stronger selection strength it is necessary to use some other form of ranking, such as exponential ranking [5]. Here, the case where $C = 0$ will be considered.

Unfortunately, the expression for the selection weight in equation (3.27) is difficult to analyse, as it does not allow the average in equation (3.4) to decouple in a simple way. It would be much more convenient to consider a function of $F_\alpha$ alone. Instead of assigning rank according to fitness within the population, a reasonable approximation is to assign rank according to the fitness distribution of the infinite population from which the population is a finite sample, $\mathcal{P}(F)$. In this case the selection weight is given by,

$$w_\alpha \simeq \int_{-\infty}^{F_\alpha} \mathrm{d}F\, \mathcal{P}(F) \tag{3.28}$$

This simplification was considered by Prügel-Bennett, who provided the following result[3].

For directional selection ($F_\alpha = R_\alpha$) the cumulants of an infinite population after selection are given by equation (3.5) with,

$$f(t,\gamma) = \int \mathrm{d}R\, p(R) \exp\left[-t\left(\mathrm{e}^{\gamma R}\int_{-\infty}^{R} \mathrm{d}R'\, p(R')\right)\right] \tag{3.29}$$

As in truncation selection, it is possible to apply the expansion described in section 3.2.2 for typical population sizes. The cumulants up to $O(1/P)$ after linear ranking selection are then given by equation (3.9),

$$K_n^{\mathrm{s}} = \lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\left[\log(\psi_1(\gamma)) - \frac{1}{2P}\left(\frac{\psi_2(\gamma)}{\psi_1^2(\gamma)}\right)\right]$$

where,

$$\psi_n(\gamma) = \int \mathrm{d}R\, p(R)\left(\mathrm{e}^{\gamma R}\int_{-\infty}^{R}\mathrm{d}R'\, p(R')\right)^n \tag{3.30}$$

In general, the expressions for the cumulants after selection are rather complex and require numerical enumeration. For the first two cumulants after selection from a Gaussian distribution one finds,

$$K_1^{\mathrm{s}} = K_1 + \sqrt{\frac{K_2}{\pi}}\left(1 - \frac{2}{3P}\right) \tag{3.31a}$$

$$K_2^{\mathrm{s}} = \left(1 - \frac{1}{\pi} - \frac{0.795199}{P}\right)K_2 \tag{3.31b}$$

[3] private communication.

### 3.4.3 Tournament selection

Under tournament selection, small groups of population members compete to decide which will be selected for the new population. This may be useful, as it allows selection to be executed in parallel and does not require sorting or normalization of the population. Typically, a small number of population members are drawn at random from the population and the fittest individual among them is selected for the new population. This process is repeated until a new population has been selected. Binary tournament selection will be considered, although the method presented here may easily be generalized to larger sized tournaments which would lead to stronger selection. It is also possible to introduce noise into the tournament, so that the winner is assigned a higher probability of selection than the loser, leading to weaker selection. Any such generalization can be considered under the procedure presented here.

In order to make the calculation straightforward, $2P$ independent population members are present before selection. In practice, this can be achieved to a good approximation by doubling the population size before crossover and this leads to a slight increase in correlation, as described in section 3.5. The population is then paired off at random and the individuals in each pair, or tournament, are assigned indices $\alpha$ and $\alpha + P$ respectively. The selection weights for population members $\alpha$ and $\alpha + P$ are complementary,

$$w_\alpha = 1 - w_{\alpha+P} = \Theta(F_\alpha - F_{\alpha+P}) \tag{3.32}$$

In this case, the partition function for selection is (see equation (3.2)),

$$Z_{\rm s} = \sum_{\alpha=1}^{P} \left(\Theta(F_\alpha - F_{\alpha+P})\mathrm{e}^{\gamma R_\alpha} + \Theta(F_{\alpha+P} - F_\alpha)\mathrm{e}^{\gamma R_{\alpha+P}}\right) \tag{3.33}$$

Averaging over $2P$ population members in equation (3.4) leads to the familiar form of generating function for the cumulants of an infinite population after selection. For directional selection ($F_\alpha = R_\alpha$) this is given by equation (3.5) with,

$$f(t,\gamma) = 2\int \mathrm{d}R\, p(R) \left(\mathrm{e}^{-t\mathrm{e}^{\gamma R}} \int_{-\infty}^{R} \mathrm{d}R'\, p(R')\right) \tag{3.34}$$

Unlike the previous selection calculations, for this form of selection the cumulants after selection can be determined exactly, in closed form. In fact, finite population corrections are the same as for flat selection, since each of the $P$ tournament winners has exactly equal probability of being selected. Therefore, the cumulants after selection are given by (for $n < 4$),

$$K_n^{\rm s} = P_n \left( \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \log(\psi_1(\gamma)) \right) \tag{3.35}$$

where $P_n$ is the finite population correction to the $n$th cumulant of a finite sample and $\psi_n(\gamma)$ is the same as for ranking selection, as defined in equation (3.30). Here, $P_1 = 1$ and $P_n$ is given in equation (2.10) for $n = 2$ and $n = 3$. The fourth cumulant has finite population corrections analogous to those in equation (2.9d).

In the infinite population limit this selection scheme is equivalent to linear ranking, which was discussed in the previous section. The first two cumulants after selection from a Gaussian distribution are,

$$K_1^{\rm s} = K_1 + \sqrt{\frac{K_2}{\pi}} \tag{3.36a}$$

$$K_2^{\rm s} = \left(1 - \frac{1}{P}\right)\left(1 - \frac{1}{\pi}\right) K_2 \tag{3.36b}$$

Comparing this with equations (3.31a) and (3.31b) it is clear that there are small differences in the two selection schemes due to finite population effects.

Since linear ranking and binary tournament selection differ only in finite population terms, it is interesting to ask which scheme is the most effective. One measure of effectiveness is to consider how the correlation increases under selection, since an excessive increase in correlation may lead to premature convergence and reduced performance [18]. Under tournament selection, the duplication term defined in equation (3.10) is always equal to $1/P$, which is the duplication contribution to the correlation expected under flat selection on $P$ individuals. This will always be less than the duplication term for linear ranking in a population of size $P$, where fitter population members are always more likely than average to be duplicated under selection. However, this is a misleading comparison as the population size is taken to be $2P$ before tournament selection (two population members in each tournament). It is then more appropriate to consider linear ranking where population members are selected from a population of size $2P$, which leads to reduced sampling errors. In this case binary tournament selection gives a slightly higher correlation due to duplication. In practice, it is unlikely that there will be much difference between the methods. Tournament selection is often the preferred method because it does not require sorting of the population and is easy to execute in parallel.

## 3.5 Reducing the sampling error

The selection procedure described in section 3.2 uses roulette wheel sampling, with population members selected independently with replacement for the new population. This is a rather noisy form of selection and other less noisy forms are often preferred in practice. One common method is stochastic universal sampling which was described at the beginning of section 3.2 [3]. Under this method, the number of each individual selected for the new population is as close to the desired proportion as possible. This is a difficult form of sampling to model exactly in general, as the selection of each individual is no longer an independent event.

One selection scheme in which the two different forms of sampling can be compared is in tournament selection, in this case binary tournament selection. A population of size $2P$ is required after selection, which can then undergo mutation and crossover before being divided into $P$ tournaments for further selection. This ensures high probability that duplicates do not appear within the same tournaments. The procedure which corresponds to stochastic universal sampling is to select exactly two of each tournament winner in the population after selection. Roulette wheel sampling corresponds to selecting $2P$ randomly from an infinite pool containing equal proportions of each tournament winner. It is simple to calculate the increase in correlation under both forms of sampling.

Let $q$ be the mean correlation between different tournament winners after selection. The correlation in an infinite population of tournament winners is $q + (1 - q)/P$, as there is a $1/P$ probability of two distinct population members being identical. This is also the expected correlation in a finite random sample of size $2P$ created by roulette wheel sampling. Now consider a population of size $2P$ which contains exactly two representatives of each tournament winner, as produced by stochastic universal sampling. The expected mean correlation in this population would be $q + (1 - q)/4P$ for large P, as there is now a $P/2P(2P - 1)$ probability of two population members being identical. Therefore, the population correlates four times as much through random duplication by using roulette wheel selection.

This result indicates that roulette wheel sampling is certainly an inefficient form of sampling and correlations may grow much less quickly under other forms of sampling. Although one might expect both forms of sampling to act similarly for strong selection, it is certainly the case that as the selection strength reduces to zero they behave very differently. This might be important when making theoretical predictions for weak selection behaviour, as there is much greater loss of diversity (genetic drift) under roulette wheel sampling than expected under stochastic universal sampling. Analysis of the asymptotic behaviour of different sampling schemes in the limit of weak selection would be very useful, but this is probably a very difficult task in general.

## 3.6 Conclusion

A general selection procedure was defined and a generating function was introduced for calculating the change in phenotype cumulants under a class of selection schemes. This work generalizes upon the results of Prügel-Bennett and Shapiro in order to cover a greater range of selection schemes and to calculate the increased correlation due to the duplication under selection [52, 53]. In contrast to other approaches, finite population effects are included explicitly under this formalism, leading to a better characterization of selection and a number of interesting observations. In general, numerical enumeration is required to generate the cumulants after selection, although it was shown how one could expand around the infinite population result for weak selection, allowing closed form results for Boltzmann, truncation and linear ranking selection in this limit. For binary tournament selection, an exact closed form result was possible in the general case. Further work is required to determine the range of applicability of the weak selection approximation and, if possible, to characterize selection under different sampling procedures.

Boltzmann selection was considered in greatest detail here, as this is the selection method of choice in the rest of this thesis. The directional selection results due to Prügel-Bennett and Shapiro were reviewed, and a calculation for the increased correlation due to duplication was also included [52, 53]. Finite population effects lead to an increase in the magnitude of higher cumulants under directional selection, resulting in a loss of variance under further selection and a faster accumulation of correlations due to duplication. These effects cannot be seen in the limit of an infinite population, emphasizing how important it is to accurately characterize finite population effects.

# Chapter 4

# Functions of an additive genotype

## 4.1 Introduction

A class of problems which are particularly amenable to analysis are functions in which alleles of the genotype contribute additively to the phenotype. These functions include a number of problems which have been discussed at length within the literature, yet the statistical mechanics formalism is the first method which accurately characterizes the dynamics in general, including inhomogeneous contributions from each site, finite population effects and non-linear fitness functions, all of which are considered in this chapter. In chapters 5 and 7 these methods are also used to model the dynamics for a simple learning problem, a diploid GA and a simple temporally varying problem.

This work was initiated by Prügel-Bennett and Shapiro, who applied the statistical mechanics formalism to two closely related problems – the random-field paramagnet and the spin-glass chain [54]. Many of the results presented here were initially derived in their analysis, although in order to achieve greater accuracy it has been necessary to follow the evolution of an extra macroscopic, the mean correlation within the population, which was defined in equation (2.14). Expressions describing the evolution of the mean correlation provide the most significant new results derived in this chapter and the explicit inclusion of this macroscopic increases the accuracy and generality of the method.

In the following sections a general form for the phenotype is defined and expressions for the effects of mutation and crossover on each macroscopic are introduced. These expressions are independent of the particular form which the fitness function takes, since mutation and crossover only affect the phenotype through the genotype and do not act on the fitness directly. In order to determine terms not explicitly related to known macroscopics, a maximum entropy ansatz due to Prügel-Bennett and Shapiro is used [54]. This ansatz is also required to determine the increased correlation under selection.

The formalism is applied to a number of fitness functions, leading to solutions for the dynamics under directional selection (one-max and the random-field paramagnet) and stabilizing selection (the subset sum problem). In each case the mean evolution of the macroscopics and best population member are accurately determined, as long as the maximum entropy ansatz is justified. In some cases the ansatz does not hold and there are systematic errors in the theoretical predictions.

## 4.2 The phenotype

The phenotype of population member $\alpha$ is defined,

$$R_\alpha = \sum_{i=1}^{N} J_i S_i^\alpha \tag{4.1}$$

Here, the $J_i$ are fixed weights at each site which are chosen from some arbitrary distribution. The allele at site $i$ in population member $\alpha$ is an Ising spin $S_i^\alpha \in \{-1, 1\}$.

The cumulants of the phenotype distribution are defined in chapter 2, section 2.4.2 and for this phenotype the first two cumulants of an infinite population are,

$$K_1 = \sum_{i=1}^{N} J_i \langle S_i^\alpha \rangle_\alpha \tag{4.2a}$$

$$\begin{aligned} K_2 &= \left\langle \left( \sum_{i=1}^{N} J_i S_i^\alpha \right)^2 \right\rangle_\alpha - \left( \sum_{i=1}^{N} J_i \langle S_i^\alpha \rangle_\alpha \right)^2 \\ &= \sum_{i=1}^{N} J_i^2 (1 - \langle S_i^\alpha \rangle_\alpha^2) + \sum_{i=1}^{N} \sum_{j \neq i} J_i J_j (\langle S_i^\alpha S_j^\alpha \rangle_\alpha - \langle S_i^\alpha \rangle_\alpha \langle S_j^\alpha \rangle_\alpha) \end{aligned} \tag{4.2b}$$

The angled brackets denote population averages as defined in equation (2.6). The expectation value for the cumulants of a finite population sampled from an infinite population can be found from equations (2.9a) to (2.9d) for the first four cumulants.

The initial population is randomly generated, with each allele chosen uniformly from $\{-1, 1\}$. In this case the mean correlation and odd cumulants of such a distribution are zero, while the first two even cumulants of an infinite, random population are,

$$K_2^{\mathrm{i}} = \sum_{i=1}^{N} J_i^2 \qquad K_4^{\mathrm{i}} = -2 \sum_{i=1}^{N} J_i^4 \tag{4.3}$$

## 4.3 Mutation

Under mutation, each allele within the population is flipped with probability $p_{\mathrm{m}}$. Introducing an independent binary variable for each allele within the population provides a natural way of describing this operator,

$$S_i^\alpha \to M_i^\alpha S_i^\alpha \qquad M_i^\alpha = \begin{cases} 1 & \text{with probability } 1-p_{\mathrm{m}} \\ -1 & \text{with probability } p_{\mathrm{m}} \end{cases} \tag{4.4}$$

So, for example, the first cumulant after mutation is,

$$K_1^{\mathrm{m}} = \sum_{i=1}^{N} J_i \langle M_i^\alpha S_i^\alpha \rangle_\alpha \tag{4.5}$$

Averaging over all mutations gives the expectation value for the first cumulant after mutation,

$$\langle K_1^{\mathrm{m}} \rangle_{\mathrm{mut}} = \Gamma \sum_{i=1}^{N} J_i \langle S_i^\alpha \rangle_\alpha = \Gamma K_1 \tag{4.6a}$$

where $\Gamma = 1 - 2p_{\mathrm{m}}$. This calculation can be generalized to the higher cumulants and Prügel-Bennett and Shapiro determine expectation values for the first four cumulants after mutation [54] [1],

$$\begin{aligned} K_2^{\mathrm{m}} &= \Gamma^2 K_2 + (1-\Gamma^2)\sum_{i=1}^{N} J_i^2 && \text{(4.6b)} \\ K_3^{\mathrm{m}} &= \Gamma^3 K_3 - 2\Gamma(1-\Gamma^2)\sum_{i=1}^{N} J_i^3 \langle S_i^\alpha \rangle_\alpha && \text{(4.6c)} \\ K_4^{\mathrm{m}} &= \Gamma^4 K_4 - 2(1-4\Gamma^2+3\Gamma^4)\sum_{i=1}^{N} J_i^4 && \text{(4.6d)} \\ &\quad -8\Gamma^2(1-\Gamma^2)\left[\sum_{i=1}^{N} J_i^4 (1 - \langle S_i^\alpha \rangle_\alpha^2) + \sum_{i=1}^{N}\sum_{j\neq i} J_i^3 J_j (\langle S_i^\alpha S_j^\alpha \rangle_\alpha - \langle S_i^\alpha \rangle_\alpha \langle S_j^\alpha \rangle_\alpha)\right] \end{aligned}$$

where these are cumulants of an infinite population. Similarly, the expected mean correlation after mutation is,

$$q_{\mathrm{m}} = \Gamma^2 q \tag{4.7}$$

A number of terms in the expressions for the third and fourth cumulants cannot be expressed in terms of the cumulants or the correlation within the population, unless the weights are equal at every site, as is the case for the one-max problem which is introduced in section 4.7. In this case the expressions for every cumulant after mutation can be written in terms of the cumulants before [52]. Otherwise, on-site terms can be calculated by assuming maximum entropy

[1] In a private communication, Nick Barton points out that Prügel-Bennett and Shapiro did not include off-site terms in the fourth cumulant.

with constraints on the mean phenotype and correlation within the population, as described in section 4.5. Unfortunately, the off-site term in the fourth cumulant cannot be determined by this method, as the maximum entropy result does not cater for off-site terms. In the problems considered here this term does not have any significant impact, however, as the effect of mutation on the higher cumulants is negligible compared to the effects of crossover (described in the next section). For our purposes, off-site terms can be neglected to a good approximation. For asexual dynamics, or very non-disruptive forms of crossover, such an approximation may not be justified.

## 4.4 Crossover

Under crossover, genetic material is exchanged between population members. This is usually carried out by pairing off the population at random, with each pair crossed to produce two children. There are many possible crossover schemes available and which is most appropriate depends on the problem under consideration, and on how the problem is encoded within the genotype [18]. For problems with strong spatial interactions between alleles it is often important to minimize disruption of the genotype. In this case single-point crossover might be most appropriate, where parent genotypes are broken at one point and the segments on one side of this point are swapped.

In the problems under consideration in this chapter there is no such spatial order and neighbouring alleles are of no more importance than spatially distant alleles. There may still be some cost involved in shuffling alleles, however, so it is often convenient to allow more or less disruption to the parent's genotypes. In this case crossover is a generalized version of uniform crossover and the alleles of a child produced by parents $\alpha$ and $\beta$ are given by,

$$S_i^{\mathrm{c}} = X_i^{\alpha\beta} S_i^{\alpha} + (1 - X_i^{\alpha\beta}) S_i^{\beta} \qquad X_i^{\alpha\beta} = \begin{cases} 1 & \text{with probability } a \\ 0 & \text{with probability } 1-a \end{cases} \tag{4.8}$$

where $a$ is the parameter which determines the relative number of alleles taken from each parent. Under uniform crossover $a = 0.5$ is a common choice, in which case alleles are taken from either parent with equal probability.

The expectation value for each cumulant after crossover can be calculated by averaging over the $X_i^{\alpha\beta}$ variables in each cumulant. Prügel-Bennett and Shapiro show that the expectation values for the first four cumulants after crossover are given by [54],

$$K_1^{\mathrm{c}} = K_1 \tag{4.9a}$$

$$K_2^{\mathrm{c}} = K_2 + 2\mathcal{A}(K_2^{\max} - K_2) \tag{4.9b}$$

$$K_3^{\mathrm{c}} = K_3 + 3\mathcal{A}(K_3^{\max} - K_3) \tag{4.9c}$$

$$K_4^{\mathrm{c}} = K_4 + 2\mathcal{A}(2 - \mathcal{A})(K_4^{\max} - K_4) \tag{4.9d}$$

Here, $\mathcal{A} = a(1-a)$ and $K_n^{\max}$ is the fixed point of the $n$th cumulant under crossover alone. This is the state where off-site averages within and between population members are equal on average; so, for example $\langle S_i^\alpha \rangle_\alpha \langle S_j^\alpha \rangle_\alpha = \langle S_i^\alpha S_j^\alpha \rangle_\alpha$ and the second term in equation (4.2b) disappears.

$$K_2^{\max} = \sum_{i=1}^{N} J_i^2 (1 - \langle S_i^\alpha \rangle_\alpha^2) \tag{4.10a}$$

$$K_3^{\max} = -2\sum_{i=1}^{N} J_i^3 (\langle S_i^\alpha \rangle_\alpha - \langle S_i^\alpha \rangle_\alpha^3) \tag{4.10b}$$

$$K_4^{\max} = -2\sum_{i=1}^{N} J_i^4 (1 - 4\langle S_i^\alpha \rangle_\alpha^2 + 3\langle S_i^\alpha \rangle_\alpha^4) \tag{4.10c}$$

These expressions can be calculated by making a maximum entropy ansatz, as described in the next section. Crossover relaxes the cumulants towards the fixed point defined by equations (4.10a) to (4.10c), often leading to a much more rapid reduction in the magnitude of the higher cumulants than could be achieved through mutation alone. In fact, for directional selection, which is discussed in section 4.7, crossover leaves the first two cumulants unchanged to a reasonable approximation and substantially reduces higher cumulants introduced by selection. This leads to much improved progress under further selection, while mutation has a relatively small effect (for practical mutation rates).

The mean correlation is unchanged by crossover, because although crossover changes the alleles within each population member, it conserves the mean number of alleles at each site within the population.

A rather extreme form of crossover is bit-simulated crossover (BSC), which is only appropriate in problems where there is a very low cost associated with crossover in terms of reduced fitness [71]. In this case it is practicable to relax the population straight to the fixed point of standard crossover. This can be achieved by selecting alleles for each site in a child from a randomly selected population member, so that the population is effectively randomized with a constraint on the mean magnetization per site (the mean allele per site within the population). If this form of crossover is used, then one can accurately describe the dynamics of problems with an additive genotype in terms of only the two macroscopics required to constrain the maximum entropy distribution. This form of crossover also allows a special limit to be developed, which facilitates a solution to the dynamics for a number of non-trivial problems. This limit is developed in section 4.9 and is applied to diploid systems and a temporally varying fitness measure in chapter 7.

## 4.5 Maximum entropy ansatz

In order to calculate terms which are not trivially related to known macroscopics, it is necessary to make some assumption about how alleles are distributed at each site. Prügel-Bennett and Shapiro have introduced a maximum entropy ansatz in order to calculate these terms [54]. They used two constraints, the mean phenotype and correlation within the population, although they did not choose the correlation as an explicit macroscopic (they estimate it from the variance). The simple correlation measure defined in equation (2.14) is used here, although it is also possible to use a different correlation measure which includes a weight factor within the sum over sites, as in reference [56]. The simpler choice of correlation measure was found to characterize the population better in the problems considered here. A comparison of the theoretical prediction with experimental results is required as *a posteriori* justification of the ansatz.

### 4.5.1 Allele distribution at maximum entropy

To estimate the non-trivial on-site terms in equations (4.6c), (4.6d) and (4.10a) to (4.10c), it is necessary to estimate how alleles are distributed at each site. This will be achieved by calculating the expected mean allele at each site in a population at maximum entropy, with constraints on the mean phenotype and correlation within the population.

Define $\tau_i$ to be the mean allele within the population (magnetization) at site $i$,

$$\tau_i = \langle S_i^\alpha \rangle_\alpha = \frac{1}{P} \sum_{\alpha=1}^{P} S_i^\alpha \tag{4.11}$$

The single-site density of states $\Omega(\tau_i)$ is the proportion of allele combinations compatible with this magnetization,

$$\Omega(\tau_i) = \frac{1}{2^P} \binom{P}{P(1+\tau_i)/2} \tag{4.12}$$

One can define an entropy $\mathcal{S}(\tau_i)$ which is the logarithm of this quantity. Using Stirling's approximation for large $P$ one finds,

$$\begin{aligned} \mathcal{S}(\tau_i) &= \log \Omega(\tau_i) \\ &\sim -\frac{P}{2} \log(1-\tau_i^2) + \frac{P\tau_i}{2} \log\left(\frac{1-\tau_i}{1+\tau_i}\right) \end{aligned} \tag{4.13}$$

Lagrange multipliers enforce constraints on the mean phenotype and correlation (these expressions are for large $P$),

$$zPK_1 = z \sum_{\alpha=1}^{P} \sum_{i=1}^{N} J_i S_i^\alpha = zP \sum_{i=1}^{N} J_i \tau_i \tag{4.14a}$$

$$\tfrac{1}{2}(xP)^2 q = \frac{x^2}{2N} \sum_{\alpha=1}^{P} \sum_{\beta=1}^{P} \sum_{i=1}^{N} S_i^\alpha S_i^\beta = \frac{(xP)^2}{2N} \sum_{i=1}^{N} \tau_i^2 \tag{4.14b}$$

A probability distribution for the $\{\tau_i\}$ configuration can then be defined which decouples at each site,

$$\mathcal{P}(\{\tau_i\}) = \prod_{i=1}^{N} p(\tau_i) = \prod_{i=1}^{N} \exp\left(\mathcal{S}(\tau_i) + zPJ_i\tau_i + \frac{(xP\tau_i)^2}{2}\right) \tag{4.15}$$

and a Gaussian integral removes the square in the exponent,

$$p(\tau_i) = \int \frac{\mathrm{d}\eta_i}{\sqrt{2\pi}} \exp\left(\frac{-\eta_i^2}{2} + PG(\tau_i, \eta_i)\right) \tag{4.16}$$

where,

$$G(\tau_i, \eta_i) = \mathcal{S}(\tau_i)/P + (zJ_i + x\eta_i)\tau_i \tag{4.17}$$

The maximal value of $G$ with respect to $\tau_i$ gives the maximum entropy distribution for $\tau_i$, in which case,

$$\tau_i = \tanh(zJ_i + x\eta_i) \tag{4.18}$$

where $\eta_i$ is drawn from a Gaussian distribution with zero mean and unit variance.

The constraints can be used to obtain values for the Lagrange multipliers,

$$\begin{aligned} K_1 &= \sum_{i=1}^{N} J_i \, \overline{\tanh(zJ_i + x\eta_i)} && \text{(4.19a)} \\ q &= \frac{1}{N}\sum_{i=1}^{N} \overline{\tanh^2(zJ_i + x\eta_i)} && \text{(4.19b)} \end{aligned}$$

where bars denote averages over the Gaussian noise $\eta_i$. The average over $J_i$ and $\eta_i$ will usually be computed numerically by Gaussian quadratures [51], depending on the particular distribution of weights. Once the Lagrange multipliers have been determined, the expressions for mutation and crossover which involve non-trivial on-site terms can be calculated,

$$\begin{aligned} \sum_{i=1}^{N} J_i^n \langle S_i^\alpha \rangle_\alpha^m &= \int \prod_i \big(\mathrm{d}J_i \, p(J_i)\big) \sum_{i=1}^{N} J_i^n \, \overline{\tanh^m(J_i z + x\eta_i)} \\ &= N \int \mathrm{d}J \, p(J) \, J^n \, \overline{\tanh^m(Jz + x\eta)} && \text{(4.20)} \end{aligned}$$

Although these averages have to be computed numerically, it should be noticed that the computation does not scale with problem size or population size.

The fixed point of an infinite population under crossover is assumed to be a maximum entropy distribution, whose cumulants and correlation may be naturally generated from a single-site partition function (this function will be useful later).

$$\begin{aligned} K_n^{\max} &= \sum_{i=1}^{N} \lim_{\gamma\to 0} \frac{\partial^n}{\partial\gamma^n} \log \mathcal{Z}_i(\gamma, 0) && \text{(4.21a)} \\ q &= \frac{1}{N}\sum_{i=1}^{N} \left(1 - \lim_{\epsilon\to 0} \frac{\partial^2}{\partial\epsilon^2} \log \mathcal{Z}_i(0,\epsilon)\right) && \text{(4.21b)} \end{aligned}$$

where,

$$\mathcal{Z}_i(\gamma,\epsilon) = \left(\frac{1+\tau_i}{2}\right) \mathrm{e}^{\gamma J_i + \epsilon} + \left(\frac{1-\tau_i}{2}\right) \mathrm{e}^{-(\gamma J_i + \epsilon)} \tag{4.22}$$

### 4.5.2 Testing the ansatz

The maximum entropy ansatz requires some justification, as it may not always provide a good approximation. Figure 4.1 shows the averaged microscopic distribution of alleles within the population for a standard GA under directional selection on an additive genotype, for two mutation rates. Snapshots were taken every 25 generations for the first 75 generations, with the mean allele and mean squared allele per site within the population shown as a function of the weight at each site. The GA with the lower mutation rate is most accurately described by the maximum entropy result, while the GA with a higher mutation rate is eventually only in qualitative agreement. This is reflected in the theoretical predictions for the dynamical trajectories which are shown in figures 4.5 and 4.6.

These results can be explained by noting that mutation takes the distribution away from maximum entropy. This phenomena is easily pictured by considering a population which mutates away from an initial population of optimal solutions. Mutation does not differentiate between high and low weights, and will flip alleles associated with high weights with a much greater probability than predicted by the maximum entropy result. When selection and mutation are combined, it is assumed that selection will redress this imbalance by rejecting the population members whose alleles have been flipped at high weights. Yet, there is no guarantee that selection will completely remove these mutations. Figure 4.1 shows how, for the higher mutation rate, the maximum entropy ansatz over-estimates the mean allele per site at the largest weights.

### 4.5.3 Probability of duplicates

The expression derived in equation (2.16) for estimating the best member of the population assumes that all population members are chosen independently from a continuous distribution of fitness. This approximation breaks down when the population becomes highly correlated and is certainly inapplicable when duplicates exist within the population. It is possible to calculate the probability of two population members being duplicates when randomly selected from the maximum entropy distribution described here.

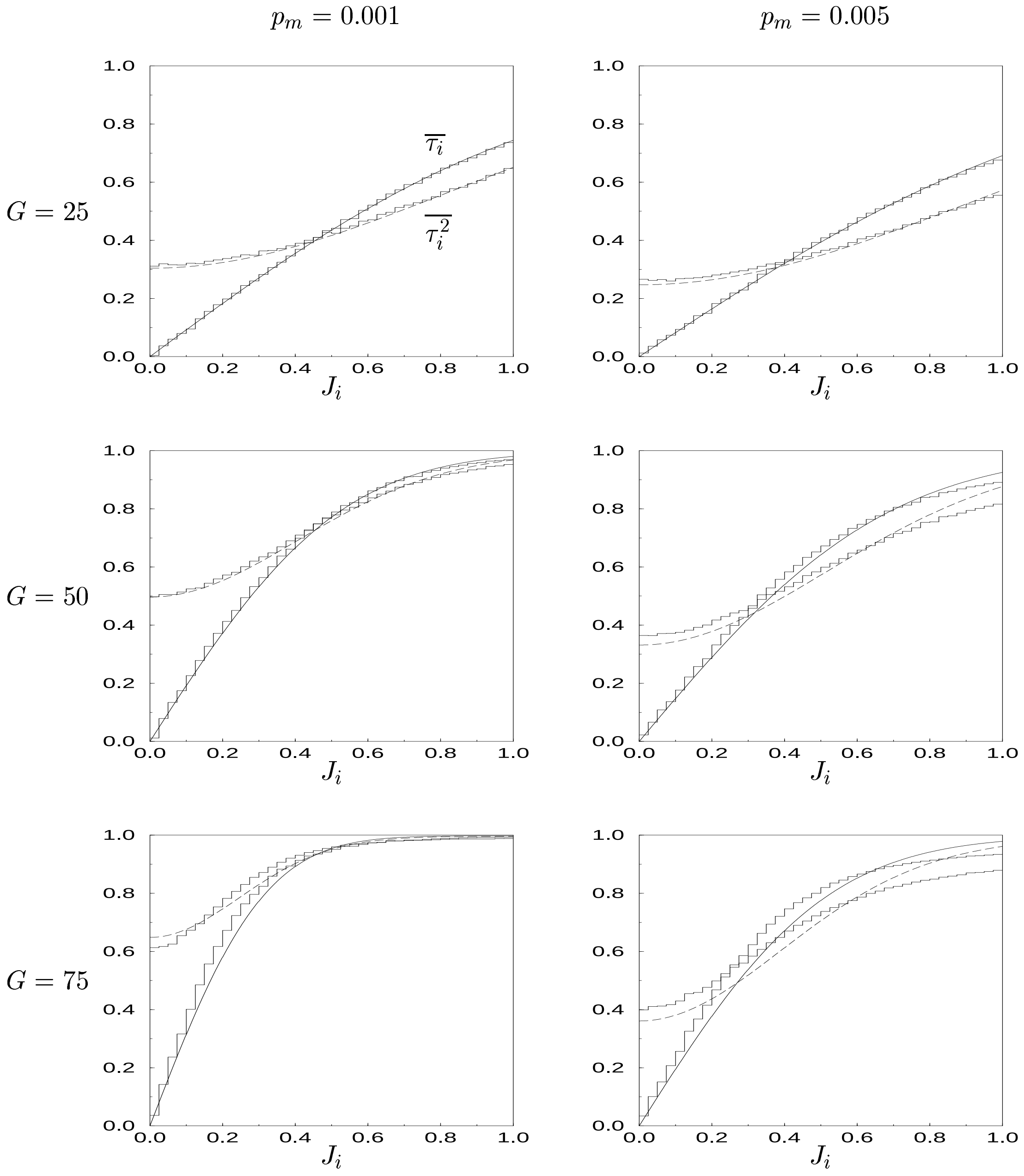

Figure 4.1: The maximum entropy result is compared to averaged results for two GAs under directional selection on an additive genotype, which differ only in their respective mutation rates. Snapshots are taken every 25 generations for the first 75 generations. The theoretical solid curves are for the mean allele per site within the population $\tau_i$ as a function of $J_i$, while the dashed curves are for the squared mean allele per site $\tau_i^2$ (the bars represent an average over all sites with the same value or range of $J_i$). The histogram results are averaged over 5000 runs of the GA and weights were uniformly distributed in the range $[0, 1]$. The simulations are the same as those used for the results presented in figures 4.5 and 4.6 and all other parameter values are given there. Notice that the histograms in the lower right of the figure cross.

If $\tau_i$ is the mean allele at site $i$, defined in equation (4.18), then the probability that population members $\alpha$ and $\beta$ have identical configurations is,

$$\Pr_{\alpha=\beta} = \prod_{i=1}^{N} \left( \frac{(1+\tau_i)^2}{4} + \frac{(1-\tau_i)^2}{4} \right) \tag{4.23}$$

Averaging the logarithm of this quantity over the Gaussian noise in $\tau_i$ one finds,

$$\Pr_{\alpha=\beta} = \exp\left( \sum_{i=1}^{N} \overline{\log[(1+\tau_i^2)/2]} \right) \tag{4.24}$$

This quantity will only be significant when $\tau_i^2 \simeq 1$, in which case,

$$\Pr_{\alpha=\beta} \simeq \exp\bigl(-\tfrac{N}{2}(1-q)\bigr) \tag{4.25}$$

where $q = \overline{\tau_i^2}$ is the mean correlation within the population.

The expected number of duplicate pairs within the population is given by this probability multiplied by the number of distinct pairs within the population,

$$\text{No. of duplicate pairs} \simeq \tfrac{1}{2}P(P-1)\exp\bigl(-\tfrac{N}{2}(1-q)\bigr) \tag{4.26}$$

When this quantity is $O(1)$ then population members can no longer be considered independent. In this case a reasonable approximation is achieved by reducing the effective population size by this amount (as long as there is negligible probability of three or more copies of the same individual being present). This is an approximation because phenotypes within the population can no longer be considered a random sample once duplicates are rejected. The effective population size is then,

$$P_{\text{eff}} = P\left[1 - \tfrac{1}{2}(P-1)\exp\bigl(-\tfrac{N}{2}(1-q)\bigr)\right] \tag{4.27}$$

As $q \to 1$ it would also be necessary to include higher order correlations (otherwise $P_{\text{eff}}$ would become negative), but this result gives a good approximation at the point when correlations first become important, as shown by the experiment described in figure 4.9.

In general, correlations will be less evenly distributed within the population than predicted by the maximum entropy ansatz and this calculation will provide a lower bound on the expected number of duplicates. There may also be functional degeneracy for integer $J_i$, in which case phenotypes may be equal without having the same configuration.

## 4.6 Evolving the mean correlation under selection

As discussed in chapter 3, section 3.2.3, the calculation for the mean correlation after selection depends on details of the problem under consideration, as it involves the relationship between genotype and phenotype. Recall equation (3.10), which describes the expected increase in correlation under selection,

$$\begin{aligned} q_{\mathrm{s}} &= \sum_{\alpha=1}^{P} p_\alpha^2 + \sum_{\alpha=1}^{P}\sum_{\beta\neq\alpha} p_\alpha p_\beta q_{\alpha\beta} \\ &= \Delta q_{\mathrm{d}} + q_{\mathrm{nat}} \end{aligned}$$

The first term is due to the duplication required in a finite population and is discussed in section 3.2.3. The second term is the natural increase in correlation as the population becomes fitter.

To simplify the calculation, it is convenient to subtract off a set of dummy variables from the first term and add the same variables to the second term,

$$\begin{aligned} q_{\mathrm{s}} &= \sum_{\alpha=1}^{P} p_\alpha^2(1-q_{\alpha\alpha}) + \sum_{\alpha=1}^{P}\sum_{\beta=1}^{P} p_\alpha p_\beta q_{\alpha\beta} \\ &= \Delta q + q_\infty \end{aligned} \tag{4.28}$$

Here, $q_{\alpha\alpha}$ is the expected correlation between distinct genotypes with the same phenotypic value $R_\alpha$. These extra variables are introduced so that $p_\alpha$ and $p_\beta$ can be treated independently in the second term (this term is denoted $q_\infty$ as it is the only contribution in the infinite population limit). The first term expresses the intuition that each duplicate pair created by selection can be thought of as replacing a pair which would otherwise be correlated by $q_{\alpha\alpha}$.

The relationship between correlations and phenotypes is required to estimate both terms in equation (4.28). It will be assumed that this relationship is well approximated by a maximum entropy distribution, as described in section 4.5, and this assumption will be justified retrospectively. Consider each contribution to the correlation in turn.

### 4.6.1 Maximum entropy result for $q_\infty$

Let $p(q_{\alpha\beta}|R_\alpha, R_\beta)$ be the conditional probability for the correlation between two population members given their phenotypes. This distribution can be determined for a maximum entropy distribution (see equation (A.3) in appendix A). The expectation value for $q_\infty$ after selection is simply the correlation averaged over this distribution and the distribution of phenotypes after selection, $p_{\rm s}(R)$,

$$q_\infty = \int \mathrm{d}q_{\alpha\beta}\,\mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_{\rm s}(R_\alpha)\,p_{\rm s}(R_\beta)\,p(q_{\alpha\beta}|R_\alpha,R_\beta)q_{\alpha\beta} \tag{4.29}$$

This integral can be computed for large $N$ by the saddle point method and in this limit the result only depends on the mean phenotype after selection. The calculation is shown in appendix A and one finds,

$$q_\infty(y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tau_i + \tanh(yJ_i)}{1+\tau_i\tanh(yJ_i)}\right)^2 \tag{4.30}$$

where $y$ is implicitly related to the mean phenotype after selection,

$$K_1^{\rm s} = \sum_{i=1}^{N} J_i\left(\frac{\tau_i + \tanh(yJ_i)}{1+\tau_i\tanh(yJ_i)}\right) \tag{4.31}$$

Here, $\tau_i$ is the mean allele at site $i$ for a maximum entropy distribution (before selection), as defined in equation (4.18). The average over the Gaussian noise in $\tau_i$ is taken over each site in both expressions. In general, it is necessary to compute these expressions numerically, first computing $y$ from equation (4.31) by numerical root finding and then substituting this value of $y$ into equation (4.30) in order to determine $q_\infty(y)$.

It is instructive to expand equation (4.31) in $y$, which is appropriate in the weak selection limit,

$$K_1^{\rm s} = K_1 + yK_2^{\rm max} + \frac{y^2}{2!}K_3^{\rm max} + \frac{y^3}{3!}K_4^{\rm max} + \cdots \tag{4.32}$$

Here, $K_n^{\rm max}$ are cumulants of the maximum entropy distribution which are defined in equation (4.21a). Truncating this expression provides a good approximation for $y$ in the weak selection limit, avoiding the need for numerical reversion of equation (4.31). This value of $y$ could then be substituted into equation (4.30) in order to determine $q_\infty(y)$.

By comparing this expansion to the Boltzmann directional selection result in equation (3.20), one finds that $y$ is equal to the selection strength $\beta$ in the infinite population limit, if the population is at maximum entropy before selection. For weak directional selection it may well be reasonable to choose $y \simeq \beta$ when approximating the dynamics, although for the simulations presented in this thesis the exact expressions were used.

### 4.6.2 Maximum entropy result for $\Delta q$

Recall the definition of $\Delta q$ given in equation (4.28),

$$\Delta q = \sum_{\alpha=1}^{P} p_\alpha^2 (1 - q_{\alpha\alpha}) \tag{4.33}$$

By averaging over each population member (as in chapter 3, section 3.2.3) one finds,

$$\Delta q = P \int_0^\infty \mathrm{d}t\, t\, f(t)\, g^{P-1}(t) \tag{4.34}$$

where,

$$\begin{aligned} f(t) &= \int \mathrm{d}R\, p(R) \int \mathrm{d}q\, p(q|R,R)\, (1-q)\, w^2(R) \exp\bigl(-tw(R)\bigr) && (4.35\text{a}) \\ g(t) &= \int \mathrm{d}R\, p(R) \exp\bigl(-tw(R)\bigr) && (4.35\text{b}) \end{aligned}$$

In general, it would be necessary to calculate these integrals numerically, but the correlation distribution is difficult to deal with as it requires the numerical reversion of a saddle point equation (see appendix A).

Instead, it is possible to expand in $1/P$ as shown in section 3.2.3, which is appropriate for weak Boltzmann selection. To leading order one finds,

$$\Delta q = \frac{\psi_2 - \hat{\psi}_2}{P\psi_1^2} + O\left(\frac{1}{P^2}\right) \tag{4.36}$$

where,

$$\begin{aligned} \psi_n &= \int \mathrm{d}R\, p(R)\, w^n(R) && (4.37\text{a}) \\ \hat{\psi}_n &= \int \mathrm{d}R\, p(R) \int \mathrm{d}q\, p(q|R,R)\, q\, w^n(R) && (4.37\text{b}) \end{aligned}$$

Notice that $\hat{\psi}_n$ can be expressed in terms of the characteristic function for the conditional distribution of correlations, which is defined in equation (A.2),

$$\frac{\hat{\psi}_n}{\psi_n} = \lim_{t\to 0} \frac{\partial}{\partial t} \log\left( \int \mathrm{d}R\, p(R) \rho(t\,|R,R) w^n(R) \right) \tag{4.38}$$

This expression depends on the particular form of selection weight, $w(R)$.

Consider directional Boltzmann selection, in which case $w(R) = \exp(\beta R)$. In this case, the expression on the right hand side of equation (4.38) can be calculated for large $N$ by the saddle point method. This follows the calculation in appendix A closely, and eventually one finds,

$$\Delta q = \frac{(1 - q_\infty[y(k)])\rho(2\beta)}{P\rho^2(\beta)} + O\left(\frac{1}{P^2}\right) \tag{4.39}$$

where $q_\infty(y)$ is defined in equation (4.30) and $\rho(\beta)$ is the characteristic function of the phenotype distribution, defined in equation (2.7). Here, $y(k)$ is found by substituting $k$ for $K_1^{\mathrm{s}}$ in equation (4.31), where $k$ is defined,

$$k = \int \mathrm{d}R\, p(R)\, R\, \mathrm{e}^{2\beta R} \tag{4.40}$$

### 4.6.3 Justifying the approximation

In the previous two sections it was shown how the mean correlation after selection may be calculated if the population is taken to be at maximum entropy before selection. This is a greater assumption than in the crossover and mutation calculations, where the maximum entropy ansatz was only required to compute on-site terms (neglecting the off-site term in the fourth cumulant after mutation). The relationship between the phenotypes of two population members and their correlation can change under crossover (unlike on-site averages), and the distribution of correlations may therefore depend on non-trivial off-site contributions. In this case it is necessary to justify the assumption that the change in correlation under selection is well described by the expressions derived here.

It is assumed that the fixed point of crossover is well modelled by the maximum entropy distribution described in section 4.5 and situations where this approximation breaks down are discussed there. Recall that $q_\infty$ was calculated for large $N$ and in this limit was found to depend only on the mean phenotype after selection (see equation (4.30)). This is an asymptotic result and one would expect terms in other cumulants to come in at $O(1/N)$ or less. This shows that the maximum entropy distribution for correlations (defined in equation (A.3)) is self-consistent in the limit of large $N$, as it returns the correct mean correlation given the mean phenotype of the population in this limit, irrespective of the variance and higher cumulants. It is therefore at least consistent to assume that higher cumulant effects (and therefore off-site terms) are of secondary importance, although this is not necessarily so. The assumption is that there is no significant systematic bias in the distribution of correlations within the population.

This approximation will break down in certain situations. For example, when very little crossover is used or when crossover is not very disruptive (small $a$ in equation (4.8)) then it is unlikely that correlations will redistribute sufficiently quickly to ensure a smooth distribution within the population. In this case the very fit individuals within the population might be closely related and duplication will lead to a much greater increase in correlation than would be estimated by assuming evenly distributed correlations. Under these circumstances, the relationship between genotypes within the population would probably be unpredictable with information from only a small number of macroscopics. Further experiments are required to determine when the maximum entropy distribution of correlations will accurately characterize the population. The results presented in this chapter indicate that these results are at least accurate when uniform crossover is used over the whole population.

## 4.7 Directional selection

If fitness equals the phenotype then selection is directional.

$$F_\alpha = R_\alpha = \sum_{i=1}^{N} J_i S_i^\alpha \tag{4.41}$$

Figure 4.2 shows the typical averaged dynamics of a GA under this fitness measure, where the population increases in fitness and moves into a state of progressively lower entropy each generation. Eventually the population may reach an equilibrium state, where the effects of selection and mutation are balanced. Without mutation, the population will eventually converge on a state where each population member is identical.

There is only one optimum configuration, which is given by the state with $S_i J_i \geq 0$ at each site, and there are no sub-optimal local fitness maxima. If the weights are chosen from a distribution then this problem is the random-field paramagnet, which was considered under the statistical mechanics formalism by Prügel-Bennett and Shapiro [54]. They also considered a closely related problem, the spin-glass chain, where nearest neighbour interactions contribute additively to fitness. These problems are equivalent under a trivial gauge transformation, although the dynamics differs due to the existence of an interface energy in the spin-glass which leads to a large number of local fitness maxima under single spin-flip dynamics. They solved

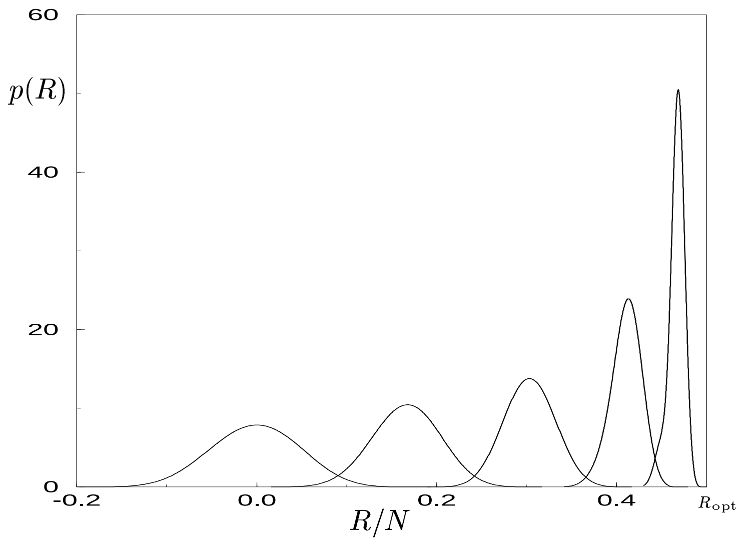

Figure 4.2: Evolution of a GA under directional selection, averaged over 5000 runs. The phenotype distribution is shown at $0, 20, 40, 70$ and $120$ generations. Weights were selected from a uniform distribution in the range $[0, 1]$, so that the optimum phenotype $R_{\rm opt}$ was $N/2$ on average. The other parameters were $P = 80, N = 150, p_{\rm m} = 0.002, \beta_{\rm s} = 0.25$ and uniform crossover was used with $a = 0.5$.

the dynamics of the paramagnet under the assumption that crossover leaves the variance of the fitness distribution unchanged. This seemed to be a reasonable approximation in some situations, but is incorrect in general. Here, a more exact approach is used, in which the mean correlation within the population is evolved as an explicit macroscopic according to the expressions derived in sections 4.3 and 4.6. Before considering the random-field paramagnet it is instructive to consider the simpler one-max problem, where the weights at every site are equal.

### 4.7.1 One-max

The fitness for the one-max problem is given by equation (4.41) with $J_i = 1$ at every site,

$$F_\alpha = \sum_{i=1}^{N} S_i^\alpha \tag{4.42}$$

Under Boltzmann selection, the alleles contribute to the selection probability multiplicatively and biologists call this a multiplicative fitness landscape (although they use a different notation) [11],

$$w_\alpha = \exp(\beta F_\alpha) = \prod_{i=1}^{N} \exp(\beta S_i^\alpha) \tag{4.43}$$

This problem has been studied extensively in the GA literature and a number of results have recently been obtained which predict the trajectory of mean fitness within the population for a number of selection schemes [45, 67, 72]. These methods rely on the population being sufficiently large so that the distribution of alleles is accurately modelled by a binomial distribution. This is a maximum entropy distribution with a constraint on mean fitness alone. These models break down in a finite population, because in this case the population will become more correlated under selection than predicted by a binomial distribution. In the infinite population limit the results presented here reduce to the results for these simpler, but less general, models.

In order to simulate the dynamics, expressions for the change in the first four cumulants and mean correlation under each operator were iterated in sequence (see sections 3.3, 4.3, 4.4 and 4.6). The theory is compared to averaged results from a standard GA in figure 4.3, for two different population sizes. The mean and variance of the fitness distribution and the highest fitness are shown, averaged over 1000 runs. Error bars were typically smaller than the symbols and are not shown. These results show good agreement with the theory, which accurately describes finite population effects. The skewness and kurtosis are shown in figure 4.4 for one population size, also agreeing well with the theory (more samples were required to obtain good averages for the higher cumulants).

Notice that the results in figure 4.3 for the smallest population size show small systematic errors. The theory eventually breaks down for very small populations and for strong selection. This is thought to be mainly because a weak selection approximation was required to calculate the duplication contribution to the increased correlation under selection (see section 4.6.2), although there might also be errors due to non-self-averaging effects. The approximation in determining the correlation after selection was considered most important, as theoretical results for the correlation were first to break down. To minimize this source of error, the contribution to equation (4.33) which does not involve the correlation was calculated numerically.

### 4.7.2 Random-field paramagnet

The fitness of the random-field paramagnet is given by equation (4.41) with weights chosen from some distribution. Here, the case where weights are chosen uniformly from the interval

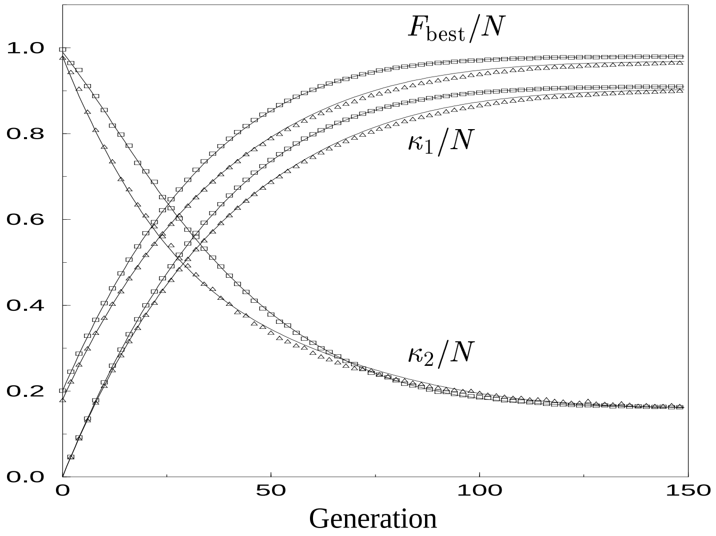

Figure 4.3: The theory is compared to averaged results for one-max with population sizes $50(\triangle)$ and $100(\Box)$. The mean ($\kappa_1$) and variance ($\kappa_2$) of the fitness distribution and the highest fitness are shown, averaged over 1000 runs, with solid lines showing the theory. The other parameters were $N = 155, p_{\rm m} = 0.005, \beta_{\rm s} = 0.3$ and uniform crossover was used with $a = 0.5$.

$[0, 1]$ is considered, although there is no significant difference to the dynamics if a Gaussian distribution is used.

As in the the previous section, the GA dynamics was simulated by iterating the difference equations describing the effects of each operator on the first four cumulants and the mean correlation within the population (see sections 3.3, 4.3, 4.4 and 4.6). In section 4.5.2 it was noted that the maximum entropy ansatz might break down in some cases, most notably when mutation is likely to flip large weights and selection is not sufficiently strong to ensure such mutations are removed from the population. Figures 4.5 and 4.6 compare the theory to averaged results from a standard GA for weak and moderate mutation rates. As expected, the results for weak mutation show better agreement and it seems that the formalism as it stands is only accurate in describing the GA with a low mutation rate for this problem. This was not the case in the one-max problem, where the weights at every site are equal.

It is not known whether the addition of extra macroscopics might provide a better characterization of the population. Experiments were performed to determine if the inclusion of a third constraint on the mean allele within the population (the mean magnetization) would characterize the population better, but the results showed no significant improvement over the

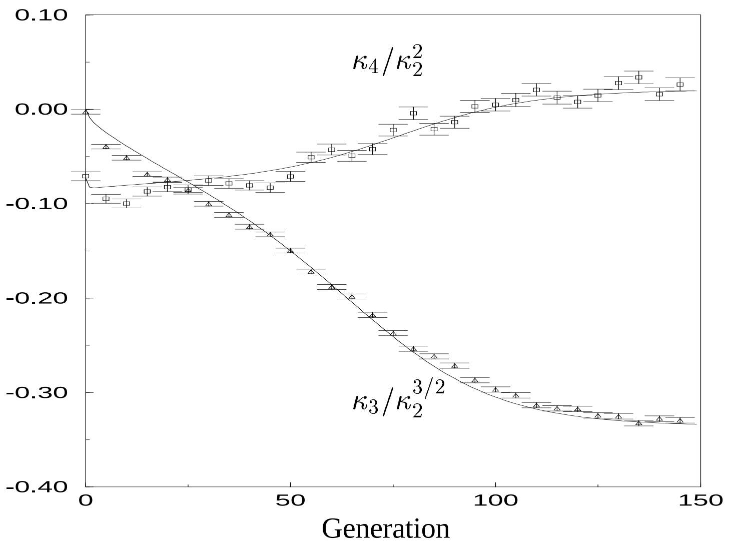

Figure 4.4: The skewness (△) and kurtosis (□) are shown for the same parameter values as the results presented in figure 4.3 for population size 100. The results were averaged over 10 000 runs. The solid lines give the theoretical result and averages were taken over cumulants, rather than the ratios shown.

two constraint model.

## 4.8 Stabilizing selection

A rather different dynamical behaviour is possible if the optimum fitness is given by a phenotype of intermediate value which lies in a high entropy region of the phenotype space. A possible fitness measure in this case might be,

$$F_\alpha = -\frac{1}{N}\left(R_\alpha - R_{\text{opt}}\right)^2 = -\frac{1}{N}\left(\sum_{i=1}^{N} J_i S_i^\alpha - R_{\text{opt}}\right)^2 \qquad (4.44)$$

where the factor of $1/N$ is chosen to make the fitness typically $O(N)$. Here, $R_{\text{opt}}$ is the optimum possible phenotype. There may be no configuration which gives a phenotype exactly equal to $R_{\text{opt}}$, in which case the closest obtainable phenotype provides the optimum fitness. Notice that the fitness defined here is never greater than zero and in this case it may be more natural to use energy (negative fitness); however, the fitness convention is retained for consistency.

Figure 4.7 shows the typical averaged evolution of a GA under this form of selection. Initially selection is directional, as the population mean moves towards the optimum phenotype.

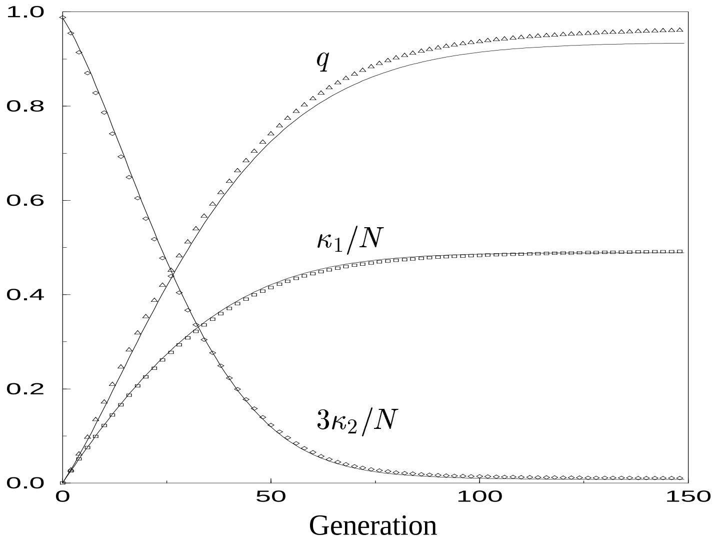

Figure 4.5: The theory is compared to averaged results for the paramagnet with a low mutation rate. The mean (□), variance (◇) and the correlation (△) are shown averaged over 5000 runs, with solid lines showing the theory. The weights were chosen from a uniform distribution in the range $[0, 1]$ so that the optimum was $N/2$ on average. The other parameters were $P = 80, N = 120, p_{\rm m} = 0.001, \beta_{\rm s} = 0.25$ and uniform crossover was used with $a = 0.5$.

After some time the population stabilizes around the optimum phenotype and the phenotypic variance decreases, as the population converges.

As in directional selection, the population may reach a balance between selection and mutation, or in the absence of mutation the population will eventually converge onto a state where all population members are identical. The dynamical behaviour is significantly different here, however, because solutions within the population are in a dense region of the search space, while under directional selection the population moves into an increasingly sparse region of the search space. Depending on the particular distribution of $J_i$ over sites, there may be many local maxima of high fitness, whereas under directional selection the only fitness maximum is the optimum configuration.

The fitness measure defined in equation (4.44) provides an appropriate algorithm for solving the subset sum problem. The subset sum problem asks whether a set of numbers, here the weight vector $\{J_1, J_2, \dots, J_N\}$, contains a subset which exactly sums to some goal value. Posed as an optimization problem one wishes to find the subset which comes as close to the goal value as possible. Clearly, the subset sum problem is more appropriately defined in terms

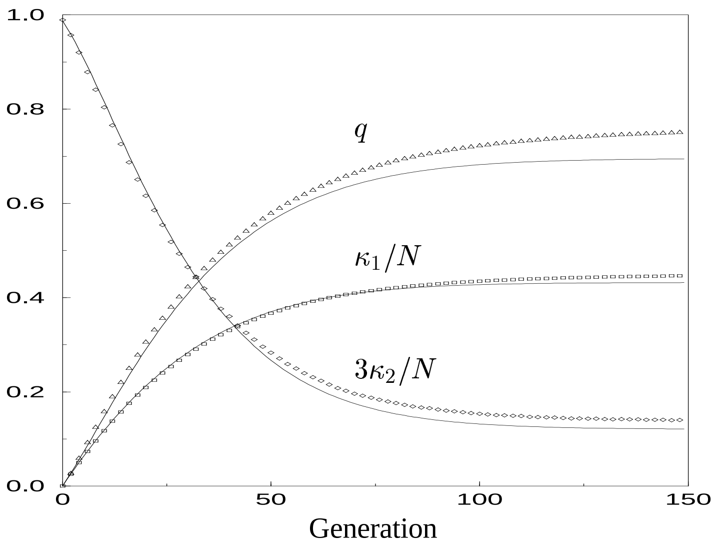

Figure 4.6: The theory is compared to averaged results for the paramagnet with a moderate mutation rate. The mean (□), variance (◇) and the correlation (△) are shown averaged over 5000 runs, with solid lines showing the theory. All the other search parameters are as in figure 4.5, except that the mutation rate is $p_{\rm m} = 0.005$.

of alleles taking the values 1 or 0, denoting whether or not a weight is selected for the subset. However, the problem can be cast in terms of Ising spins under a change in variables $X_i^\alpha = \frac{1}{2}(S_i^\alpha + 1)$. Then the optimum phenotype $R_{\rm opt}$ can be chosen so that the goal value for the subset sum problem is $\frac{1}{2}(R_{\rm opt} + \sum J_i)$.

Although the subset sum problem is strictly NP-hard, it is pseudo-polynomial and for typical weight distributions can be solved in polynomial time by standard methods [17]. The GA is not expected to outperform polynomial time algorithms and the aim of this study is not a comparison of methods on this problem. However, there are related strong NP-hard problems, such as bin-packing, to which GAs have been successfully applied [38]. It is hoped that a solution of the dynamics for this problem might provide some insight into these harder problems. This problem is also of some interest as a model of stabilizing selection in quantitative genetics (see, for example, reference [7]).

### 4.8.1 Cumulants after selection

Under this fitness measure, the selection weight for Boltzmann selection is,

$$w_\alpha = \exp\left(-\frac{\beta}{N}\left(R_\alpha - R_{\rm opt}\right)^2\right) \tag{4.45}$$

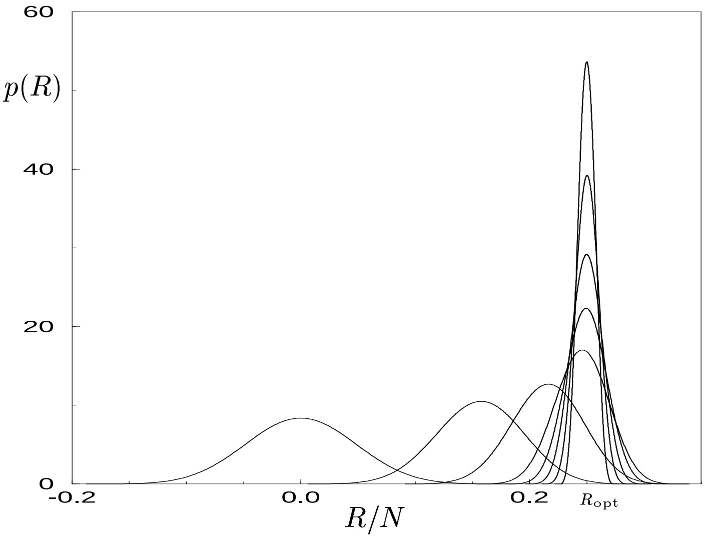

Figure 4.7: Evolution of a GA under stabilizing selection, averaged over 2000 runs. The phenotype distribution is shown at generation $0$, $20$, $40$, and every $40$ generations up to $240$. The weights where selected from a uniform distribution in the range $[0,1]$ and the optimum phenotype was $R_{\rm opt} = N/4$. The other parameters were $P = 80, N = 150, p_{\rm m} = 0, \beta_{\rm s} = 0.025$ and uniform crossover was used with $a = 0.5$.

In general, the cumulants after selection can be determined numerically from equation (3.5) using Gaussian quadratures.

The weak selection expansion described in section 3.2.2 is instructive, as it shows the contribution to each cumulant after selection explicitly. The Gram-Charlier expansion in equation (2.11) can be used to parameterize the distribution of phenotypes. For the first three cumulants up to first order in $\beta$ and to $O(1/P)$ one finds,

$$K_1^{\rm s} = K_1 + \beta_{\rm s}\left(1-\frac{1}{P}\right)\left(2(R_{\rm opt}-K_1) - \frac{K_3}{K_2}\right) + O(\beta^2) \tag{4.46a}$$

$$K_2^{\rm s} = \left(1-\frac{1}{P}\right)K_2 - 2\beta_{\rm s}\left(1-\frac{3}{P}\right)\left(K_2 - (R_{\rm opt}-K_1)\frac{K_3}{K_2}\right) + O(\beta^2) \tag{4.46b}$$

$$K_3^{\rm s} = \left(1-\frac{3}{P}\right)K_3 - 6\beta_{\rm s}\left[\left(1-\frac{8}{P}\right)K_3 + \frac{2}{P}(R_{\rm opt}-K_1)K_2\right] + O(\beta^2) \tag{4.46c}$$

where $\beta_{\rm s} = \beta K_2/N$ is a scaled selection parameter (this differs from the scaling used for directional selection).

During the initial, directional dynamics when $K_1 \neq R_{\rm opt}$ the mean phenotype moves towards the optimal phenotype, as expected. If the mean is increasing, then the third cumulant becomes negative as it does in directional selection. This leads to reduced variance under further selection and a loss of diversity within the population. This effect is also observed in the correlation expression which is presented in the next section.

After a number of generations the mean becomes arbitrarily close to $R_{\rm opt}$ and the magnitude of the third cumulant is reduced as the population becomes more symmetrical (as long as mutation is weak). Ignoring higher cumulants, the ratio of the variance after and before selection at this stage is,

$$\frac{K_2^{\rm s}}{K_2} = \frac{1}{1+2\beta_{\rm s}} - \frac{1}{P(1+4\beta_{\rm s})^{\frac{3}{2}}} \tag{4.47}$$

One can keep this ratio fixed by scaling $\beta$ in order to keep $\beta_{\rm s}$ constant and maintain selective pressure. As in directional selection, this requires an increased selection strength as the GA converges.

### 4.8.2 Mean correlation after selection

The mean correlation after selection can be calculated as described in section 4.6. The only difference between the present case and the directional selection calculation presented there is in the duplication term, defined in equation (4.33). The expression in equation (4.38) can again be calculated for large $N$ by the saddle point method in a similar calculation to that presented in appendix A. This yields the following expression for $\Delta q$,

$$\Delta q = \frac{(1-q_\infty[y(k)])\psi(2\beta)}{P\psi^2(\beta)} + O\left(\frac{1}{P^2}\right) \tag{4.48}$$

where,

$$\psi(\beta) = \int {\rm d}R\, p(R) \, \exp\left(-\frac{\beta}{N}(R-R_{\rm opt})^2\right) \tag{4.49}$$

Here, $q_\infty(y)$ is defined in equation (4.30) and $y(k)$ is found by substituting $k$ for $K_1^{\rm s}$ in equation (4.31), where $k$ is defined,

$$k = \int {\rm d}R\, p(R)\, R \, \exp\left(-\frac{2\beta}{N}(R-R_{\rm opt})^2\right) \tag{4.50}$$

These expressions can be calculated by parameterizing the distribution of phenotypes using the Gram-Charlier expansion given in equation (2.11).

Expanding in $\beta$ shows the relevant contributions from each cumulant and up to second order one finds (ignoring terms of $O(1/\sqrt{N})$ and less),

$$\Delta q \simeq \frac{1-q_\infty[y(k)]}{P}\left[1+2\beta_{\rm s}^2\left(1+\frac{2(R_{\rm opt}-K_1)^2}{K_2} - \frac{2(R_{\rm opt}-K_1)K_3}{K_2^2}\right)\right] \tag{4.51}$$

where $\beta_{\mathrm{s}}$ is the scaled selection strength which was defined in section 4.8.1. This shows that when the mean phenotype is increasing under selection, the negative third cumulant introduced by selection results in an increased correlation under further selection.

Once the mean phenotype within the population stabilizes around $R_{\mathrm{opt}}$, then the main contribution to the increased correlation is through the duplication term, since for large $N$ the natural increase term defined in section 4.6 depends only on the mean phenotype. If the population size increases exponentially with $N$, however, it might be necessary to go beyond leading order in the saddle point calculation given in appendix A. This refinement has not been pursued here because the population sizes under consideration are generally of $O(N)$ or less.

### 4.8.3 Best population member

One can estimate the best individual within the population by assuming population members are independently sampled from an infinite population, as described in chapter 2, section 2.5. For stabilizing selection the phenotype distribution $p(R)$ and fitness distribution $\mathcal{P}(F)$ are related through equation (2.2) which yields the following expression,

$$\mathcal{P}(F) = \frac{1}{2}\sqrt{\frac{N}{|F|}}\left(p(R_{\mathrm{opt}} - \sqrt{N|F|}) + p(R_{\mathrm{opt}} + \sqrt{N|F|})\right)\Theta(-F) \tag{4.52}$$

where $|F|$ is the magnitude of the fitness (here, $F \leq 0$). Eventually, the phenotype distribution is centred around $R_{\mathrm{opt}}$ and substituting the above expression into equation (2.16) for a Gaussian phenotype distribution one finds,

$$F_{\mathrm{best}} = -\frac{2P}{N}\int_0^\infty \frac{\mathrm{d}R}{\sqrt{2\pi K_2}}\,\mathrm{erfc}^{P-1}(R)\,R^2\,\exp(-R^2/2K_2) \tag{4.53}$$

Other cumulants can also be included by parameterizing the phenotype distribution using the Gram-Charlier expansion in equation (2.11). In general, this integral must be computed numerically.

A good approximation can be achieved by using a flat distribution with the same height as the phenotype distribution at $R_{\mathrm{opt}}$. This does not affect the best population member significantly, since the population is locally flat in the neighbourhood of the mean phenotype when it is close to $R_{\rm opt}$. Using four cumulants, the height at the mean when $K_1 = R_{\rm opt}$ is,

$$H = \frac{1}{\sqrt{2\pi K_2}}\left(1+\frac{K_4}{8K_2^2}\right) \tag{4.54}$$

The fitness of the best population member is then,

$$\begin{aligned} F_{\rm best} &\simeq -\frac{2HP}{N}\int_0^{\frac{1}{2H}} {\rm d}R\, R^2\,(1-2HR)^{P-1} \\ &= \frac{-\pi K_2}{N(P+1)(P+2)(1+K_4/8K_2^2)^2} \end{aligned} \tag{4.55}$$

This will be an upper bound, because there is a larger probability within the neighbourhood of $R_{\rm opt}$ for a flat distribution than for a Gaussian, but it should become exact for large $P$.

As discussed in section 4.5.3, the assumption that population members are independently sampled from a continuous distribution may break down when the population becomes highly correlated. This is remedied by using the smaller effective population size expression in equation (4.27). There is also the possibility that when the population becomes very narrow, the fine grain structure of the phenotype space may become important. This feature of the search is not described by the macroscopics under consideration in this work.

### 4.8.4 Simulating the dynamics

The dynamics of the GA was simulated by iterating difference equations describing the effect of each operator on the first four cumulants and the mean correlation within the population (see sections 4.3, 4.4, 4.8.1 and 4.8.2). The theory is compared to averaged results from a standard GA under stabilizing selection in figures 4.8 and 4.9, with weights taken uniformly from the interval $[0, 1]$.

The theory shows good agreement, although there is an underestimate in the correlation and a corresponding overestimate in the variance during the later stages of the dynamics. This can mostly be attributed to an underestimate in the increased correlation under selection, which may be due to bias in the distribution of correlations, as discussed in section 4.6.3.

Notice that the fitness of the best individual eventually drops, as shown in figure 4.9, and this drop is accurately predicted by the theoretical result from section 4.8.3 with the effective population size chosen according to equation (4.27). This is a consequence of the increased

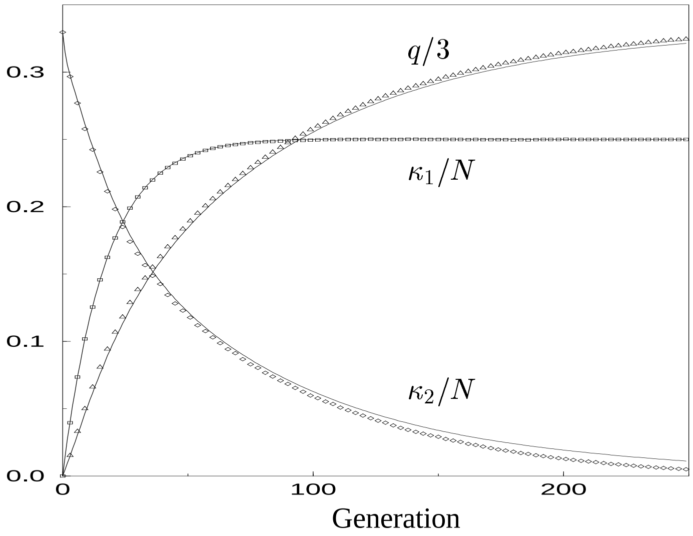

Figure 4.8: The theory is compared to averaged results from a GA under stabilizing selection. The mean (□), variance (◇) and the correlation (△) are shown averaged over 2000 runs, with the solid lines showing the theory. The weights where selected from a uniform distribution in the range $[0, 1]$ and the optimum phenotype was $R_{\text{opt}} = N/4$. The other parameters were $P = 80, N = 150, p_{\text{m}} = 0, \beta_{\text{s}} = 0.03$ and uniform crossover was used with $a = 0.5$.

correlation within the population, which leads to a large number of duplicates and a corresponding reduction in the effective population size. The search becomes ineffective after this point.

Unfortunately, the validity of the maximum entropy ansatz was not ascertained for different mutation rates, as in the case of the random-field paramagnet. Mutation was not used in these simulations because it was not thought to be of critical importance when these experiments were carried out and because this led to interesting behaviour when the correlation was very high, as described above. The effect of mutation moving the population away from maximum entropy, as described in section 4.5.2, may not be so important under stabilizing selection, because the higher weights are not so critical when the population is not close to the extreme of all ones (or all negative ones). However, further experiments should be carried out in order to test this view.

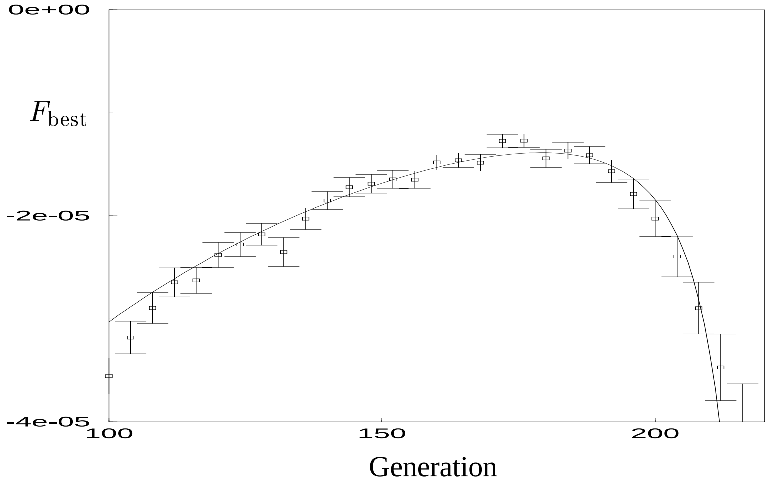

Figure 4.9: The best population member each generation is averaged over the same runs as in figure 4.8. The solid line gives the theoretical result. As the population becomes highly correlated the number of independent population members drops, leading to a corresponding drop in fitness of the best individual.

## 4.9 Bit-simulated crossover limit

It is useful to define a limit which can be used if bit-simulated crossover (BSC) is an appropriate crossover operator (see the final paragraph of section 4.4) [71]. This is usually only the case if sites contribute independently to the fitness, for example under directional selection on an additive genotype. This limit allows a microscopic description of the population after crossover, which facilitates the solution to the dynamics for more involved problems in chapter 7. It also allows the correlation after selection to be calculated directly from the selection partition function. A nice feature of these results is that they do not use the large $N$ limit, which was required to calculate the increased correlation under selection in section 4.6. However, to decouple the average over the distribution of alleles at each site it is necessary to use a weak selection approximation.

### 4.9.1 Cumulants after selection

Under BSC, the population is brought straight to the fixed point of standard crossover, which is assumed to be the maximum entropy distribution described in section 4.5. In this case the distribution of alleles at each site decouples and it is possible to average the cumulant generating function for selection over this distribution. For weak Boltzmann selection the $1/P$ expansion described in chapter 3, section 3.2.2 is appropriate. For directional selection the cumulants after selection are then given by

$$K_n^{\mathrm{s}} \simeq \frac{\partial^n}{\partial \beta^n}\left[\log(\rho(\beta)) - \frac{1}{2P}\left(\frac{\rho(2\beta)}{\rho^2(\beta)}\right)\right] \tag{4.56}$$

where $\rho(\beta)$ is now averaged over alleles, rather than the distribution of phenotypes,

$$\rho(\beta) = \left\langle \exp\left(\beta \sum_{i=1}^{N} J_i S_i\right)\right\rangle_{\{S_i\}} \tag{4.57}$$

The alleles are distributed according to equation (4.18),

$$p(S_i) = \left(\frac{1+\tau_i}{2}\right)\delta(S_i - 1) + \left(\frac{1-\tau_i}{2}\right)\delta(S_i + 1) \tag{4.58}$$

Completing the average,

$$\begin{aligned} \rho(\beta) &= \prod_{i=1}^{N}\left[\left(\frac{1+\tau_i}{2}\right)\mathrm{e}^{\beta J_i} + \left(\frac{1-\tau_i}{2}\right)\mathrm{e}^{-\beta J_i}\right] \\ &= \prod_{i=1}^{N} \mathcal{Z}_i(\beta, 0) \;=\; \exp\left(\sum_{n=1}^{\infty} \frac{\beta^n K_n^{\max}}{n!}\right) \end{aligned} \tag{4.59}$$

where $\mathcal{Z}_i(\beta, \epsilon)$ is the single-site partition function defined in equation (4.22) and $K_n^{\max}$ is the $n$th cumulant of the maximum entropy phenotype distribution, which is assumed to describe the population after BSC. Thus, $\rho(\beta)$ turns out to be the characteristic function for the maximum entropy phenotype distribution. This gives the same result obtained by averaging directly over the phenotype distribution, which is shown in chapter 3, section 3.20.

Writing the results in terms of the mean allele at each site, one finds that the mean phenotype after selection is,

$$\begin{aligned} K_1^{\mathrm{s}} = \sum_{i=1}^{N} J_i &\left(\frac{\tau_i + \tanh(\beta J_i)}{1 + \tau_i \tanh(\beta J_i)}\right) \\ &- \frac{\rho(2\beta)}{P\rho^2(\beta)} \sum_{i=1}^{N} J_i \left(\frac{\tau_i + \tanh(2\beta J_i)}{1 + \tau_i \tanh(2\beta J_i)} - \frac{\tau_i + \tanh(\beta J_i)}{1 + \tau_i \tanh(\beta J_i)}\right) \end{aligned} \tag{4.60}$$

This expression will be used again in chapter 7.

### 4.9.2 Mean correlation after selection

As well as generating the cumulants after selection, it is also possible to generate the mean correlation after selection. Although this provides a more elegant means of calculating the correlation than the discussion in section 4.6, it is less general and is only relevant in the BSC limit considered here.

The mean correlation after selection can be generated by including a new term in the selection partition function,

$$q_{\rm s} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \lim_{\epsilon\to 0}\frac{\partial^2}{\partial\epsilon^2}\log Z_q(\epsilon)\right) \tag{4.61}$$

where,

$$Z_q(\epsilon) = \sum_{\alpha=1}^{P} w_\alpha \exp\left(\epsilon S_i^\alpha\right) \tag{4.62}$$

Again, for weak directional selection the $1/P$ expansion in section 3.2.2 is appropriate. The averaged logarithm of the selection partition function is then given by,

$$\langle\log Z_q(\epsilon)\rangle \;\simeq\; \log(\rho(\beta,\epsilon)) - \frac{1}{2P}\left(\frac{\rho(2\beta,2\epsilon)}{\rho^2(\beta,\epsilon)}\right) \tag{4.63}$$

where,

$$\rho(\beta,\epsilon) = \mathcal{Z}_i(\beta,\epsilon)\prod_{j\neq i}^{N-1}\mathcal{Z}_j(\beta,0) \tag{4.64}$$

Here, $\mathcal{Z}_i(\beta,\gamma)$ is the single-site partition function defined in equation (4.22). Differentiating out one finds,

$$\begin{aligned} q_{\rm s} = {} & \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tau_i + \tanh(\beta J_i)}{1+\tau_i\tanh(\beta J_i)}\right)^2 \\ & + \frac{\rho(2\beta)}{NP\rho^2(\beta)}\sum_{i=1}^{N}\frac{(1-\tau_i^2)(2-\cosh(2\beta J_i) - \tau_i\sinh(2\beta J_i))}{(\cosh(\beta J_i)+\tau_i\sinh(\beta J_i))^2(\cosh(2\beta J_i)+\tau_i\sinh(2\beta J_i))} \end{aligned} \tag{4.65}$$

where $\rho(\beta)$ is the characteristic function for the phenotype distribution at maximum entropy defined in equation (4.59). Notice that as $P \to \infty$ the first term becomes equal to the natural increase contribution to the correlation after selection which was previously derived in equation (4.30) ($y \to \beta$ in this limit).

### 4.9.3 Linkage disequilibrium for one-max

For the one-max problem, where $J_i = 1$ at every site, the correlation and variance are related by a particularly simple relationship after BSC, since off-site terms in equation (4.2b) cancel.

$$K_2^{\rm max} = N(1-q) \tag{4.66}$$

This relationship no longer holds after selection. In quantitative genetics the deviation from this equality is known as the second order linkage disequilibrium, which is denoted $D$ [11].

$$\begin{aligned} D &= \sum_{i \neq j} \langle S_i^\alpha S_j^\alpha \rangle_\alpha - \langle S_i^\alpha \rangle_\alpha \langle S_j^\alpha \rangle_\alpha \\ &= K_2 - N(1-q) \end{aligned} \tag{4.67}$$

Uniform crossover reduces this quantity by a factor of $1 - 2\mathcal{A}$ each generation on average, where $\mathcal{A}$ is a parameter which determines how disruptive crossover is (see section 4.4). In reference [54], Prügel-Bennett and Shapiro worked under the assumption that this quantity remains small, so that the variance remains fixed under crossover. Using the expressions derived in equations (3.21b) and (4.65) it is possible to expand the linkage disequilibrium after selection in $\beta$. Only finite population terms of order $\beta^2$ and above remain,

$$D_{\mathrm{s}} = -\frac{\beta^2}{P}\left(3K_2^2 + \frac{K_4 - 4K_2}{3}\right) + O(\beta^3) \tag{4.68}$$

Typically, $\beta$ is $O(1/\sqrt{N})$ and the cumulants are $O(N)$, so this term is typically $O(N/P)$. If the population size is $O(N)$ or greater then one might expect this term to be negligible for large $N$. However, if less disruptive forms of crossover are used then this may not be the case, as effects will be cumulative. Often $P$ is smaller than $N$, in which case the assumption that the linkage disequilibrium is negligible will certainly be unfounded.

## 4.10 Conclusion

Results due to Prügel-Bennett and Shapiro [53, 54] for describing the effects of mutation and crossover on the phenotype distribution for a problem with an additive genotype were reproduced. A maximum entropy ansatz was required in order to deduce terms not trivially related to the given macroscopics and certain conditions under which this ansatz might break down were described. These results were then combined with new results for evolving the correlation as an explicit macroscopic and this provided a more accurate model of the dynamics than the simplification required in the previous formulation. The most important new result was the expression describing the change in the mean correlation under selection. This was divided into two contributions – a duplication term represents increases in correlation due to the duplication of population members required in a finite population, while a natural increase term represents changes in correlation as the population moves into a new region of state space. The second contribution requires information about the mapping between genotype and phenotype and this was provided by the maximum entropy ansatz. Conditions under which this result might break down were discussed and it was suggested that further work be carried out to determine limits of applicability.

Results were presented for directional selection with homogeneous weights at each site (one-max) and inhomogeneous weights at each site (the random-field paramagnet). The theory agreed well with averaged results from a real GA, so long as the maximum entropy ansatz was a good approximation. Unfortunately, this was not the case for the paramagnet with a significant mutation rate, as mutation was shown to take the population away from maximum entropy. Stabilizing selection with inhomogeneous weights at each site (the subset sum problem) was also considered and again the theory showed good predictive power. As the population became highly correlated, the effective population size was reduced by the existence of duplicates and the fitness of the best solution eventually stopped increasing and began to degrade.

The results in this chapter are encouraging, although there was also a realization that caution is required when using a maximum entropy ansatz. Assumptions with no *a priori* justification must always be checked for validity. Nevertheless, the formalism was shown to give powerful predictions and accurately accounted for finite population effects. The inclusion of an extra macroscopic, the mean correlation within the population, was certainly an essential ingredient and marks an important departure from the infinite population idealization which is often used (explicitly or implicitly) in theoretical models.

The work presented in this chapter provides the basis for analysing a number of other interesting problems. In chapters 5 and 7 these results will be used to model the dynamics for a simple learning problem, a diploid GA and a problem with temporally varying fitness. Some work has also been done on two-well potentials, although this is a difficult problem because the population may become bi-modal and strongly non-Gaussian [64]. In this case a cumulant expansion is not ideal, although the superposition of two Gaussian distributions (or approximately Gaussian distributions) sometimes provides a reasonable approximation. In another interesting study, Prügel-Bennett used the formalism to model an asexual GA for the one-max problem by extending the expressions presented in this chapter to include covariances. The dynamics was then described by the evolution of a non-interacting ensemble of populations [52]. This refinement was necessary because fluctuations from mean behaviour are large when crossover no longer suppresses the growth of higher cumulants under selection. This study showed explicitly how the continued inclusion of more macroscopics (successively higher cumulants) lead to a steady improvement in theoretical predictions.

Although the power of the statistical mechanics formalism has been demonstrated, there is still more work required to determine when the results in this chapter may break down. For example, if recombination is very non-disruptive or selection is strong then it is unlikely that correlations will be sufficiently well distributed within the population for the correlation expressions to work well. It should also be pointed out that most of the work in this chapter, and in the remainder of the thesis, is concerned with deriving equations of motion for the GA. No significant effort has yet been made to analyse these expressions. The aim of this work is to provide a compact description of the dynamics and once this has been achieved it is hoped that analysis of the resulting expressions will lead to greater understanding and, hopefully, quantitative results for optimizing performance. This latter goal is achieved to some extent in the next chapter, where expressions are derived for the optimal training batch size in a simple learning problem, but further analysis is required to gain more general insight.

# Chapter 5

# Noise corrupted fitness and a simple learning problem

## 5.1 Introduction

It is important to understand the effects of noise in the fitness evaluation, as this has implications for many optimization and machine learning problems [4, 13, 44]. The fitness measure used in these problems often involves some uncertainty, or noise, due to the limited or corrupted data available for determining fitness. For example, one common paradigm from machine learning involves generalization of a mapping given an incomplete set of training examples. This can be achieved through supervised learning, which typically involves the minimization of some form of training error, such as the number of misclassified training examples in a batch. The training batch will typically not contain every example and is therefore susceptible to random bias, or noise. Of course, there are other sources of error when attempting to generalize, such as overfitting to a particular training set or poor performance of the chosen learning algorithm. Here, discussion is limited to the effects of noise in fitness evaluation.

It has previously been argued that GAs perform well in the presence of noise compared to other search methods [13]. Indeed, it has recently been shown that a GA can outperform simple local search methods on a class of additive problems related to one-max when fitness is corrupted by noise [4]. In another recent study, Miller and Goldberg determined the effect of noise on the change in mean fitness under selection for a continuous Gaussian fitness distribution [44]. However, although they chose the population size in order to remove finite population effects, this choice was based on a conservative predictor rather than an exact result [20]. A more appropriate method for modelling selection on a finite population was introduced in chapter 3. The inclusion of finite population effects proves to be of crucial importance when characterizing the subtle effects of noise in the evaluation of fitness.

In this chapter the statistical mechanics formalism is extended to describe selection on a general stochastic fitness measure. For the case of additive Gaussian noise and weak Boltzmann selection, an increase in population size is shown to completely remove the effects of noise. Since the other genetic operators do not depend on population size, this resizing effectively maps the noisy dynamics onto the zero noise case. The theory is tested on the one-max problem corrupted by noise and agrees closely to averaged results from a real GA. Under Boltzmann selection, Gaussian noise only affects finite population terms and this emphasizes the importance of accounting for these terms accurately.

As well as introducing a general result for calculating the effects of noise in fitness evaluation, a simple problem from learning theory is considered – generalization in a perceptron with binary weights. The perceptron attempts to learn from examples produced by a teacher perceptron, also with binary weights. Baum *et al* show that this problem is similar to the one-max problem corrupted by noise, so long as an independent batch of training examples are presented each time the training error is calculated [4]. This simplifies the analysis considerably, as it avoids overfitting to a particular training set, allowing the dynamics to be solved under the present formalism. A limit is then identified for which the optimal training batch size can be determined.

## 5.2 Selection on a stochastic fitness measure

Let the stochastic relationship between fitness and phenotype be described by a conditional probability distribution $p(F|R)$. If fitness is a deterministic function of the phenotype then this distribution is a delta function, while for noise corrupted fitness the distribution has some variance. Expectation values for the phenotype cumulants after selection can be calculated as described in chapter 3, section 3.2.1, except that the average is now also taken over fitness, which may no longer be a deterministic function of the phenotype. Equation (3.5) provides the result (for $n > 0$),

$$K_n^{\mathrm{s}} = -\lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\int_0^\infty \mathrm{d}t\,\frac{f^P(t,\gamma)}{t} \tag{5.1}$$

where $f(t,\gamma)$ now includes an average over the conditional fitness distribution,

$$f(t,\gamma) = \int \mathrm{d}R\,\mathrm{d}F\,p(R)\,p(F|R)\exp\left(-tw(F)\,\mathrm{e}^{\gamma R}\right) \tag{5.2}$$

Here, $w(F)$ is the selection weight which was introduced in chapter 3, section 3.2. For Boltzmann selection one chooses $w(F) = \exp(\beta F)$.

### 5.2.1 Gaussian noise and Boltzmann selection

Consider the case where fitness is described by a Gaussian distribution centred around the phenotype,

$$p(F|R) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(F-R)^2}{2\sigma^2}\right) \tag{5.3}$$

This is equivalent to directional selection on a phenotype corrupted by additive Gaussian noise with mean zero and standard deviation $\sigma$.

For Boltzmann selection the integrals in equations (5.1) and (5.2) must be computed numerically in general, as for the zero noise case. However, for sufficiently small $\beta\sqrt{K_2+\sigma^2}$ the $1/P$ expansion described in chapter 3, section 3.2.2 is appropriate. In this case, recall that the cumulants after selection are generated by equation (3.9),

$$K_n^{\mathrm{s}} = \lim_{\gamma\to 0} \frac{\partial^n}{\partial\gamma^n}\left[\log(\psi(\gamma,\beta)) - \frac{1}{2P}\left(\frac{\psi(2\gamma,2\beta)}{\psi^2(\gamma,\beta)}\right)\right]$$

where $\psi(\gamma,\beta)$ now includes an average over the noise,

$$\begin{aligned}\psi(\gamma,\beta) &= \int \mathrm{d}R\,\mathrm{d}F\, p(R)\, p(F|R)\, \mathrm{e}^{\beta F+\gamma R} \\ &= \mathrm{e}^{(\beta\sigma)^2/2}\int \mathrm{d}R\, p(R)\, \mathrm{e}^{(\beta+\gamma)R}\end{aligned} \tag{5.4}$$

Using the cumulant expansion in equation (2.8) one finds that the cumulants after selection are given by,

$$K_n^{\mathrm{s}} = \lim_{\gamma\to 0} \frac{\partial^n}{\partial\gamma^n}\left[\sum_{i=1}^{\infty} \frac{(\gamma+\beta)^i K_i}{i!} - \frac{\mathrm{e}^{(\beta\sigma)^2}}{2P}\exp\left(\sum_{i=1}^{\infty}\frac{(2^i-2)(\gamma+\beta)^i K_i}{i!}\right)\right] \tag{5.5}$$

The duplication contribution to the correlation after selection (see chapter 3, section 3.2.3) is similarly found to be,

$$\Delta q_{\mathrm{d}} \simeq \frac{\mathrm{e}^{(\beta\sigma)^2}}{P}\left(1 + K_2\beta^2 - K_3\beta^2 + O(\beta^4)\right) \tag{5.6}$$

The noise increases finite population effects but has no effect on the infinite population result. For zero noise, equations (5.5) and (5.6) reduce to equations (3.20) and (3.22) as expected and the qualitative discussion in sections 3.3.2 and 3.3.3 still holds. Selection introduces higher cumulants into a finite population, which increases convergence under further selection and leads to reduced performance in general. The addition of noise to the fitness measure increases the finite population effects and correspondingly the performance will fall off even more rapidly. For Boltzmann selection and Gaussian noise this is solely a finite population effect.

Under other forms of selection or noise there may also be systematic effects on the infinite population results due to noise [44]. These effects would be much harder to characterize in a simple way, although the present formalism is still able to accurately determine the change in each cumulant under selection. Boltzmann selection provides a particularly transparent model for understanding the effects of Gaussian noise precisely because there are no effects in the infinite population limit.

### 5.2.2 Resizing the population to remove noise

The detrimental effects of Gaussian noise can be removed in the weak selection limit by increasing the population size appropriately,

$$P = P_0 \exp\left[(\beta\sigma)^2\right] \tag{5.7}$$

where $P_0$ is the population size for zero noise. Here, $\beta$ can depend on the phenotype cumulants in an arbitrary way, but must be independent of the noise. Since the other genetic operators do not involve finite population effects, this choice of population size maps the whole dynamics onto the trajectory of a GA without noise and with population size $P_0$. In section 5.4.4 it will be shown how this population resizing allows the optimal batch size to be determined for a simple learning problem.

In the absence of noise, the selection strength should be chosen inversely proportional to the standard deviation of phenotypes within the population (see chapter 3, section 3.3.2). The scaled selection strength is defined $\beta = \beta_{\rm s}/\sqrt{\kappa_2}$, where $\beta_{\rm s}$ is fixed. If this scaling is used and the noise is $O(\sqrt{N})$, then the population size defined above is $O(P_0 {\rm e}^{\beta_{\rm s}})$ and the GA can remove noise without an excessive increase in computation time.

In a more realistic scenario only the measured, noisy fitness would be known. Choosing the selection strength inversely proportional to the standard deviation of fitness (rather than phenotype) leads to the selection strength varying with the level of noise. In this case the population resizing expression in equation (5.7) does not apply, as noise will affect terms other than the finite population corrections to equation (5.5). However, the resizing expression can be applied with any fixed schedule for determining the selection strength. Scaling the selection strength inversely with the standard deviation of phenotypes is equivalent (on average) to choosing the schedule which is most appropriate for a GA without noise and with population size $P_0$. Of course, the results derived here do not depend on any particular scheme for choosing the selection strength.

## 5.3 Noisy one-max

Consider the one-max problem defined in chapter 4, section 4.7. Under noisy fitness evaluation the expressions for crossover and mutation are unchanged, because noise only affects the selection procedure. The expectation values for the cumulants after selection are shown in equation (5.1) and it only remains to calculate the correlation after selection. This calculation almost exactly follows the discussion in section 4.6. The only difference is in the calculation of $\Delta q$, which is defined in equation (4.33), since the averages in equations (4.37a) and (4.37b) now include integrals over the noise distribution. For Gaussian noise and Boltzmann selection the integrals are simply Gaussian integrals and one finds that to $O(1/P)$, $\Delta q$ is increased by a factor of $\mathrm{e}^{(\beta\sigma)^2}$. Notice that this is the same factor which appears in the finite population terms for the cumulants after selection, given in equation (5.5). Recall that this expression is only exact for weak selection and low noise. To improve accuracy in simulations, the term which does not involve the correlation (see equation (4.35a)) can be determined through numerical integration, where now there is also an average over noise.

### 5.3.1 Simulating the dynamics

The dynamics of the GA can be simulated by combining the selection results derived in the preceding sections with the crossover and mutation results derived in chapter 4, sections 4.3 and 4.4. Bit-simulated crossover is used (see the last paragraph of section 4.4 in chapter 4), which allows the dynamics to be described in terms of only two parameters, the mean phenotype and correlation within the population, and therefore avoids the need to follow higher cumulants.

The higher cumulants are still required before selection and these are calculated by the maximum entropy ansatz described in section 4.5. The results presented here can be generalized to describe a GA with uniform crossover using the methods developed in chapter 4.

In figure 5.1 the theoretical results are compared to averaged results from a GA for a typical choice of parameters. Trajectories are shown for the mean and variance of the phenotype distribution. The zero noise case is compared to noisy one-max with $\sigma^2 = 6\kappa_2$ and $\sigma^2 = 12\kappa_2$, showing how increased noise leads to reduced performance. The noise variance was chosen proportional to the phenotypic variance as this provides the most natural units for measuring noise. This may seem a rather artificial choice, although in many situations the noise will fall off as the mean phenotype increases (for example, this is true for the perceptron problem considered in the next section). In view of this, a fixed noise variance might be an equally artificial construction. These considerations are not of critical importance here, however, as the aim is to verify the theoretical model and a more realistic situation is considered in the next section.

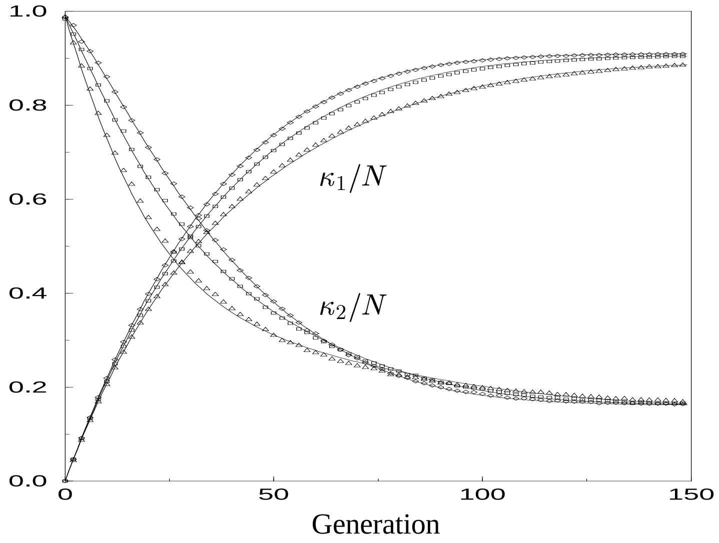

Figure 5.1: The theory for noisy one-max is compared to results averaged over 1000 runs of a GA. The mean ($\kappa_1$) and variance ($\kappa_2$) are shown, with solid lines giving theoretical predictions. The result for zero noise ($\diamond$) is compared to results with additive Gaussian noise of strength $\sigma^2 = 6\kappa_2$ ($\Box$) and $\sigma^2 = 12\kappa_2$ ($\triangle$). The other parameters were $N = 155$, $\beta_{\rm s} = 0.3$, $p_{\rm m} = 0.005$, $P = 100$ and bit-simulated crossover was used.

Notice that the noise variance is significantly greater than the phenotype variance in this example, which emphasizes how robust the GA is even with high levels of noise. For very high levels of noise the theory breaks down however, probably because a weak selection, low noise approximation is required to calculate the duplication contribution to the correlation after selection. There may well be a better approximation for this term, although the approximation used here seems to be accurate for reasonable levels of noise. It may also be the case that when noise levels are high the dynamics do not average well, since there are large fluctuations from mean behaviour.

## 5.4 A simple learning problem

Generalization in a perceptron with binary weights provides a very simple example of a learning problem. The perceptron comprises one computational unit, in this case a McCulloch-Pitts neuron [41], which fires if the summed inputs exceeds some predefined threshold value. The perceptron is trained on examples produced by a teacher perceptron, also with binary weights. This problem has received some considerable attention, including a thermodynamic study of the state space in the limit of large problem size which shows that there is a first order transition to perfect generalization as the number of training examples is increased [26, 62]. The threshold number of training examples at which this transition occurs is $O(N)$ and above this threshold the teacher is the only perceptron compatible with every training example (although a learning algorithm may still fail to find this solution). Below this threshold overfitting is possible, so that although the perceptron learns the training set it does not necessarily generalize well.

Here, the training error (the number of misclassified training examples) is calculated using an independent batch of training examples for each evaluation. This avoids dealing with correlations between a particular training set and perceptrons within the population, which would otherwise make the analysis difficult. The GA will typically require more than $O(N)$ training examples in total and overfitting is not expected to be a problem.

Baum *et al* have shown that this problem is very similar to the noisy one-max problem described in section 5.3 [4]. They analyse a GA which uses a form of truncation selection and show that the computation time of the GA scales as $O(N \log_2^2 N)$ on one-max, if the population size is chosen to be sufficiently large so that the correlation due to duplication is negligible.

They also show that this scaling is not affected when noise with $\sigma \simeq \sqrt{N}$ is introduced into the fitness evaluation. Under the selection method used here, the population improves by the same order each generation as under truncation selection and this algorithm may therefore be expected to scale in the same way as the GA used by Baum *et al*, under similar conditions. The results described here are more general, however, as they do not rely on a large population size and the full dynamical trajectories are predicted.

### 5.4.1 The perceptron

The perceptron has Ising weights $S_i \in \{-1, 1\}$ (encoded in the genotype) and maps an Ising training pattern $\{\zeta_i^\mu\}$ onto a binary output (with zero threshold),

$$O^\mu = \mathrm{sgn}\left(\sum_{i=1}^{N} S_i \zeta_i^\mu\right) \qquad \mathrm{sgn}(x) \;=\; \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{5.8}$$

where $N$ is the number of weights and $\mu$ labels patterns. Let $T_i$ be weights of the teacher perceptron and $S_i$ be weights of the student. The stability of a pattern is a measure of how well it is stored by the perceptron and the stability of pattern $\mu$ for the teacher and student are $\Lambda_{\mathrm{t}}^\mu$ and $\Lambda_{\mathrm{s}}^\mu$ respectively,

$$\Lambda_{\mathrm{t}}^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i \zeta_i^\mu \qquad \Lambda_{\mathrm{s}}^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i \zeta_i^\mu \tag{5.9}$$

The training error will be defined as the number of patterns the student misclassifies,

$$E = \sum_{\mu=1}^{\lambda N} \Theta(-\Lambda_{\mathrm{t}}^\mu \Lambda_{\mathrm{s}}^\mu) \qquad \Theta(x) \;=\; \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{5.10}$$

where $\lambda N$ is the number of training patterns presented. Here, a new batch of training examples are presented each time the training error is calculated. The training error plays the role of an inverse fitness in the GA.

Define the phenotype $R$ to be the overlap between student and teacher. It is possible to choose $T_i = 1$ at each site without loss of generality, in which case $R$ is defined,

$$R = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{5.11}$$

This is simply the phenotype of the one-max problem (normalized to be of order unity).

In order to apply the selection results from section 5.2, it is necessary to find an expression for the training error in terms of the phenotype. If a statistically independent pattern is presented to a perceptron, then for large $N$ the stabilities of the teacher and student are Gaussian variables each with zero mean and unit variance, and with covariance $R$,

$$p(\Lambda_{\rm t}, \Lambda_{\rm s}) = \frac{1}{2\pi\sqrt{1-R^2}} \exp\!\left(\frac{-(\Lambda_{\rm t}^2 - 2R\Lambda_{\rm t}\Lambda_{\rm s} + \Lambda_{\rm s}^2)}{2(1-R^2)}\right) \tag{5.12}$$

The conditional probability distribution for the training error given the overlap is,

$$p(E|R) = \left\langle \delta\!\left(E - \sum_{\mu=1}^{\lambda N} \Theta(-\Lambda_{\rm t}^\mu \Lambda_{\rm s}^\mu)\right)\right\rangle_{\{\Lambda_{\rm t}^\mu, \Lambda_{\rm s}^\mu\}} \tag{5.13}$$

where $\delta(x)$ is the Dirac delta function and brackets denote an average over stabilities distributed according to the joint distribution in equation (5.12). The characteristic function for this distribution is,

$$\begin{aligned} \rho(t\,|R) &= \int \mathrm{d}E\, p(E|R)\, \mathrm{e}^{tE} \\ &= \left\langle \prod_{\mu=1}^{\lambda N} \exp\left[t\Theta(-\Lambda_{\rm t}^\mu \Lambda_{\rm s}^\mu)\right]\right\rangle_{\{\Lambda_{\rm t}^\mu, \Lambda_{\rm s}^\mu\}} \\ &= \left(1 + \frac{1}{\pi}(\mathrm{e}^t - 1)\cos^{-1}(R)\right)^{\lambda N} \end{aligned} \tag{5.14}$$

The logarithm of this characteristic function generates the cumulants of the distribution (see equation (2.7)). The higher cumulants are $O(\lambda N)$ and it turns out that the shape of the distribution is not critical so long as $\lambda$ is $O(1)$. A Gaussian distribution will be a good approximation in this case,

$$p(E|R) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\!\left(\frac{-(E - E_{\rm g})^2}{2\sigma^2}\right) \tag{5.15}$$

where the mean and variance are functions of the overlap between student and teacher,

$$E_{\rm g}(R) = \frac{\lambda N}{\pi} \cos^{-1}(R) \tag{5.16a}$$

$$\sigma^2(R) = E_{\rm g}(R)\left(1 - \frac{E_{\rm g}(R)}{\lambda N}\right) \tag{5.16b}$$

Here, $E_{\rm g}(R)$ is the generalization error, which is the probability of misclassifying a randomly chosen training example, multiplied by the batch size (errors are chosen proportional to $N$

here). The variance expresses the intuition that there is noise in the error evaluation due to the finite size of each training batch.

### 5.4.2 Selection

In the previous section a conditional probability distribution relating the training error (negative fitness) to overlap (phenotype) was derived. The cumulants of the overlap distribution after selection are found from equation (5.1) and the integrals must be calculated numerically in general. All the integrals where computed by Gaussian quadratures in the simulation results presented in section 5.4.5 [51].

For weak selection and large $N$ it is possible to apply the $1/P$ expansion described in chapter 3, section 3.2.2. Since the variance of overlaps within the population is $O(1/N)$ it is reasonable to expand the mean of $p(E|R)$ around the mean of the population in this limit ($R \simeq K_1$). It is also assumed that the variance of $p(E|R)$ is well approximated by its leading term and this assumption may break down if the noise gradient becomes important. Under these simplifying assumptions one finds,

$$E_{\mathrm{g}}(R) \simeq \frac{\lambda N}{\pi}\left(\cos^{-1}(K_1) - \frac{(R-K_1)}{\sqrt{1-K_1^2}}\right) \tag{5.17a}$$

$$\sigma^2 \simeq \frac{\lambda N}{\pi}\cos^{-1}(K_1)\left(1 - \frac{1}{\pi}\cos^{-1}(K_1)\right) \tag{5.17b}$$

Now the problem has been transformed into directional selection corrupted by Gaussian noise, which was described in section 5.2.1. The only significant difference is that here the standard deviation of the noise is a function of the mean overlap (phenotype) within the population. Following the calculation in section 5.2.1 closely, one finds that the cumulants of the overlap distribution after selection are,

$$K_n^{\mathrm{s}} = \lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\left[\sum_{i=1}^{\infty}\frac{(\gamma+k\beta)^i K_i}{i!} - \frac{\mathrm{e}^{(\beta\sigma)^2}}{2P}\exp\left(\sum_{i=1}^{\infty}\frac{(2^i-2)(\gamma+k\beta)^i K_i}{i!}\right)\right] \tag{5.18}$$

where,

$$k = \frac{\lambda N}{\pi\sqrt{1-K_1^2}} \tag{5.19}$$

This is equivalent to selecting on the phenotype directly (see equation (5.5)) where $k\beta$ is the effective selection strength and $\sigma/k$ is the effective standard deviation of the noise.

The calculation for the correlation after selection almost exactly follows the derivation in chapter 4, section 4.6. As in the case of noisy one-max the only difference is in the $\Delta q$ term defined in equation (4.33). Making use of the weak selection, large $N$ approximation for $p(E|R)$ yields the same result as for noisy one-max (see section 5.3) with the effective selection strength and standard deviation defined above.

### 5.4.3 Resizing the population

The noise due to the finite size of each training batch increases the magnitude of detrimental finite population terms in selection. In the limit of weak selection and large problem size discussed in the previous section, this can be compensated for by increasing the population size according to equation (5.7). This expression is valid if the effective selection strength $k\beta$ is independent of batch size (which determines the noise strength). For this to be the case $\beta$ must be chosen proportional to $1/\lambda$, which is the most natural scaling in any case because the training error is proportional to $\lambda$. It is then convenient to rewrite equation (5.7),

$$P = P_0 \exp\left(\frac{\lambda_o}{\lambda}\right) \tag{5.20}$$

where,

$$\lambda_o \;=\; \lambda(\beta\sigma)^2 \;=\; \frac{(\lambda\beta)^2 N}{\pi}\cos^{-1}(K_1)\left(1 - \frac{1}{\pi}\cos^{-1}(K_1)\right) \tag{5.21}$$

Here, $\lambda_o$ is independent of $\lambda$ because of the $\beta$ scaling described above. Choosing $P$ according to this expression removes the effects of noise due to the finite batch size and in principle maps the dynamical trajectory onto the infinite training set dynamics (where $E = E_{\mathrm{g}}(R)$) for a GA with population size $P_0$. Typically $\beta$ is $O(1/\sqrt{N})$ so that the exponent here is of order unity, in which case this population resizing will not blow up with increases in problem size (for fixed $\lambda$). This is consistent with the result due to Baum *et al*, although they provide a rigorous proof for the scaling of their algorithm [4].

Both the selection strength and noise variance will change over time, and it would therefore be necessary to change the population size each generation in order to apply the above expression. However, this is problematic when the population size has to be increased, as this leads to an increased correlation[1]. In this case the dynamics will no longer exactly map onto the infinite training set situation.

Instead of varying the population size, one can fix the population size and vary the size of each training batch. In this case one finds,

$$\lambda = \frac{\lambda_o}{\log(P/P_0)} \tag{5.22}$$

Figure 5.2 shows how choosing the batch size each generation according to this expression leads to the dynamics converging onto the infinite training set trajectory of a GA with a smaller population. The infinite training set result for the largest population size is also shown, as this gives some measure of the potential variability of trajectories available under different batch sizing schemes. Any deviation from the weak selection, large $N$ limit is not apparent here.

In these experiments the effective selection strength was scaled inversely to the standard deviation of the overlap distribution ($\beta = \beta_{\rm s}/k\sqrt{\kappa_2}$). This is a rather artificial choice, as it requires information about the overlap statistics which would not be known in general. However, as discussed in section 5.2.2, the population resizing in equation (5.20) and the corresponding batch sizing expression in equation (5.22) are valid given any fixed schedule for determining selection strength. The choice of selection scaling used here is equivalent (on average) to an appropriate schedule for the infinite training batch problem, but it should be emphasized that these results do not rely on any particular scheme for choosing selection strength (as long as the effective selection strength $k\beta$ is independent of the batch parameter $\lambda$).

### 5.4.4 Optimal batch size

In the previous section it was shown how the population size could be increased in order to remove the effects of noise associated with a finite training batch. Fitzpatrick and Grefenstette also identified the existence of such a tradeoff between population size and batch size, and they suggested that there is often an optimal choice of batch size (or measurement accuracy) [13]. If the population resizing in equation (5.20) is used, then it is possible to identify such an

[1] This is a problem for a real GA which produces a finite population after selection. The theoretical model described in chapter 2, section 2.3 does not have this problem, as the population size is infinite after selection. In a real GA one might overcome this by creating a large but finite population after selection, some members of which could be discarded before the next round of selection.

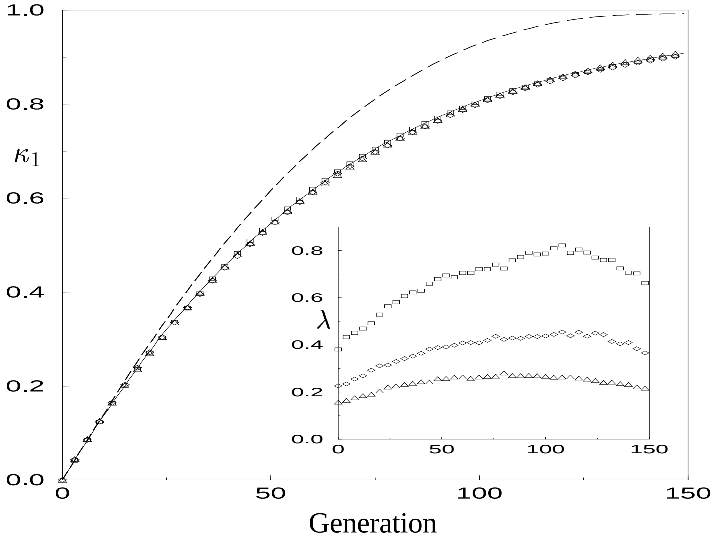

Figure 5.2: The mean overlap between teacher and student within the population is shown each generation, averaged over 100 runs of a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. Training batch sizes were chosen according to equation (5.22), leading to trajectories converging onto the infinite training set result where $E = E_{\rm g}(R)$. The solid curve is for the infinite training set result with $P_0 = 60$ and finite training set results are for $P = 90$ (□), $120(\diamond)$ and $163(\triangle)$. Inset is the mean choice of batch parameter ($\lambda$) each generation. The dashed line is the infinite training set result for $P = 163$, showing that there is significant potential variability of trajectories under different batch sizing schemes. The other parameters were $N = 279$, $\beta_{\rm s} = 0.25$ and $p_{\rm m} = 0.001$.

optimal batch size, which minimizes the computational cost of training error evaluations. This choice of batch size will also minimize the total number of training examples presented when independent batches are used.

It is assumed that computation is mainly due to error evaluation and that other overheads can be neglected. There are $P$ error evaluations each generation with computation time for each scaling as $\lambda$. If the population size each generation is chosen by equation (5.20), then the computation time $\tau_c$ is related to batch size by,

$$\tau_c(\lambda) \;\propto\; \lambda \, \exp\left(\frac{\lambda_o}{\lambda}\right) \tag{5.23}$$

The optimal choice of $\lambda$ is given by the minimum of $\tau_c$, which is at $\lambda_o$ (defined in equation (5.21)). Choosing this batch size leads to the population size being constant over the whole GA run and for optimal performance one should choose,

$$P = P_0\,\mathrm{e}^1 \simeq 2.73P_0 \tag{5.24a}$$

$$\lambda = \lambda_o \tag{5.24b}$$

where $P_0$ is the population size used for the zero noise, infinite training set GA with the same dynamical trajectory. Notice that it is not necessary to determine $P_0$ in order to choose the size of each batch, since $\lambda_o$ is not a function of $P_0$ (see equation (5.21)). One of the runs in figure 5.2 is for this choice of $P$ and $\lambda$, showing close agreement to the infinite training set result ($P = 163 \simeq P_0\mathrm{e}$).

Unfortunately, the optimal batch size is a function of the mean overlap within the population, which would not be known in general (although it could be estimated from training error statistics). However, the initial optimal batch size provides an upper bound, since $\sigma^2$ is a monotonically decreasing function of the mean overlap. Setting $K_1 = 0$ in equation (5.21) provides this bound,

$$\lambda_o \leq \tfrac{1}{4}(\lambda\beta)^2 N \tag{5.25}$$

Recall that $\beta$ is proportional to $1/\lambda$, so that the right hand side of this expression is independent of $\lambda$. The selection strength is typically $O(1/\sqrt{N})$ and the optimal batch size is therefore typically $O(N)$. This is a somewhat intuitive result, as it shows how more effort should be expended in determining fitness (through increasing the batch size) when the resulting decisions are more critical (through stronger selection).

Statistics describing the overlap distribution change in a non-trivial manner each generation and their evolution can be determined by simulating the dynamics, as described in the next section.

### 5.4.5 Simulating the dynamics

The dynamics can be modelled by combining the selection results from section 5.4.2 with the expressions for mutation and crossover derived in chapter 4, sections 4.3 and 4.4. Bit-simulated crossover was used, as this allows the dynamics to be described in terms of the mean overlap and correlation alone, which simplifies the selection numerics and avoids the need to follow higher cumulants. Although the dynamics only require the iteration of expressions for these two macroscopics, the higher cumulants are required before selection and these are obtained from the maximum entropy ansatz described in section 4.5. These results can be generalized to other forms of crossover by the methods developed in chapter 4.

Figure 5.3 shows the averaged trajectories of the mean and variance of the overlap distribution and figure 5.4 shows the overlap of the best solution, for a typical choice of search parameters. The infinite training batch result, where $E = E_{\mathrm{g}}(R)$, is compared to results for two fixed batch sizes, showing how performance degrades as the batch size is reduced. The theoretical curves agree well, although there is a sight under-estimate in the maximum overlap towards the end of each run, possibly for the reasons discussed in chapter 2, section 2.5. There is also a slight systematic error in the curves for the smallest batch size. As the batch size is reduced further the theory breaks down. This is mostly for the same reasons as discussed in section 5.3.1. The duplication contribution to the increased correlation after selection required the use of a weak selection, large $N$ approximation and the dynamics may not average well when fluctuations from mean behaviour increase. It is also possible that the Gaussian approximation for $p(E|R)$ breaks down for small $\lambda$, in which case it would be necessary to expand the noise in terms of more cumulants.

## 5.5 Conclusion

The selection calculation has been extended to describe a stochastic fitness measure. This was motivated by the observation that there may be noise in the evaluation of fitness for a number of optimization and machine learning problems. A result was derived for the expected phenotype cumulants after selection given a general selection scheme and an arbitrary stochastic fitness measure. For weak Boltzmann selection and additive Gaussian noise it was possible to write down the result for each cumulant after selection in closed form. In this limit a simple increase in population size removes the effects of noise in every cumulant and in the duplication contribution to the correlation after selection. The theory agreed well with averaged results from a GA for the one-max problem corrupted by Gaussian noise.

To show how this work may be relevant to machine learning, a simple learning problem

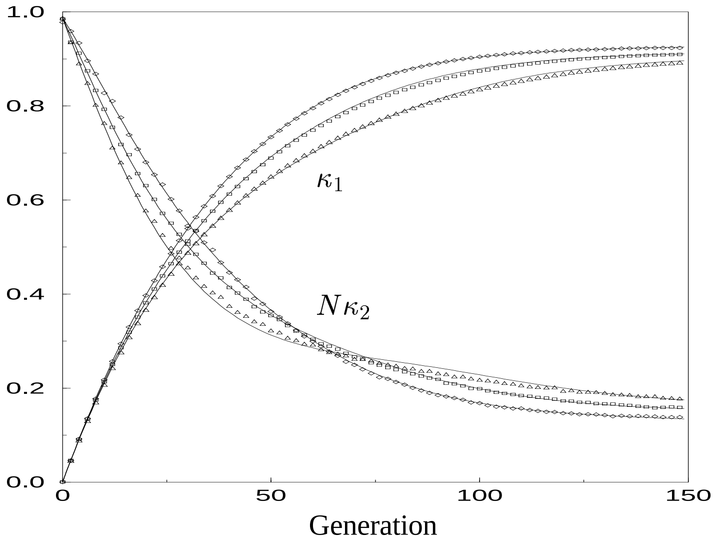

Figure 5.3: The theory is compared to averaged results from a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. The mean ($\kappa_1$) and variance ($\kappa_2$) of the overlap distribution are shown averaged over 1000 runs, with the solid lines showing theoretical predictions. The infinite training set result (◇) is compared to results for a finite training set with $\lambda = 0.65$ (□) and $\lambda = 0.39$ (△). The other parameters were $N = 155$, $\beta_{\rm s} = 0.3$, $p_{\rm m} = 0.005$, $P = 80$ and bit-simulated crossover was used.

was introduced – generalization in a perceptron with binary weights. The perceptron learns from examples produced by a teacher with the same architecture. To simplify matters, a new batch of training examples were chosen each time the training error was calculated. In this case the training error is a random variable distributed around the generalization error. For large problem size the training error distribution was shown to be well approximated by a Gaussian distribution, whose effective variance increases as the training batch size is reduced. The full dynamics was simulated by following the distribution of overlaps between the teacher and perceptrons within the population. The theory agreed closely with averaged results from a GA for a number of batch sizes. In the limit of weak Boltzmann selection and large problem size it was shown how the population size could be chosen each generation in order to remove the detrimental effects of noise due to the finite size of each training batch. This population resizing was then used to determine the optimal batch size each generation, which minimized computation time as well as the total number of training examples required.

It might be instructive to extend this work in a number of directions. The binary perceptron problem required a new batch of training examples for each error evaluation and it would be

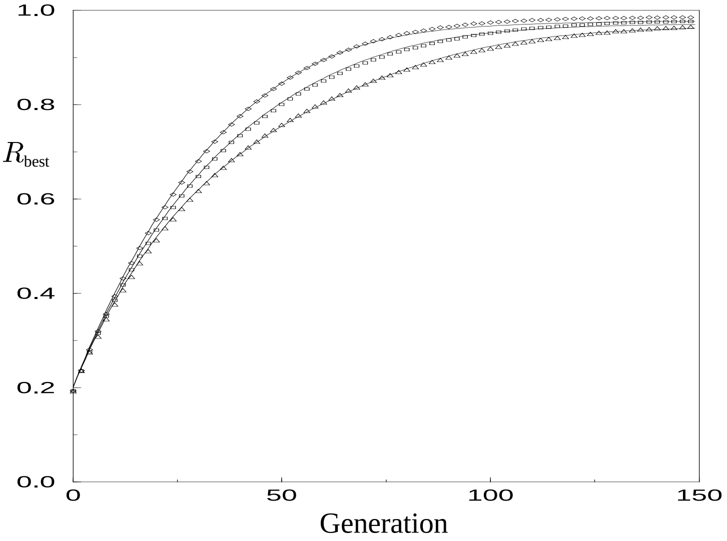

Figure 5.4: The maximum overlap between teacher and student is shown each generation, averaged over the same runs as the results presented in figure 5.3. The solid lines show theoretical predictions and the symbols are as in figure 5.3.

interesting to consider the case were batches are recycled, leading to the possibility of over-fitting. One could also consider a multi-layer perceptron, in which case the phenotype might be a vector of order parameters describing overlaps between different nodes within the teacher and student. This would be especially interesting as the GA would have to break symmetry within the space of solutions and this symmetry breaking would have to be incorporated by the theory. It might then be interesting to compare the dynamics of the GA with on-line gradient descent in networks with continuous weights, for which closed form expressions describing the dynamics have recently been obtained [59]. There are many other situations where the fitness measure has a stochastic component and it is hoped that the results described in this chapter will provide a framework for analysing such problems.

# Chapter 6

# Attempting a strong NP-hard problem

## 6.1 Introduction

Although a variety of problems have been considered under the present formalism, these have so far only come from the rather restricted class where alleles of the genotype contribute additively to the phenotype. An interesting question is to ask how far the formalism can move beyond this restriction, in order that it may describe truly hard problems. In this chapter the formalism is applied to the problem of storing random binary patterns in a perceptron with binary weights. This problem is NP-hard in the strong sense if the number of patterns is proportional to the number of weights and no algorithm exists which can solve it in polynomial time [50]. It is an appropriate problem to study because the GA finds optimal solutions with reasonable efficiency, although simulated annealing seems to be the most effective algorithm to date [35, 48, 55]. The perceptron is also naturally encoded as a binary vector, so there are no representational difficulties. This is in marked contrast to the travelling salesman problem, which is one of the most commonly used NP-hard bench marks.

Although no solution is found for the dynamics of the GA in general, the effect of mutation can be accurately modelled under certain assumptions. The most important assumption is that individuals within the population are equally likely to take any configuration given their particular training error. That is, the state space is not biased towards a particular kind of solution. Of course, the population correlates under selection and this is a potential source of bias, but because mutation does not involve interactions between different individuals this effect is not necessarily critical. The assumption of an unbiased population allows the cumulants after mutation to be calculated in the limit of large problem size, using the replica method to average over random disorder in the training patterns. For low capacity the replica symmetric solution reduces to the much simpler annealed result, which was previously derived in reference [55] for the simplest error measure considered here. This limit allows closed form results for mutation.

Unfortunately, the assumption that taking an unbiased average should accurately model mutation in general is shown to be unjustified. History effects play an important role in the dynamics of mutation, so that the training error alone is insufficient to accurately characterize a perceptron configuration. This seems to be most important for the simplest training error, which is the number of misclassified examples. For a training error which also includes information about the stability of each unstored example the theory is shown to characterize the mean change under mutation more accurately, although not perfectly. Given this evidence, it seems unlikely that an accurate characterization of mutation is possible in general without including some extra features into the theory.

## 6.2 Storing random patterns in a binary perceptron

The perceptron was introduced in chapter 5, section 5.4, where the problem of learning patterns produced by a teacher perceptron was considered. Here, the perceptron attempts to store a set of random and uncorrelated binary mappings. Recall the definition of the perceptron given in equation (5.8). The condition for pattern $\mu$ to be stored is,

$$O^{\mu}\sum_{i=1}^{N} S_i\zeta_i^{\mu} \geq 0 \tag{6.1}$$

where patterns map an Ising vector with components $\zeta_i^{\mu} \in \{-1, 1\}$ onto a single Ising output $O^{\mu} \in \{-1, 1\}$. The role of a training algorithm is to find the weights which satisfy this inequality for as many patterns as possible. Since the patterns are randomly generated binary vectors, a trivial gauge transformation can be applied without changing the nature of the problem,

$$\xi_i^{\mu} = O^{\mu}\zeta_i^{\mu} \tag{6.2}$$

Here, $\xi_i^{\mu} \in \{-1, 1\}$ is also a random Ising spin which satisfies the following conditions,

$$\lim_{N\to\infty}\langle\xi_i^{\mu}\rangle_i = 0 \qquad \lim_{N\to\infty}\langle\xi_i^{\mu}\xi_i^{\nu}\rangle_i = \delta^{\mu\nu} \tag{6.3}$$

where brackets denote site averages and $\delta^{\mu\nu}$ is the Kronecker delta,

$$\delta^{\mu\nu} = \begin{cases} 1 & \mu = \nu \\ 0 & \mu \neq \nu \end{cases} \tag{6.4}$$

Often, the patterns are required to have a finite basin of attraction, in which case the stability of each pattern $\Lambda^{\mu}$ is required to exceed some threshold $\mathcal{T}$,

$$\Lambda^{\mu} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} S_i\xi_i^{\mu} \geq \mathcal{T} \tag{6.5}$$

Here, the factor of $1/\sqrt{N}$ is chosen to make the stability of order unity in the typical case. It has been shown that the optimal threshold for learning is often greater than the threshold required at retrieval [34, 48, 55].

### 6.2.1 Training error

It is necessary to define a training error which plays the role of negative fitness in the GA. There is no simpler phenotype available from which the training error can be derived and it is therefore necessary to model the distribution of errors directly. For the generalization problem introduced in chapter 5, the number of incorrectly classified patterns was used. Storage is often a harder problem for the learning algorithm because the patterns are completely random and uncorrelated, and therefore contain less structure than those produced by a teacher. To store $O(N)$ random patterns it seems to be necessary to include some information about how far each pattern is from being stored. One form of training error which incorporates this feature is defined by,

$$E = \sum_{\mu=1}^{\lambda N} u_l(\mathcal{T} - \Lambda^\mu) \quad \text{where} \quad u_l(x) = x^l \Theta(x) \tag{6.6}$$

Here, $\lambda N$ is the total number of pattern being stored and $\lambda$ is called the capacity[1]. For $l = 0$ this training error reduces to the number of misclassified patterns and this will be called the step error. With $l = 2$ this is the error used in the most successful algorithm to date, which is a simulated annealing procedure due to Patel [48]. This will be called the summed square error. The whole set of patterns is presented to the GA each time the training error is calculated.

### 6.2.2 Storage capacity

Krauth and Mézard have determined the critical capacity of the binary perceptron in an extension to Gardner's seminal work on the perceptron with continuous weights [15, 16, 37]. They find that for random patterns a perceptron can store up to $\lambda_c N$ patterns (for large $N$), where $\lambda_c \simeq 0.83$ is the critical capacity. This result has been confirmed numerically to within a high degree of accuracy [48]. They employed the replica method, which is often used in statistical

[1] The capacity is usually denoted $\alpha$, but $\lambda$ is used here to avoid any ambiguity.

mechanics to average over quenched random disorder. It is instructive to consider their work here, as the calculation is closely related to the mutation calculation described in section 6.3.

In an ingenious formalism, Gardner showed how one could average over the configuration space of the perceptron in order to calculate the number of states compatible with a set of training examples. The volume of the configuration space compatible with the condition in equation (6.5) is given by,

$$\Omega = \left\langle \prod_{\mu=1}^{\lambda N} \Theta(\Lambda^{\mu} - \mathcal{T}) \right\rangle_{\{S_i\}} \tag{6.7}$$

where the brackets denote an average over all weight configurations. The logarithm of this volume corresponds to the entropy of configurations, which is assumed to be a self-averaging quantity. The patterns are quenched, or fixed, random vectors and the average over this randomness can only be taken over a self-averaging quantity. This is the familiar problem in statistical mechanics of averaging over a logarithm.

To compute the average over the logarithm the replica method is used (see, for example, reference [43]). This makes use of the following identity,

$$\langle \log \Omega \rangle = \lim_{n \to 0} \frac{\langle \Omega^n \rangle - 1}{n} \tag{6.8}$$

where brackets denote an average over the quenched patterns. The method assumes validity of the analytical continuation from positive integer values of $n$ through the reals to zero. The right hand side of equation (6.8) can be calculated for integer $n$ by making $n$ replicas of the system,

$$\begin{aligned} \langle \Omega^n \rangle &= \left\langle \prod_{\alpha=1}^{n} \Omega_{\alpha} \right\rangle \\ &= \left\langle \prod_{\alpha=1}^{n} \left\langle \prod_{\mu} \Theta\Big(\sum_i S_i^{\alpha} \xi_i^{\mu} - \mathcal{T}\sqrt{N}\Big) \right\rangle_{\{S_i^{\alpha}\}} \right\rangle_{\{\xi_i^{\mu}\}} \end{aligned} \tag{6.9}$$

where $\alpha$ labels replicas. The inner average is over the weight configuration for each replica while the outer average is over the quenched patterns. The calculation can now be completed by commuting the order of averaging and using the saddle point method in the limit of large $N$ [40]. Some care must be taken when exchanging the order of the limits $n \to 0$ and $N \to \infty$. This exchange of limits can be justified in the closely related SK spin-glass problem and is thought to also be valid here [28].

To complete the calculation, some assumption has to be made about the relationship between replicas. The simplest assumption is to assume symmetry between replicas. In this case the order parameter describing the correlation between replicas takes a single value. For continuous weights the correlation approaches one and the entropy vanishes as the capacity increases to the critical capacity. In this case the replica symmetric ansatz is consistent, although only up until this point [6, 15]. For Ising weights the entropy vanishes before the replica symmetric correlation reaches one and then becomes negative, indicating an unphysical interpretation. In fact, Krauth and Mézard show that consistent results are obtained by one step of symmetry breaking according to Parisi's ansatz [37]. The replica symmetry breaking occurs at the critical capacity $\lambda_c \simeq 0.83$ where the replica symmetric entropy vanishes. More interesting behaviour is observed by introducing a temperature and moving into the canonical ensemble. This leads to a physical interpretation of replica symmetry breaking in terms of ergodicity breaking, where many meta-stable states are formed whose escape times diverge with the problem size.

A dynamical study by Horner shows that simulated annealing encounters meta-stable states for all capacities of $O(N)$, which is compatible with the problem being NP-hard for all capacities of this order [34]. He concludes that the replica treatment, which is essentially an equilibrium thermodynamics approach, is not sufficient to capture all of the interesting dynamical features of the training algorithm. This also provides some motivation for studying the dynamics of search by other learning algorithms, such as the GA. It has been proposed that the GA may be able to avoid the meta-stable states which trap local search procedures, although simulated annealing has proved to be the more successful algorithm to date [35, 48, 55].

Replica symmetry is thought to hold for capacities right up until the critical capacity in this problem. In the following section the mutation calculation will be carried out under the assumption that $\lambda \ll \lambda_c$ and the replica symmetric assumption is assumed to be valid.

## 6.3 Microcanonical mutation calculation

From chapter 2, section 2.3, recall that under the present formalism mutation can be carried out within the infinite population produced by selection. To calculate the effect of mutation on the distribution of training errors within this population, it is first necessary to determine a conditional probability for the training error after mutation in terms of the error before, $p(E_{\rm m}|E)$. The distribution of errors within an infinite population after mutation is then given by,

$$p(E_{\rm m}) = \int {\rm d}E \, p(E) p(E_{\rm m}|E) \tag{6.10}$$

The cumulants after mutation can be obtained from the characteristic function of this distribution. To calculate $p(E_{\rm m}|E)$ it is necessary to make some assumption about the microscopic configuration of perceptrons within the population. It will be assumed that configurations are typical of perceptrons with a given training error, in which case $p(E_{\rm m}|E)$ can be computed by an unbiased average over the entire configuration space. This is essentially maximizing entropy with a constraint on individual configurations rather than the entire population, and corresponds closely to the microcanonical formulation of statistical mechanics.

Let $\Lambda^\mu$ and $\Lambda^\mu_{\rm m}$ be the stability of pattern $\mu$ before and after mutation respectively,

$$\Lambda^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i \xi_i^\mu \tag{6.11}$$

$$\Lambda^\mu_{\rm m} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} M_i S_i \xi_i^\mu \qquad M_i = \begin{cases} 1 & \text{with probability } 1 - p_{\rm m} \\ -1 & \text{with probability } p_{\rm m} \end{cases} \tag{6.12}$$

Here, $M_i$ are random variables which determine the probability of a weight being flipped under mutation. Recall the definition of training error given in equation (6.6). If the distribution of weight configurations is unbiased, then the conditional probability $p(E_{\rm m}|E)$ is given by,

$$\begin{aligned} p(E_{\rm m}|E) &= \frac{\Omega(E_{\rm m}, E)}{\Omega(E)} \\ &= \frac{\langle \delta(E_{\rm m} - \sum_\mu u_l(\mathcal{T} - \Lambda^\mu_{\rm m}))\, \delta(E - \sum_\mu u_l(\mathcal{T} - \Lambda^\mu)) \rangle_{\{S_i, M_i\}}}{\langle \delta(E - \sum_\mu u_l(\mathcal{T} - \Lambda^\mu)) \rangle_{\{S_i\}}} \end{aligned} \tag{6.13}$$

where $\delta(x)$ is the Dirac delta function and the angled brackets denote an average over all weight configurations and mutation variables. It will be assumed that the cumulants of this distribution are self-averaging. The cumulants are generated from the logarithm of a characteristic function (see equation (2.7)) which is defined,

$$\begin{aligned}\rho(t\,|E) &= \int \mathrm{d}E_{\mathrm{m}}\, p(E_{\mathrm{m}}|E)\, \mathrm{e}^{tE_{\mathrm{m}}} \\ &= \frac{\langle \mathrm{e}^{t\sum_\mu u_l(\mathcal{T}-\Lambda^\mu_{\mathrm{m}})}\, \delta(E-\sum_\mu u_l(\mathcal{T}-\Lambda^\mu))\rangle_{\{S_i,M_i\}}}{\langle \delta(E-\sum_\mu u_l(\mathcal{T}-\Lambda^\mu))\rangle_{\{S_i\}}} \\ &= \frac{\rho(t,E)}{\rho(0,E)} \qquad (6.14)\end{aligned}$$

where $\rho(t,E)$ is the characteristic function of the joint distribution for the training error before and after mutation. Taking the logarithm decouples the fraction so that it is only necessary to average the logarithm of the numerator (the logarithm of the denominator is retrieved for $t=0$).

### 6.3.1 Replica symmetric result

Recall the replica trick, which made use of the identity in equation (6.8).

$$\log \rho(t,E) = \lim_{n\to 0} \frac{\rho^n(t,E)-1}{n} \qquad (6.15)$$

Writing the power as a product over replicas one finds,

$$\rho^n(t,E) = \prod_{\alpha=1}^{n} \left\langle \mathrm{e}^{t\sum_\mu u_l(\mathcal{T}-\Lambda^{\mu\alpha}_{\mathrm{m}})}\, \delta(E-\sum_\mu u_l(\mathcal{T}-\Lambda^{\mu\alpha})) \right\rangle_{\{S_i^\alpha,M_i^\alpha\}} \qquad (6.16)$$

Now the average over quenched patterns can be computed by making a replica symmetric ansatz, as shown in appendix B. The calculation is for large $N$ and relies on the mutation probability being of order unity in this limit, which is unfortunate as a smaller probability is often used in practice. It is unclear how well this result approximates the effects of weaker mutation, although any differences are probably manifested in the higher cumulants.

Eventually one finds (ignoring irrelevant multiplicative constants),

$$\rho^n(t,E) \;=\; \exp\!\big(-n\nu E + \tfrac{1}{2}nN\phi q + \lambda N G_0 + N G_1\big) \qquad (6.17)$$

Here, $G_1$ and $G_0$ are defined in equations (B.14) and (B.17) respectively,

$$\begin{aligned} G_1 &= n\int Du\, \log\!\big(2\cosh(u\sqrt{\phi})\big) - \frac{n\phi}{2} \qquad (6.18)\\ G_0 &= n\int D\eta_x\, D\eta_z\, D\eta_{xz}\, \log\Bigg[\bigg(\int_{-\infty}^{\mathcal{T}} \mathrm{d}\epsilon_t\, \mathrm{e}^{t(\mathcal{T}-\epsilon_t)^l} + \int_{\mathcal{T}}^{\infty} \mathrm{d}\epsilon_t\bigg) \\ &\qquad \times \bigg(\int_{-\infty}^{\mathcal{T}} \mathrm{d}\epsilon_\nu\, \mathrm{e}^{\nu(\mathcal{T}-\epsilon_\nu)^l} + \int_{\mathcal{T}}^{\infty} \mathrm{d}\epsilon_\nu\bigg) \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}x\mathrm{d}z}{-4\pi^2} \exp(F)\Bigg] \qquad (6.19)\end{aligned}$$

where,

$$F = x\epsilon_\nu + z\epsilon_t + \tfrac{1}{2}(1-q)(x^2+z^2+2\Gamma xz) + \eta_{xz}\sqrt{q\Gamma}(x+z) + \sqrt{q(1-\Gamma)}(x\eta_x + z\eta_z)$$

Here, $\Gamma = 1 - 2p_m$. The saddle point equations fix the values of $\nu$, $q$ and $\phi$,

$$E = N\lambda\frac{\partial}{\partial\nu}\left(\frac{G_0}{n}\right) \tag{6.20a}$$

$$\frac{\phi}{2} + \lambda\frac{\partial}{\partial q}\left(\frac{G_0}{n}\right) = 0 \tag{6.20b}$$

$$\frac{q}{2} + \frac{\partial}{\partial\phi}\left(\frac{G_1}{n}\right) = 0 \tag{6.20c}$$

In general, these expressions are rather unwieldy and would require numerical enumeration in many cases, even to first order in $\lambda$. Rather than continuing with the most general situation, it is more instructive to consider a much simpler limit.

### 6.3.2 Low capacity limit

From equations (6.20b) and (6.20c) one can show that as $\lambda$ becomes small, $q$ and $\phi$ are both proportional to $\lambda$. For sufficiently small $\lambda$ it is then reasonable to take $q = 0$ and $\phi = 0$. In this case the replica symmetric result reduces to the annealed result, which was previously calculated for the step error ($l = 0$) in reference [55]. Although the summed square error ($l = 2$) is a more useful choice in practice, the step error provides a simple measure with which to test the theoretical results. Extensions to other values of $l$ should be possible in principle, as results up till this point have been for general $l$. The annealed result corresponds to averaging $\rho(t, E)$ directly over patterns, rather than averaging the logarithm, which is expected to be incorrect in general because unusual pattern configurations will dominate the average and give untypical results.

With $q$ and $\phi$ equal to zero, the expression for the characteristic function is much simplified,

$$\log\rho(t,E) = -\nu E + \lambda N G_{\rm ann} \tag{6.21}$$

where $G_{\rm ann}$ is the annealed equivalent of $G_0$ in the replica symmetric expression,

$$G_{\rm ann} = \log\Bigg[\left(\int_{-\infty}^{\mathcal{T}} \mathrm{d}\epsilon_t \mathrm{e}^{t(\mathcal{T}-\epsilon_t)^l} + \int_{\mathcal{T}}^{\infty}\mathrm{d}\epsilon_t\right)\left(\int_{-\infty}^{\mathcal{T}} \mathrm{d}\epsilon_\nu \mathrm{e}^{\nu(\mathcal{T}-\epsilon_\nu)^l} + \int_{\mathcal{T}}^{\infty}\mathrm{d}\epsilon_\nu\right) \times \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{\mathrm{d}x\mathrm{d}z}{-4\pi^2}\exp\big(x\epsilon_\nu + z\epsilon_t + \tfrac{1}{2}(x^2+z^2+2\Gamma xz)\big)\Bigg] \tag{6.22}$$

Now only one saddle point equation is required to fix $\nu$ as a function of $t$ and $E$,

$$E = N\lambda \frac{\partial G_{\mathrm{ann}}}{\partial \nu} \tag{6.23}$$

In the next two sections the step error and the summed square error are considered. The former measure is the simplest and the the first four cumulants of the population after mutation can be calculated. The latter case is more involved and only the mean error within the population after mutation is calculated here.

### 6.3.3 Step error ($l = 0$)

For $l = 0$ the integrals in equation (6.22) are standard integrals and for $\mathcal{T} = 0$ one finds [55],

$$G_{\mathrm{ann}} = \log\left[\tfrac{1}{2}(1+\mathrm{e}^{\nu}) + (\mathrm{e}^{t}-1)\left(\tfrac{1}{2}\mathrm{e}^{\nu} + \frac{(1-\mathrm{e}^{\nu})}{2\pi}\tan^{-1}\left(\frac{\sqrt{1-\Gamma^2}}{\Gamma}\right)\right)\right] \tag{6.24}$$

The saddle point equation (6.23) fixes $\nu$ as a function of $t$ and $E$.

The cumulants of the error distribution after mutation are generated from the characteristic function of the error distribution (see equation (2.7)),

$$\begin{aligned} \rho_{\mathrm{m}}(t) &= \int \mathrm{d}E_{\mathrm{m}}\, p(E_{\mathrm{m}})\, \mathrm{e}^{tE_{\mathrm{m}}} \\ &= \int \mathrm{d}E\, p(E)\, \rho(t\,|E) \\ &= \int \mathrm{d}E\, p(E)\, \frac{\rho(t,E)}{\rho(0,E)} \end{aligned} \tag{6.25}$$

Here, $\rho(t,E)$ is defined by equations (6.21) and (6.24). To complete this integral one can represent $p(E)$ as a Fourier transform,

$$p(E) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}k}{2\pi\mathrm{i}} \exp\left(\sum_{n} \frac{k^n}{n!} K_n - kE\right) \tag{6.26}$$

where $K_n$ is the $n$th cumulant of the training error distribution before mutation. Substituting this expression into equation (6.25) allows the integrals over $E$ and $k$ to be computed for large $N$ by the saddle point method, as long as the cumulants are $O(N)$. The calculation can be completed by expanding the relevant parameters in $t$, as described in reference [55]. This is appropriate for determining the cumulants after mutation, which are given by the coefficients of the expansion of $\log \rho_{\rm m}(t)$. For the first four cumulants one finds[2],

$$
\begin{aligned}
K_1^{\rm m} &= (1-2\Delta)K_1 + \Delta\lambda N &\quad (6.27a)\\
K_2^{\rm m} &= (1-2\Delta)^2 K_2 + \Delta(1-\Delta)\lambda N &\quad (6.27b)\\
K_3^{\rm m} &= (1-2\Delta)^3 K_3 + \Delta(1-\Delta)(1-2\Delta)(\lambda N - 2K_1) &\quad (6.27c)\\
K_4^{\rm m} &= (1-2\Delta)^4 K_4 - \Delta(1-\Delta)\left[8K_2(1-2\Delta)^2 - \lambda N(1-6\Delta+6\Delta^2)\right] &\quad (6.27d)
\end{aligned}
$$

where,

$$\Delta = \frac{1}{\pi}\tan^{-1}\left(\frac{\sqrt{1-\Gamma^2}}{\Gamma}\right) \tag{6.28}$$

and $\Gamma = 1 - 2p_m$.

It is interesting to compare these results to the expressions for the additive problems which were introduced in chapter 4 (see equations (4.6a) to (4.6d)). The expressions here are very similar, where $\Delta$ corresponds closely to $p_m$ in the additive genotype results. There is an exact correspondence for the first two cumulants if the expressions are rewritten in terms of cumulants from a population of random configurations (the fixed point of mutation). Notice that $\Delta \simeq 2\sqrt{p_m}/\pi$ to first order, for small mutation rates. Typically $p_m \ll 1$ and $\sqrt{p_m} \gg p_m$, so that the effect of mutation is clearly much greater here than for the additive problems. If the benefit of mutation is to increase diversity without too much cost in terms of lost fitness then this is a significant penalty, as the reduced correlation within the population due to mutation is independent of the particular problem under consideration. However, the conclusions which can be derived from these results are severely limited, because they do not describe mutation accurately in general (this will be demonstrated in section 6.4).

### 6.3.4 Summed square error ($l = 2$)

The most successful choice of training error to date is given by equation (6.6) with $l = 2$ [48]. Again, the integrals in equation (6.22) are standard integrals, but the final expression for $G_{\rm ann}$ is more complex than for the step error. For $\mathcal{T} = 0$ one finds,

$$G_{\rm ann} = \log\left[I(0,0,\Gamma) + I(t,0,-\Gamma) + I(0,\nu,-\Gamma) + I(t,\nu,\Gamma)\right] \tag{6.29}$$

[2] This calculation was automated using *Mathematica*, a symbolic programing language [76].

where,

$$I(t,\nu,\Gamma) = \int_0^\infty \mathrm{d}\epsilon_t \, \mathrm{e}^{t\epsilon_t^2} \int_0^\infty \mathrm{d}\epsilon_\nu \, \mathrm{e}^{\nu\epsilon_\nu^2} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}x\mathrm{d}z}{-4\pi^2} \exp\big(x\epsilon_\nu + z\epsilon_t + \tfrac{1}{2}(x^2+z^2+2\Gamma xz)\big)$$
$$= \frac{1}{2\sqrt{1-2\nu-2t(1-2\nu(1-\Gamma^2))}} \left[1 - \frac{1}{\pi}\tan^{-1}\!\left(\frac{\sqrt{1-\Gamma^2}}{\Gamma}\sqrt{1-2\nu-2t(1-2\nu(1-\Gamma^2))}\right)\right]$$

The saddle point equation (6.23) fixes $\nu$ as a function of $t$ and $E$.

In principle, the cumulants after mutation can be computed by the same methods discussed in the previous section. Unfortunately, the resulting series expansions soon become rather cumbersome and a number of terms seem to require a numerical solution. The calculation for the mean error after mutation is straightforward, however, as this only requires the solution of the saddle point equation for $t = 0$.

The expectation value for the error after mutation is given by,

$$\begin{aligned} \langle E_\mathrm{m} \rangle &= \lim_{t\to 0} \frac{\mathrm{d}}{\mathrm{d}t} \log \rho(t,E) \\ &= \lim_{t\to 0} \lambda N \frac{\partial G_\mathrm{ann}}{\partial t} \end{aligned} \tag{6.30}$$

Differentiating out one finds,

$$\begin{aligned} \langle E_\mathrm{m} \rangle = E\left(x^2(1-\Gamma^2)+\Gamma^2\right)\left(1 - \frac{1}{\pi}\tan^{-1}\!\left(x\frac{\sqrt{1-\Gamma^2}}{\Gamma}\right)\right) \\ + \frac{x\lambda N}{\pi(1+x)}\left(\tan^{-1}\!\left(\frac{\sqrt{1-\Gamma^2}}{\Gamma}\right) - \Gamma\sqrt{1-\Gamma^2}\left(1-\frac{1}{x^2}\right)\right) \end{aligned} \tag{6.31}$$

where $x = \sqrt{1-2\nu}$ and the saddle point equation fixes $x$ as a function of $E$,

$$E = \frac{\lambda N}{x^2(1+x)} \tag{6.32}$$

The expression for the mean error after mutation for large $N$ is simply found by replacing $E_\mathrm{m}$ and $E$ by $K_1^\mathrm{m}$ and $K_1$ respectively.

## 6.4 How good was the assumption ?

In the preceding sections mutation expressions were derived by a microcanonical formulation which involved averaging over all configurations with a given training error. Unfortunately, the assumption that this form of averaging is appropriate appears to be false in most cases. Comparing the theoretical predictions for the effect of mutation in a GA to simulation results shows only qualitative agreement. A simple experiment clearly shows the discrepancy between theory and simulations.

### 6.4.1 Mutating away from an unbiased sample

A random sample of configurations are created, whose training error lies below a pre-determined threshold. This serves as an unbiased population whose cumulants can be measured. This population then undergoes repeated mutations with a fixed mutation rate. Any theoretical model of mutation should certainly be able to describe this situation accurately.

Figure 6.1 shows averaged results from this experiment for the step error. The first two cumulants are shown and solid lines give the theoretical predictions according to the expressions in equations (6.27a) and (6.27b). As expected, the theory accurately describes the behaviour for the first generation since the population is initially an unbiased sample. After this, however, the theory and simulation results diverge. The experiment was repeated for a range of $\lambda$ values in order to ensure that there was no significant error due to the small $\lambda$ approximation. Clearly, the history of the population is important. Configurations within the population are no longer typical of configurations with a given training error even after only one generation of mutation.

Figure 6.2 shows the same experiment for the summed square error measure and although the theory gives a better prediction here, there is still significant deviation from the experimental result. One explanation for the better agreement in this case is that the summed square error measure contains information about the stability of unstored patterns. This measure therefore provides a more constrained characterization of configurations, so that the averages in equation (6.13) are more representative than for the simpler step error. Unfortunately, only the change in mean has been determined so far for this training error.

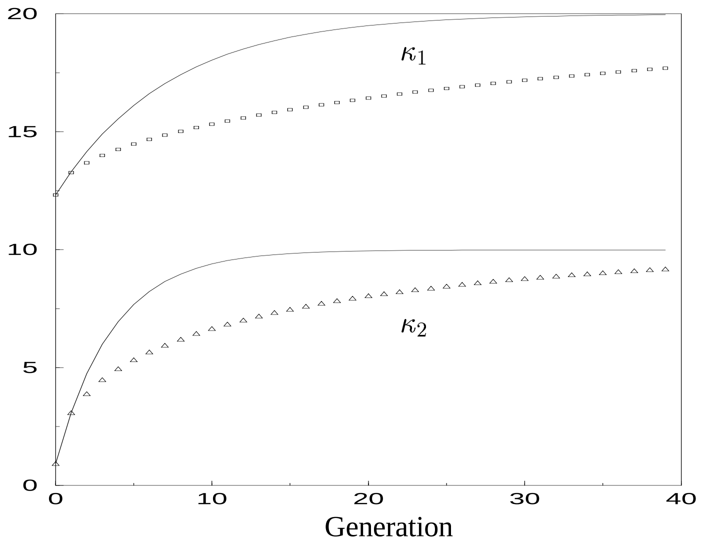

Figure 6.1: A population of $1000$ randomly generated configurations with step error below $\lambda N/3$ undergoes repeated generations of mutation with $p_{\mathrm{m}} = 0.01$. The mean (□) and variance (△) of the step error are shown each generation, averaged over $500$ samples. Solid curves show the prediction from the microcanonical theory. The problem size was $N = 341$ and the number of patterns was $\lambda N = 40$.

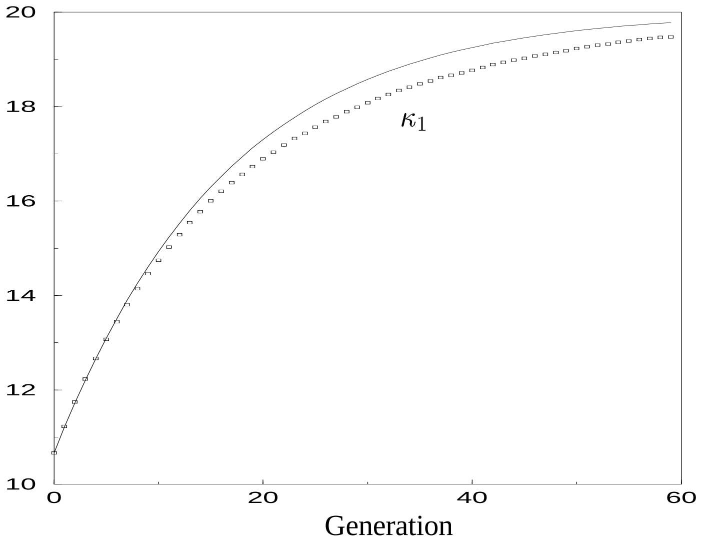

Figure 6.2: A population of $1000$ randomly generated configurations with summed square error below $\lambda N/3$ undergoes repeated generations of mutation with $p_{\mathrm{m}} = 0.02$. The mean (□) summed square error is shown each generation, averaged over $100$ samples. The solid curve shows the prediction from the microcanonical theory. The problem size was $N = 341$ and the number of patterns was $\lambda N = 40$.

### 6.4.2 Showing inconsistencies in the mutation results

The experiment in the preceding section showed empirically that the mutation expressions derived in section 6.3 do not accurately describe the effect of mutation in general. It is also possible to show analytically that these mutation results are inconsistent. One way that this can be achieved is by comparing the change under mutation of some macroscopic quantity other than the training error with the change predicted by the microcanonical approach.

An appropriate quantity to consider is the mean stability of patterns for a particular perceptron,

$$\begin{aligned} \overline{\Lambda} &= \langle \Lambda^{\mu} \rangle_{\mu} \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i \langle \xi_i^{\mu} \rangle_{\mu} \end{aligned} \tag{6.33}$$

where $\Lambda_{\mu}$ is the stability of pattern $\mu$ and the angled brackets denote an average over training patterns. It is straightforward to calculate the expected value for this quantity after mutation,

$$\begin{aligned} \overline{\Lambda}_{\mathrm{m}} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \langle M_i S_i \rangle \langle \xi_i^{\mu} \rangle_{\mu} \\ &= \Gamma \overline{\Lambda} \end{aligned} \tag{6.34}$$

where $M_i$ are the mutation variables defined in equation (6.12) which are averaged out (as denoted by the first set of brackets) to give $\Gamma = 1 - 2p_{\mathrm{m}}$.

It is also relatively straightforward to calculate the expectation value for the mean stability given the training error if one assumes an unbiased average over configurations. Under this assumption one can define a conditional probability for the mean stability given the training error,

$$p(\overline{\Lambda}|E) = \frac{\langle \delta(\overline{\Lambda} - \frac{1}{\lambda N} \sum_{\mu} \Lambda^{\mu})\, \delta(E - \sum_{\mu} u_l(\mathcal{T} - \Lambda^{\mu})) \rangle_{\{S_i\}}}{\langle \delta(E - \sum_{\mu} u_l(\mathcal{T} - \Lambda^{\mu})) \rangle_{\{S_i\}}} \tag{6.35}$$

The cumulants of this distribution can be calculated in general using the replica method, as was the case for the $p(E_{\mathrm{m}}|E)$ calculation in section 6.3. The annealed result holds for sufficiently low $\lambda$ and for large $N$ one finds that for the step error with $\mathcal{T} = 0$ the expected value of the mean stability is a simple linear function of the training error,

$$\overline{\Lambda} = \sqrt{\frac{2}{\pi}} \left( 1 - \frac{2E}{\lambda N} \right) \tag{6.36}$$

As expected, the mean stability increases from an initial value of zero as the error is reduced towards zero.

Under the same assumptions the expected training error after mutation is given by equation (6.27a) (except that a single perceptron is under consideration here, rather than a population),

$$E_{\mathrm{m}} = (1 - 2\Delta)E + \Delta\lambda N \tag{6.37}$$

where $\Delta$ is defined in equation (6.28).

Equations (6.34), (6.36) and (6.37) are inconsistent. Equation (6.37) shows how the training error changes by a non-linear function of $\Gamma$ under mutation, while equations (6.34) and (6.36) require a linear relationship. Since equation (6.34) is exact and equation (6.37) only requires an unbiased configuration space before mutation, then equation (6.36) must be incorrect after mutation from an randomly selected configuration. The assumption of an unbiased distribution of configurations after mutation must then be false. As well as being inconsistent in general, these expressions remain inconsistent in the most relevant limit of weak mutation where $\Delta \simeq 2\sqrt{p_{\mathrm{m}}}/\pi$.

Another inconsistency in the mutation results becomes apparent by observing that for a low mutation rate and large $N$ the application of mutation twice should be equivalent to a single mutation with a doubled mutation probability. This is because mutating the same bit twice is vanishingly unlikely in this limit. This is certainly the case for the mutation results in chapter 4, section 4.3. For the step error expressions in section 6.3 the expected training error after mutation to first order in $\sqrt{p_{\mathrm{m}}}$ is,

$$E^{\mathrm{m}} \simeq \left(1 - \frac{4\sqrt{p_{\mathrm{m}}}}{\pi}\right)E + \frac{2\sqrt{p_m}}{\pi}\lambda N \tag{6.38}$$

The expected step error after two mutations to first order in $\sqrt{p_{\mathrm{m}}}$ is then,

$$(E^{\mathrm{m}})^{\mathrm{m}} \simeq \left(1 - \frac{8\sqrt{p_{\mathrm{m}}}}{\pi}\right)E + \frac{4\sqrt{p_m}}{\pi}\lambda N \tag{6.39}$$

which is greater than the expected step error after a single mutation with probability $2p_{\mathrm{m}}$. This is confirmed by the results of figure 6.1, which shows that the theory significantly overestimates the mean change in training error within the population after two rounds of mutation.

For the summed square error these kinds of inconsistencies are expected to be much smaller (compare figure 6.1 with figure 6.2). Unfortunately, there were technical difficulties in calculating the higher cumulants after mutation for this error measure and the analysis has therefore not been pursued further.

## 6.5 Conclusion

In this chapter the statistical mechanics formalism was applied to the strong NP-hard problem of storing random binary patterns in a perceptron with binary weights. This provides a stiff test for any theoretical approach, as the analysis of this problem is very difficult even in a thermodynamics framework, where the powerful assumption of thermal equilibrium can be used. In the limit of small batch size and large problem size it was possible to characterize mutation under the assumption that configurations were typical of configurations with a given training error – a microcanonical formulation.

Unfortunately, the assumption of an unbiased population of configurations was found to be false in most cases. To verify this finding a simple experiment was conducted, where an initially unbiased population with training errors below some pre-determined threshold was subjected to successive mutations. The microcanonical prediction diverged rapidly from averaged simulation results after the first generation for the step error measure, and more slowly for the summed square error. The latter error measure contained information about the stability of unstored pattern and it is argued that this may lead to a more constrained characterization of configurations and correspondingly better averaging. However, due to technical difficulties the higher cumulants after mutation were not calculated for this measure and the theory could not be properly tested. For the simpler step error the mutation results were shown to be completely inconsistent in at least two ways.

The evidence presented in this chapter suggests that for this problem the perceptron is not sufficiently well characterized by training error alone to allow a general description of mutation. It may therefore be necessary to include more information. For example, one could use statistics describing the distribution of pattern stabilities associated with each configuration, as suggested in a previous study [55]. The simplest such statistic would be the mean stability of training patterns, but other statistics may also be required. Then it would be necessary to follow the joint distribution of the training error and these extra statistics within the population in order to model the dynamics of the GA. It is not clear at present whether this will be achievable in practice, as it would presumably be technically very difficult. The inclusion of crossover would provide an added complication because it involves the interaction of different population members. There is a long way to go before it will be possible to accurately model the dynamics of even the simplest GA in general for this problem.

# Chapter 7

# Increasing biological realism: diploidy and temporally varying fitness

## 7.1 Introduction

Many natural organisms have both a haploid and a diploid stage in their development, with the diploid stage often predominating in higher organisms. During the diploid phase there are two sets of genes available and therefore twice the necessary amount of genetic information required for development. Which alleles are expressed at each site may depend on their relative dominance. Although a diploid phase may be required to facilitate arguably beneficial biological processes, such as DNA repair, sexual recombination and assortment of chromosomes, it is an open question as to why the diploid phase is so prolonged in animals and is often the only phase represented by a multicellular organism. A number of taxa, for example some plants and fungi, can produce both diploid and haploid individuals. In some algae only the haploid phase is represented by a multicellular organism [36]. It is often argued that for diploidy to have become so common it must present some advantage. One common belief is that having two genes present allows deleterious mutant alleles to exist as recessives within the population, which might then become selectively advantageous under a change in the environment or a return to previous conditions. Since fully recessive alleles are only expressed when there are two copies at a site, then the probability of a rare harmful allele being expressed is much lower if it is recessive in a diploid population than would be the case in a haploid population. The existence of diploidy allows greater genetic diversity to exist within the population which selection can then act on.

A problem in some GA applications is the maintenance of diversity within the population and this is exacerbated if the fitness function changes over time, since genetic diversity is soon lost under continued selection pressure. A number of schemes exist in order to combat such premature convergence. For static fitness measures two of the most popular methods are island and niching models. In an island model, the population is spatially divided into subpopulations (islands) each evolving independently save for infrequent migrations which reintroduce diversity (this also allows a parallel implementation) [49]. Niching methods come in a variety of forms, but they generally invoke some form of density dependent selection, so that individuals are penalised if they are genetically similar (or sometimes phenotypically similar) to existing population members [21]. For GAs evolving in a temporally varying environment one possible way to maintain diversity is to use diploidy, or even polyploidy (for some recent examples, see references [22, 65, 77]). This is an old idea and reference [22] provides a review of early treatments.

In this chapter the statistical mechanics formalism is extended to incorporate diploid genotypes. This is desirable both because of possible GA applications and also because it brings the method closer to population genetics, which usually involves diploid models [12]. A simple temporally varying fitness measure is also considered. Because of time constraints the work is not complete and a number of interesting models have not been studied; most notably only haploids are modelled under the temporally varying fitness measure and an evolving dominance map is not considered. An adaptive dominance map would be most desirable in a GA, where one would not know *a priori* which map to choose for a non-trivial problem. However, the dynamics of a number of simple systems are solved and the strength and potential of the formalism are demonstrated. Possible extensions to more involved situations are discussed, including the evolution of the dominance map and simple parasite-host interactions, which are of interest in natural populations as well as in artificial genetic search [27, 31, 39].

## 7.2 A simple diploid GA

A highly idealized diploid GA is considered, which is only very roughly analogous to any biological system. A diploid genotype comprises a pair of binary strings, called gametes[1]. Initially, a random population of diploids are created and the genetic operators are then applied as follows,

1. A population of $2P$ gametes is selected from $P$ diploids, with each gamete selected according to the fitness of its associated diploid. Each diploid can only generate two types of gamete – there is no assortment or recombination at this stage.

2. The gametes undergo crossover and mutation at random, with no regard for which diploid the gametes originate from.

[1] An abuse of biological terminology – real gametes (eggs or sperm) are created by diploids through assortment of and recombination between chromosomes from each diploid parent. The gametes from two parents then fuse to create zygotes (fertilized eggs) which develop into diploid adults. Gametes are not contained within the diploid and certainly do not participate in recombination, as in the highly idealized situation described above.

3. Pairs of gametes fuse at random to produce $P$ diploids for the new population.

These steps are iterated over a number of generations in much the same way as for the familiar haploid GA. Of course, this procedure differs from the biological picture in a number of respects. Most notably, the recombination phase (crossover) is separated from the selection phase in a rather artificial manner. A more realistic situation would be to only allow recombination between gametes from the same diploid, and then to randomly fuse gametes in order to create the new diploid population. However, in this chapter crossover will generally be so disruptive that such a distinction makes little difference. The essential feature from the point of view of genetic search is that selection acts on the diploid. Each diploid produces genetic material for the next generation in proportion to its selective weight. The phenotype of the diploid is some function of the two constituent gametes and may involve some form of dominance.

For the purposes of modelling it is convenient to create an infinite pool of gametes after selection, as in the haploid GA. The dynamics can then be followed in terms of statistics from this infinite population. This does not change the nature of the problem and the two algorithms are essentially equivalent. The theoretical algorithm only differs in the first step above, which now reads,

1. An infinite population of gametes is selected from $P$ diploids, with each gamete selected according to the fitness of its associated diploid.

Steps 2 and 3 are the same as above.

### 7.2.1 A diploid phenotype

Recall the definition of the phenotype for the additive haploid genotype, defined in equation 4.1. Consider the case where $J_i = 1$ at every site, as in the one-max problem,

$$R_\alpha = \sum_{i=1}^{N} S_i^\alpha$$

where $S_i^\alpha \in \{-1, 1\}$ are alleles of the haploid genotype. The diploid genotype is made up of two haploid genotypes which will be called gametes ($R_\alpha$ will be called the gamete phenotype). One way to define the diploid phenotype associated with gametes $\alpha$ and $\beta$ and with dominance is,

$$R_{\alpha\beta} = \tfrac{1}{2}\sum_{i=1}^{N}\left[S_i^{\alpha} + S_i^{\beta} + h_i(1 - S_i^{\alpha}S_i^{\beta})\right] \tag{7.1}$$

where $h_i$ is the dominance coefficient which determines the contribution from site $i$ when $S_i^{\alpha}$ and $S_i^{\beta}$ differ. The vector of dominance coefficients is called a dominance map. This phenotype has been studied in quantitative genetics for the case where $h_i$ is the same at every site, leading to a number of exact results for stationary distributions in the infinite population limit (see, for example, reference [30]). In the context of genetic search it is important to be able to characterize the dynamics for more general dominance maps, as it is not known in general which sites should be dominant. The goal is to eventually be able to describe a GA with an adaptive dominance map, although this is beyond the scope of the present analysis.

For zero dominance one chooses $h_i = 0$ for all $i$, in which case the diploid phenotype is the average of the two gamete phenotypes. This is the only situation when the diploid phenotype can be written in terms of its two associated gamete phenotypes. In general, details about the configuration of each gamete are required in order to determine the diploid phenotype. For example, if $h_i = 1$ for all $i$ then the final term in equation (7.1) is the correlation between gametes.

### 7.2.2 Modelling the dynamics

It is most convenient to follow the dynamics of the distribution of gametes within the population. The gamete phenotype is the same as for the additive haploid problems which were considered in chapter 4. The only difference between those problems and the simple diploid considered here is in the selection phase. Thus, the expressions for mutation and crossover given in sections 4.3 and 4.4 still hold (although the maximum entropy distribution may require extra constraints). In the following two sections, expressions are derived for the dynamics in the case of directional selection without dominance and with a fixed binary dominance map. The former situation is the most straightforward, as the diploid fitness in this case is simply the mean phenotype of its two constituent gametes. The latter situation is more involved and cannot be addressed under the present formalism without resorting to the bit-simulated crossover (BSC) limit, which was introduced for the haploid case in chapter 4, section 4.9.

## 7.3 Directional selection without dominance

There are $P$ diploids within the population and each is associated with two gametes, which are randomly chosen from the infinite pool of gametes before selection. Label the gametes in each associated pair $\alpha$ and $\alpha + P$ respectively. Recall that under directional selection the fitness of an individual equals the phenotype. From equation (7.1) the fitness of the diploid with gametes $\alpha$ and $\alpha + P$ under zero dominance is,

$$F_{\alpha,\alpha+P} = \tfrac{1}{2}(R_\alpha + R_{\alpha+P}) \tag{7.2}$$

where $R_\alpha$ is the familiar one-max phenotype. Boltzmann selection is used, in which case the selection weight for both gametes associated with a diploid is,

$$w_\alpha = w_{\alpha+P} = \exp\bigl(\tfrac{1}{2}\beta\bigl(R_\alpha + R_{\alpha+P}\bigr)\bigr) \tag{7.3}$$

The partition function for selection is (from equation (3.2)),

$$\begin{aligned} Z_{\mathrm{s}} &= \sum_{\alpha=1}^{2P} w_\alpha \exp(\gamma R_\alpha) \\ &= \sum_{\alpha=1}^{P}\left[\exp\bigl(\tfrac{1}{2}\beta(R_\alpha + R_{\alpha+P})\bigr)\left(\mathrm{e}^{\gamma R_\alpha} + \mathrm{e}^{\gamma R_{\alpha+P}}\right)\right] \end{aligned} \tag{7.4}$$

The logarithm of this quantity generates the cumulants of the infinite gamete population after selection. This can be averaged over $2P$ gametes randomly sampled from the gamete population before selection in order to calculate the expectation values for the cumulants, which are then given by equation (3.5),

$$K_n^{\mathrm{s}} = -\lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\int_0^\infty \mathrm{d}t\,\frac{f^P(t,\gamma)}{t}$$

where,

$$f(t,\gamma) = \int \mathrm{d}R_\alpha\,\mathrm{d}R_{\alpha'}\,p(R_\alpha)\,p(R_{\alpha'})\,\exp\left(-t\mathrm{e}^{\beta(R_\alpha+R_{\alpha'})/2}(\mathrm{e}^{\gamma R_\alpha} + \mathrm{e}^{\gamma R_{\alpha'}})\right) \tag{7.5}$$

These integrals must be computed numerically in general, as was the case for haploid selection. The correlation calculation given in chapter 4, section 4.6.2 can be similarly generalized to the diploid case.

### 7.3.1 Weak selection expansion

For weak selection (small $\beta\sqrt{K_2}$) it is possible to apply the $1/P$ expansion described in section 3.2.2. In this case the cumulants after selection are given by equation (3.9),

$$K_n^{\rm s} = \lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\left[\log(\psi_1(\beta,\gamma)) - \frac{1}{2P}\left(\frac{\psi_2(\beta,\gamma)}{\psi_1^2(\beta,\gamma)}\right)\right]$$

where,

$$\psi_n(\beta,\gamma) = \int \mathrm{d}R_\alpha\,\mathrm{d}R_{\alpha'}\,p(R_\alpha)\,p(R_{\alpha'})\,(\mathrm{e}^{\gamma R_\alpha} + \mathrm{e}^{\gamma R_{\alpha'}})^n \mathrm{e}^{\frac{n}{2}\beta(R_\alpha + R_{\alpha'})} \tag{7.6}$$

For $n = 1$ and $n = 2$ one finds,

$$\psi_1(\beta,\gamma) = 2\rho(\beta/2)\rho(\gamma+\beta/2) \tag{7.7a}$$

$$\psi_2(\beta,\gamma) = 2\rho(\beta)\rho(2\gamma+\beta) + 2\rho^2(\gamma+\beta) \tag{7.7b}$$

where $\rho(\beta)$ is the characteristic function of the gamete distribution (see equation (2.7)). Expanding in $\beta$ for the first few cumulants one finds,

$$K_1^{\rm s} = K_1 + \frac{\beta}{2}\left(1-\frac{1}{2P}\right)K_2 + \frac{\beta^2}{8}\left(1-\frac{3}{P}\right)K_3 + \cdots \tag{7.8a}$$

$$K_2^{\rm s} = \left(1-\frac{1}{2P}\right)K_2 + \frac{\beta}{2}\left(1-\frac{2}{P}\right)K_3 + \frac{\beta^2}{8}\left[\left(1-\frac{5}{P}\right)K_4 - \frac{6K_2^2}{P}\right] + \tag{7.8b}$$

$$K_3^{\rm s} = \left(1-\frac{3}{2P}\right)K_3 + \frac{\beta}{2}\left[\left(1-\frac{4}{P}\right)K_4 - \frac{3K_2^2}{P}\right] + \cdots \tag{7.8c}$$

It is instructive to compare these expressions with the small $\beta$ results for directional Boltzmann selection in a haploid GA, which are given in equations (3.21a) to (3.21c). Diploid selection without dominance is almost equivalent to haploid selection with a population of size $2P$ (which is the number of gametes in the diploid population) and a halved selection strength. The two cases are not exactly equivalent, as there are subtle differences in the finite population terms. However, the discussion for the haploid case given in chapter 3, section 3.3.2 still holds. Selection increases the magnitude of the higher cumulants, most notably the third, which slows down progress under further selection. The other genetic operators are required to re-populate the tails of the gamete distribution and reduce the magnitude of the higher cumulants. The effects of mutation and crossover on the gamete distribution are described in chapter 4, sections 4.3 and 4.4. Although a halved selection strength would be significant in a biological population, where selection is imposed by the environment, it is not so important in the context of artificial genetic search, because the selection strength can easily be doubled if necessary. In this case there is no significant difference between a haploid GA and a diploid GA without dominance.

### 7.3.2 Simulating the dynamics

The selection expressions in section 7.3 were combined with the mutation and crossover results from chapter 4, sections 4.3 and 4.4 in order to simulate the dynamics. Bit-simulated crossover (BSC) was used, as this allows the dynamics to be described in terms of only the mean gamete and mean correlation within the population, therefore simplifying the selection numerics. The higher cumulants are still required after crossover and these are determined using the maximum entropy ansatz described in chapter 4, section 4.5 (four cumulants were used here). There is no reason why these results could not be generalized to other forms of crossover by using the methods developed in chapter 4. The selection strength was scaled inversely to the standard deviation of the gamete distribution ($\beta = \beta_{\rm s}/\sqrt{\kappa_2}$). It may be more appropriate to use the variance of the diploid fitness distribution to scale the selection strength, but the present scaling allows a meaningful comparison of the haploid and diploid results.

Figure 7.1 shows the mean and variance of the gamete distribution averaged over 1000 runs of a diploid GA without dominance. The diploid fitness is the average of its two constituent gametes, so that the expected mean fitness within the diploid population is equal to the mean gamete. These results show very good agreement with the theoretical curves, although there are very slight systematic errors which may be due to non-self-averaging effects, deviations from maximum entropy or because the weak selection approximation was required to determine the correlation after selection. Results from a haploid GA with a halved selection strength and a doubled population size are shown for comparison and the trajectories are clearly very similar to those of the diploid GA. Any differences can be attributed to subtle finite population effects.

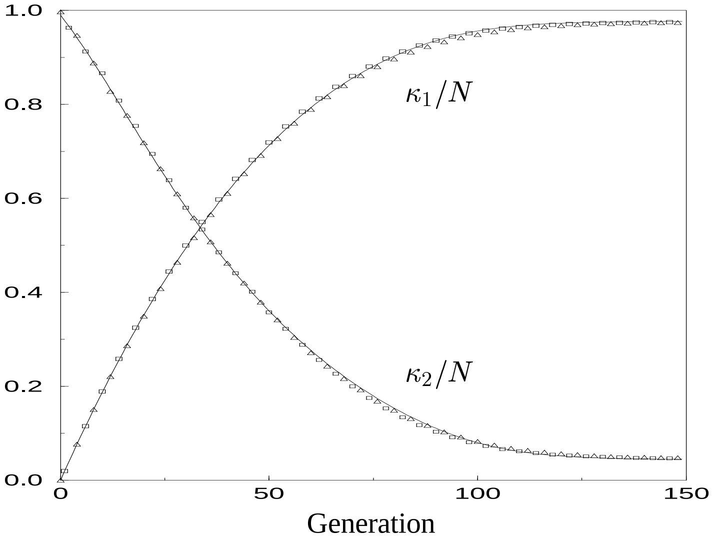

Figure 7.1: The theory is compared to averaged results from a diploid GA without dominance on the one-max problem. The diploid results (△) for the mean and variance of the gamete distribution are averaged over 1000 runs with $P = 50$ and $\beta_{\mathrm{s}} = 0.5$. The solid lines show the diploid theory. The haploid results (□) with $P = 100$ and $\beta_{\mathrm{s}} = 0.25$ are shown for comparison. The other parameters were $N = 155$, $p_{\mathrm{m}} = 0.002$ and bit-simulated crossover was used.

## 7.4 Directional selection with a fixed binary dominance map

When dominance is non-zero the fitness of a diploid can no longer be written in terms of its constituent gamete's phenotypes. For directional selection the fitness of a diploid is equal to the phenotype defined in equation (7.1),

$$F_{\alpha\beta} = \tfrac{1}{2} \sum_{i=1}^{N} \left[ S_i^\alpha + S_i^\beta + h_i (1 - S_i^\alpha S_i^\beta) \right] \tag{7.9}$$

The case where $h_i \in \{-1, 1\}$ is considered here, so that the dominance map is a binary vector. To determine this fitness it is necessary to know how alleles are distributed relative to the dominance map. One way to do this is to use the BSC limit, which was introduced in chapter 4, section 4.9. After BSC the distribution of alleles decouples at every site. The selection procedure can be averaged over this distribution in order to determine the expectation values for the relevant macroscopics after selection. It is first necessary to describe the distribution of alleles and this can be achieved by making a maximum entropy ansatz.

### 7.4.1 Maximum entropy distribution

Recall the maximum entropy calculation for the additive haploid genotype, which was introduced in chapter 4, section 4.5. To apply such an ansatz here, it is first necessary to decide which macroscopics are most important. The most obvious macroscopics to describe the gamete distribution are the mean gamete phenotype $K_1$ and the mean overlap between the gametes and the dominance map, which will be denoted $H$,

$$K_1 = \sum_{i=1}^{N} \langle S_i^\alpha \rangle_\alpha = \sum_{i=1}^{N} \tau_i \tag{7.10a}$$

$$H = \sum_{i=1}^{N} h_i \langle S_i^\alpha \rangle_\alpha = \sum_{i=1}^{N} h_i \tau_i \tag{7.10b}$$

where $\tau_i$ is the mean allele at site $i$ within the infinite gamete pool.

It will be necessary to include the correlation measure $q$, as the population is finite and will become correlated under selection. It is also desirable to know which sites are correlated and this can be achieved by including a fourth constraint, which will be denoted $Q$ (these expressions are for large $P$),

$$q = \frac{1}{N} \sum_{i=1}^{N} \langle S_i^\alpha \rangle_\alpha^2 = \frac{1}{N} \sum_{i=1}^{N} \tau_i^2 \tag{7.10c}$$

$$Q = \frac{1}{N} \sum_{i=1}^{N} h_i \langle S_i^\alpha \rangle_\alpha^2 = \frac{1}{N} \sum_{i=1}^{N} h_i \tau_i^2 \tag{7.10d}$$

For $h_i \in \{-1, 1\}$ the sum of $q$ and $Q$ gives the correlation for sites with positive dominance while the difference gives the correlation for sites with negative dominance. From equation (7.9) one finds that the expected mean fitness for a population of diploids whose gametes are randomly sampled from the gamete pool can be written in terms of $Q$ and $K_1$,

$$\langle F_{\alpha\beta} \rangle_{\alpha \neq \beta} = K_1 + \tfrac{1}{2}\Big(\sum_i h_i - Q\Big) \tag{7.11}$$

Notice that $Q$ is selected on directly, so that it may be necessary to include correlation macroscopics for this problem even in the infinite population limit.

The sites can be arranged so that $h_i = 1$ for the first $mN$ sites and $h_i = -1$ for the remaining sites. The particular ordering of the dominance coefficients is irrelevant, since there are no spatial interactions. Thus, $m$ completely parameterizes a fixed binary dominance map and determines the degree of dominance for positive alleles on average. Rewriting the expressions for $H$ and $Q$,

$$\begin{aligned} H &= \sum_{i=1}^{mN} \tau_i - \sum_{mN+1}^{N} \tau_i && \text{(7.12a)} \\ Q &= \frac{1}{N}\sum_{i=1}^{mN} \tau_i^2 - \frac{1}{N}\sum_{mN+1}^{N} \tau_i^2 && \text{(7.12b)} \end{aligned}$$

The four constraints can be enforced by Lagrange multipliers as in the haploid case. Notice that if $m = 0$ then $H = K_1$ and $Q = q$, so that only two constraints are required as in the haploid case (this is true in general if $h_i$ is the same at every site). A similar calculation to that presented in chapter 4, section 4.5 provides an expression for the mean bond at each site for the maximum entropy distribution (this result is also valid for more general dominance maps),

$$\tau_i = \tanh\left(z + h_i y + \eta_i \sqrt{x^2 + h_i w^2}\right) \tag{7.13}$$

where $w^2$, $x^2$, $y$ and $z$ are conjugate to $Q$, $q$, $H$ and $K_1$ respectively, while $\eta_i$ is a Gaussian variable with zero mean and unit variance. If $h_i = 0$ at every site, then this reduces to the haploid expression defined in equation (4.18). After BSC the alleles within the gamete pool are assumed to be distributed according to,

$$p(S_i) = \left(\frac{1+\tau_i}{2}\right)\delta(S_i - 1) + \left(\frac{1-\tau_i}{2}\right)\delta(S_i + 1) \tag{7.14}$$

The constraints fix the values of each Lagrange multiplier,

$$\begin{aligned} K_1 + H &= 2mN\,\overline{\tanh\left(z + y + \eta\sqrt{x^2 + w^2}\right)} && \text{(7.15a)} \\ K_1 - H &= 2N(1-m)\,\overline{\tanh\left(z - y + \eta\sqrt{x^2 - w^2}\right)} && \text{(7.15b)} \\ q + Q &= 2m\,\overline{\tanh^2\left(z + y + \eta\sqrt{x^2 + w^2}\right)} && \text{(7.15c)} \\ q - Q &= 2(1-m)\,\overline{\tanh^2\left(z - y + \eta\sqrt{x^2 - w^2}\right)} && \text{(7.15d)} \end{aligned}$$

where bars denote averages over the Gaussian noise. Although this is a four-dimensional root finding problem, a trivial change in variables decouples the equations into two pairs which can be solved independently. The problem is therefore no more involved than for the haploid case. In all the cases which were considered here the argument of the hyperbolic tangent remained real and the roots were unique.

### 7.4.2 Mutation

The mutation calculation is a straightforward generalization of the calculation in chapter 4, section 4.3. The expectation values for the two extra macroscopics are given by,

$$H^{\mathrm{m}} = \Gamma H \tag{7.16a}$$

$$Q^{\mathrm{m}} = \Gamma^2 Q \tag{7.16b}$$

where $\Gamma = 1 - 2p_{\mathrm{m}}$. These two equations are analogous to the results for $K_1$ and $q$ respectively.

### 7.4.3 Calculating $K_1$ and $H$ after selection

The selection calculation follows the haploid discussion closely (see chapter 4, section 4.9). For a diploid whose fitness measure is given by equation (7.9) the partition function for selection is,

$$Z_{\mathrm{s}} = \sum_{\alpha=1}^{P} \left[ \mathrm{e}^{\frac{\beta}{2} \sum_i \left[S_i^\alpha + S_i^{\alpha+P} + h_i(1 - S_i^\alpha S_i^{\alpha+P})\right]} \left( \mathrm{e}^{\gamma \sum_i S_i^\alpha} + \mathrm{e}^{\gamma \sum_i S_i^{\alpha+P}} \right) \right] \tag{7.17}$$

For weak selection the cumulants of the gamete phenotype distribution after selection are generated from the familiar $1/P$ expansion (see chapter 3, section 3.2.2),

$$K_n^{\mathrm{s}} \simeq \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \left[ \log(\psi_1(\beta,\gamma)) - \frac{1}{2P} \left( \frac{\psi_2(\beta,\gamma)}{\psi_1^2(\beta,\gamma)} \right) \right] \tag{7.18}$$

where $\psi_n(\beta,\gamma)$ is now averaged over alleles distributed according to equation (7.14),

$$\psi_n(\beta,\gamma) = \left\langle \mathrm{e}^{\frac{n\beta}{2} \sum_i \left[S_i^\alpha + S_i^{\alpha'} + h_i(1 - S_i^\alpha S_i^{\alpha'})\right]} \left( \mathrm{e}^{\gamma \sum_i S_i^\alpha} + \mathrm{e}^{\gamma \sum_i S_i^{\alpha'}} \right)^n \right\rangle_{\{S_i^\alpha, S_i^{\alpha'}\}} \tag{7.19}$$

Notice that when $h_i = 0$ this expression reduces to the zero dominance expression in equation (7.6), except that here the average is over alleles rather than the gamete distribution. Completing the average one finds,

$$\psi_1(\beta,\gamma) = 2 \prod_{i=1}^{N} \phi_i(\beta,\gamma,\gamma) \tag{7.20a}$$

$$\psi_2(\beta,\gamma) = 2 \prod_{i=1}^{N} \phi_i(2\beta,2\gamma,2\gamma) + 2 \prod_{i=1}^{N} \phi_i(2\beta,2\gamma,0) \tag{7.20b}$$

where,

$$\phi_i(\beta,\gamma,\delta) = \left(\frac{1+\tau_i}{2}\right)^2 \mathrm{e}^{\beta+\gamma} + \left(\frac{1-\tau_i}{2}\right)^2 \mathrm{e}^{-(\beta+\gamma)} + \frac{(1-\tau_i^2)}{2} \mathrm{e}^{\beta h_i} \cosh(\delta) \tag{7.21}$$

When using BSC it is only necessary to evolve the macroscopics required to describe the maximum entropy distribution. In this case only the mean gamete phenotype after selection is required, and from equation (7.18),

$$\begin{aligned} K_1^{\mathrm{s}} &= \frac{\psi_1'(\beta,\gamma)}{\psi_1(\beta,\gamma)} - \frac{1}{2P}\frac{\psi_2(\beta,\gamma)}{\psi_1^2(\beta,\gamma)}\left(\frac{\psi_2'(\beta,\gamma)}{\psi_2(\beta,\gamma)} - \frac{2\psi_1'(\beta,\gamma)}{\psi_1(\beta,\gamma)}\right)\Bigg|_{\gamma=0} \qquad (7.22)\\ &= \sum_{i=1}^{N}\frac{\phi_i'(\beta,\gamma)}{\phi_i(\beta,\gamma)} - \frac{1}{2P}\left(\sum_{i=1}^{N}\frac{\phi_i'(2\beta,2\gamma)}{\phi_i(2\beta,2\gamma)} - \frac{2\phi_i'(\beta,\gamma)}{\phi_i(\beta,\gamma)}\right)\mathrm{e}^{\sum_i \log(\phi_i(2\beta,2\gamma))-2\log(\phi_i(\beta,\gamma))}\Bigg|_{\gamma=0} \end{aligned}$$

where $\psi'(\beta,\gamma)$ and $\phi'(\beta,\gamma)$ denote differentials with respect to $\gamma$ and $\phi_i(\beta,\gamma) = \phi_i(\beta,\gamma,0)$. The average over the Gaussian variable in $\tau_i$ (see equation (7.13)) was taken over summed terms, as these are expected to self-average. This expression cannot be written as a simple function of cumulants unless $h_i = 0$ at every site, in which case the result reduces to the zero dominance case in equation (7.8a).

The mean overlap between gametes and the dominance map is $H$. The expectation value for this quantity after selection is found in a similar calculation to that given above. Recall the selection partition function defined in equation (7.17). By replacing $\gamma\sum S_i^\alpha$ by $\gamma\sum h_i S_i^\alpha$ in this expression one can generate the expectation value for $H$ after selection. This follows the result for the first cumulant closely and the result is given by the final line of equation (7.22) under the transformation $\phi_i'(\beta,\gamma) \to h_i\phi_i'(\beta,\gamma)$,

$$\begin{aligned} H^{\mathrm{s}} &= \sum_{i=1}^{N} h_i\frac{\phi_i'(\beta,\gamma)}{\phi_i(\beta,\gamma)}\\ &- \frac{1}{2P}\left(\sum_{i=1}^{N} h_i\left[\frac{\phi_i'(2\beta,2\gamma)}{\phi_i(2\beta,2\gamma)} - \frac{2\phi_i'(\beta,\gamma)}{\phi_i(\beta,\gamma)}\right]\right)\mathrm{e}^{\sum_i \log(\phi_i(2\beta,2\gamma))-2\log(\phi_i(\beta,\gamma))}\Bigg|_{\gamma=0} \qquad (7.23) \end{aligned}$$

### 7.4.4 Calculating $q$ and $Q$ after selection

As in the haploid case (see chapter 4, section 4.9.2) one can include an extra term in the selection partition function in order to generate the correlation after selection. In the case of a diploid whose fitness is given by equation (7.9) the relevant partition function is,

$$Z_q(\epsilon) = \sum_{\alpha=1}^{P}\left[\mathrm{e}^{\frac{\beta}{2}\sum_j\left[S_j^\alpha + S_j^{\alpha+P} + h_j(1-S_j^\alpha S_j^{\alpha+P})\right]}\left(\mathrm{e}^{\epsilon S_i^\alpha} + \mathrm{e}^{\epsilon S_i^{\alpha+P}}\right)\right] \qquad (7.24)$$

Using the familiar weak selection approximation leads to an expression for the correlation after selection,

$$\begin{aligned} q_{\mathrm{s}} &= \frac{1}{N}\sum_{i=1}^{N}\left(1-\lim_{\epsilon\to 0}\frac{\partial^2}{\partial\epsilon^2}\log Z_q(\epsilon)\right) \\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\left(1-\lim_{\epsilon\to 0}\frac{\partial^2}{\partial\epsilon^2}\left[\log(\psi_1(\beta,\epsilon))-\frac{1}{2P}\left(\frac{\psi_2(\beta,\epsilon)}{\psi_1^2(\beta,\epsilon)}\right)\right]\right) \qquad (7.25) \end{aligned}$$

where,

$$\psi_n(\beta,\epsilon) = \left\langle \mathrm{e}^{\frac{n\beta}{2}\sum_j\left[S_j^\alpha+S_j^{\alpha'}+h_j(1-S_j^\alpha S_j^{\alpha'})\right]}\left(\mathrm{e}^{\epsilon S_i^\alpha}+\mathrm{e}^{\epsilon S_i^{\alpha'}}\right)^n\right\rangle_{\{S_j^\alpha,S_j^{\alpha'}\}} \qquad (7.26)$$

The brackets denote averages over alleles distributed according to equation (7.14). Completing the average one finds,

$$\begin{aligned} \psi_1(\beta,\epsilon) &= 2\phi_i(\beta,\epsilon,\epsilon)\prod_{j\neq i}^{N-1}\phi_j(\beta,0,0) \qquad (7.27a) \\ \psi_2(\beta,\epsilon) &= 2\left(\phi_i(2\beta,2\epsilon,2\epsilon)+\phi_i(2\beta,2\epsilon,0)\right)\prod_{j\neq i}^{N-1}\phi_j(2\beta,0,0) \qquad (7.27b) \end{aligned}$$

where $\phi_i(\beta,\gamma,\delta)$ is defined in equation (7.21). The expression for $q_{\mathrm{s}}$ has not been differentiated out here as the resulting expression is rather cumbersome and is not particularly illuminating.

This result is easily generalized in order to calculate $Q$ after selection, by introducing a factor of $h_i$ into the outermost sum of equation (7.25).

$$Q^{\mathrm{s}} \simeq \frac{1}{N}\sum_{i=1}^{N}h_i\left(1-\lim_{\epsilon\to 0}\frac{\partial^2}{\partial\epsilon^2}\left[\log(\psi_1(\beta,\epsilon))-\frac{1}{2P}\left(\frac{\psi_2(\beta,\epsilon)}{\psi_1^2(\beta,\epsilon)}\right)\right]\right) \qquad (7.28)$$

### 7.4.5 Simulating the dynamics

The dynamics can be modelled for an arbitrary binary dominance map using the expressions which were derived in the preceding sections. Figure 7.2 compares the theory to averaged results for a completely random map ($m = 0.5$ in section 7.4.1), for two different population sizes. The mean fitness of the diploid population is shown in each case, along with results for each of the relevant macroscopics. The results for the largest population size show excellent agreement, while there is some disagreement during the later stages with the smaller population size. This may be because the weak selection approximation was required in order to calculate the selection expressions in sections 7.4.3 and 7.4.4, or because the dynamics average less well for smaller population sizes.

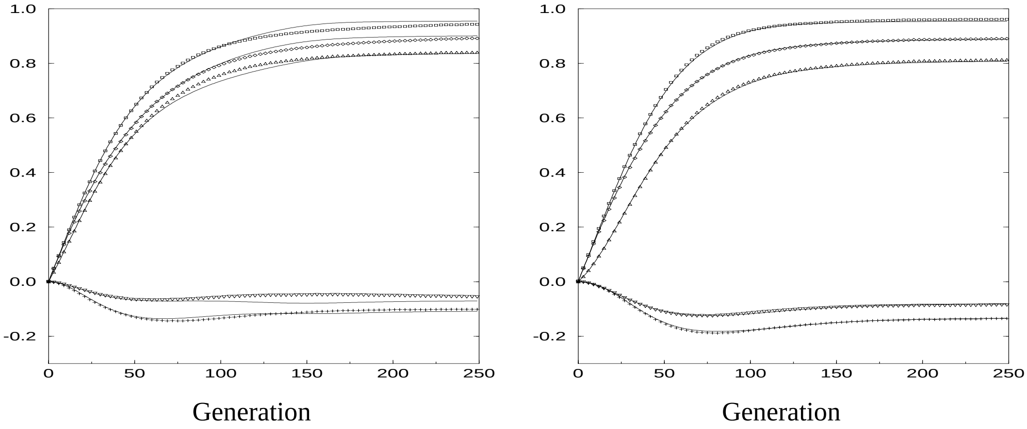

Figure 7.2: The theory is compared to averaged results from a diploid GA with a random binary dominance map on the one-max problem, for two population sizes. The population sizes were $P = 50$ (left) and $P = 100$ (right). The results for the mean fitness (□) and relevant macroscopics $K_1$ ($\diamond$), $q$ ($\triangle$), $H$ ($\triangledown$) and $Q$ ($+$) are shown (in descending order). The results are averaged over 600 runs for $P = 50$ and 400 runs for $P = 100$. The closest solid lines show the theory. The other parameters were $\beta_{\mathrm{s}} = 0.4$, $N = 155$, $p_{\mathrm{m}} = 0.002$ and bit-simulated crossover was used.

It is interesting to note that even in the case where the dominance map is completely random, then the mean diploid fitness is higher than the mean gamete ($K_1$). This has been achieved by driving $Q$ negative, which leads to an increase in the mean expressed fitness defined in equation (7.11). The mean correlation is also lower here than would be expected for a haploid population of the same mean fitness. More work is required to really understand the interplay of the relevant macroscopics and it would be most interesting to model a diploid under a temporally varying fitness measure. The next section takes us closer to this goal by describing the dynamics of a haploid GA under such a fitness measure. Along with the work in the present section, it is hoped that this might provide the basis for accurately characterizing more interesting situations.

## 7.5 Temporal changes in the fitness measure

The main motivation for using a diploid GA is to maintain diversity and retain useful information under a temporally varying fitness measure. If these changes are periodic or recurrent then a dominance map may learn information about previous states of the environment and this information may prove useful in the future. Such a scenario is beyond the scope of the formalism presented here as it stands, but it is hoped that something can be learned from a very simple example of a temporally varying problem. A haploid GA will be considered here as this simplifies the analysis, although generalization to a diploid GA with a fixed dominance map would be straightforward. In section 7.6 some possible applications of these ideas and those of the previous section will be considered.

### 7.5.1 A simple problem

A haploid GA is considered, whose phenotype is given by equation (4.1) with each weight initially set to one. The initial fitness for directional selection is then the same as for the one-max problem,

$$F_\alpha = \sum_{i=1}^{N} S_i^\alpha \tag{7.29}$$

The simplest way to change this fitness measure after some time is by introducing a new weight vector,

$$F_\alpha^{\mathrm{J}} = \sum_{i=1}^{N} J_i S_i^\alpha \qquad J_i = \begin{cases} 1 & \text{with probability } 1-v \\ -1 & \text{with probability } v \end{cases} \tag{7.30}$$

where $v$ determines the probability of introducing a negative weight. If $v = 0.5$ then the new weights are completely uncorrelated with the old and for smaller values of $v$ there is some correlation between the new and old weights. In general, this is an Ising paramagnet whose weights are flipped with probability $v$ in one generation. The weights are initially set to one without any loss of generality.

To simplify matters BSC is used, so that the distribution of alleles at each site decouples and averages can be taken directly over this distribution (see chapter 4, section 4.9). To describe the distribution of alleles after BSC it is necessary to make a maximum entropy ansatz.

### 7.5.2 Maximum entropy distribution

Before the weights of the paramagnet are flipped the problem is equivalent to one-max and constraints on the mean fitness and correlation within the population will accurately characterize the population, as described in chapter 4, section 4.5. However, once new weights are introduced at each site these two macroscopics are no longer sufficient, because the population is still correlated with the previous weight vector. It is therefore necessary to follow the overlap with the original weight vector, which will be denoted $O$. As in the diploid problem it is also desirable to know which weights are correlated and this can be achieved by including another extra macroscopic, which is denoted $Q$ in analogy to the similar macroscopic introduced in section 7.4.1. The two extra constraints required after a change in the weight vector are then,

$$O = \sum_{i=1}^{N} \langle S_i^\alpha \rangle_\alpha = \sum_{i=1}^{N} \tau_i \tag{7.31a}$$

$$Q = \frac{1}{N}\sum_{i=1}^{N} J_i \langle S_i^\alpha \rangle_\alpha^2 = \frac{1}{N}\sum_{i=1}^{N} J_i \tau_i^2 \tag{7.31b}$$

where $\tau_i$ is the mean allele at site $i$. If the mean fitness and correlation are included, then comparison with equations (7.10a) to (7.10d) shows that this is equivalent to the problem of maximizing entropy in the diploid GA with a fixed binary dominance map. The discussion in section 7.4.1 provides the result (with $v$ analogous to $1-m$) for the distribution of alleles at maximum entropy.

The mean bond at each site in this case is,

$$\tau_i = \tanh\left(z + J_i y + \eta_i \sqrt{x^2 + J_i w^2}\right) \tag{7.32}$$

where $w^2$, $x^2$, $y$ and $z$ are conjugate to $Q$, $q$, $K_1$ and $O$ respectively, while $\eta_i$ is a Gaussian variable with zero mean and unit variance. The constraints fix the Lagrange multipliers once the weights have been flipped and after averaging over the distribution of weights one finds,

$$O + K_1 = 2N(1-v)\,\overline{\tanh\left(z + y + \eta\sqrt{x^2 + w^2}\right)} \tag{7.33a}$$

$$O - K_1 = 2Nv\,\overline{\tanh\left(z - y + \eta\sqrt{x^2 - w^2}\right)} \tag{7.33b}$$

$$q + Q = 2(1-v)\,\overline{\tanh^2\left(z + y + \eta\sqrt{x^2 + w^2}\right)} \tag{7.33c}$$

$$q - Q = 2v\,\overline{\tanh^2\left(z - y + \eta\sqrt{x^2 - w^2}\right)} \tag{7.33d}$$

where bars denote averages over the Gaussian noise. Again, these equations decouple into two pairs under a trivial change in variables.

Unfortunately, each time the weight vector is changed the number of constraints is increased by a factor of two and the problem becomes progressively more complex (if all the relevant constraints are used). The root finding will still be straightforward, however, as the equations always decouple into pairs. Here, only a single change of weights is under consideration.

### 7.5.3 Evolving the macroscopics

In the previous section it was shown that the relevant macroscopics for this problem are equivalent to the macroscopics which described the population for a diploid with a fixed binary dominance map. Before the weights flip the dynamics are exactly equivalent to the one-max problem which was described in chapter 4, section 4.7.1. Once the weights have flipped it is necessary to follow the evolution of four macroscopics, as was the case for the diploid GA. Here, the dynamics are considered from the point when the weights flip. The expected values of the macroscopics at this point are found by averaging over the new weights (see equation (7.30)),

$$K_1^{\mathrm{J}} = (1-2v)K_1 \tag{7.34a}$$

$$Q^{\mathrm{J}} = (1-2v)q \tag{7.34b}$$

while $q$ remains fixed and $O = K_1$ is the overlap with the previous weights at this point.

As this is a haploid GA, expressions describing the effect of selection in the BSC limit are most closely related to those given in chapter 4, section 4.9. The expressions for $K_1^{\mathrm{s}}$ and $q_{\mathrm{s}}$ are exactly equivalent to equations (4.60) and (4.65), except that $\tau_i$ is now defined by equation (7.32). The expression for $O^{\mathrm{s}}$ is given by equation (4.60) with the factor of $J_i$ deleted in the two sums over sites. Similarly, the expression for $Q^{\mathrm{s}}$ is given by equation (4.65) with a factor of $J_i$ introduced into both sums over sites.

The mutation results for $O$ and $Q$ are analogous to the results for $H$ and $Q$ in the diploid GA, as described in section 7.4.2.

### 7.5.4 Simulating the dynamics

Figure 7.3 compares the theory to averaged results from a GA for one realization of the problem. The two extra macroscopics $O$ and $Q$ are only included after generation 70, where the weights change. At this point, $K_1$ is reduced according to equation (7.34a) and $Q$ is chosen according to equation (7.34b). The result for $K_1$ is therefore discontinuous at this point, as shown in the left hand part of the figure. The overlap with the original weight vector $O$ is initialized to the value of $K_1$ just before the weights are flipped, while $q$ is unchanged.

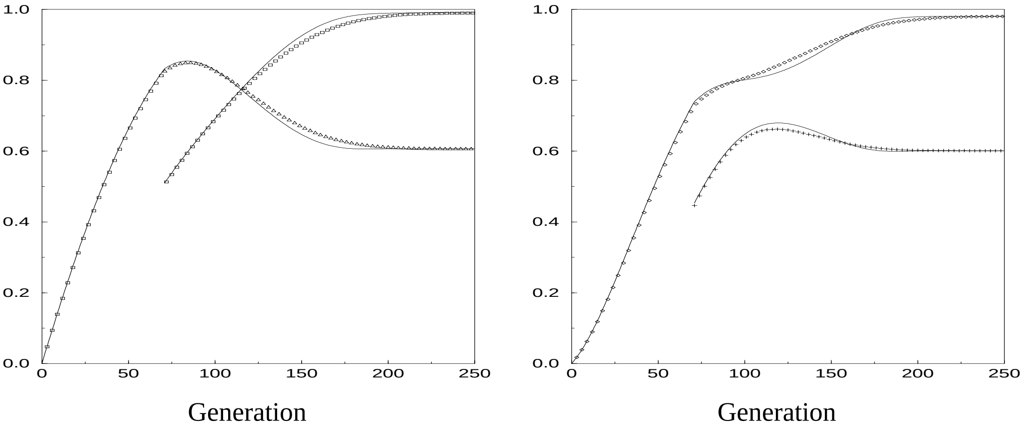

Figure 7.3: The theory is compared to averaged results from a haploid GA for a paramagnet whose Ising weight vector changes after 70 generations. The results are shown for $K_1\,(\Box)$, $O\,(\triangle)$, $q\,(\diamond)$ and $Q\,(+)$. The data points are averaged over 500 runs and solid lines show the theory. 30 weights were flipped at generation 70 ($v = 0.194$). The other parameters are $P = 200$, $\beta_\mathrm{s} = 0.2$, $N = 155$, $p_\mathrm{m} = 0.001$ and bit-simulated crossover was used.

The results show very good agreement, although there is a slight discrepancy in the predictions of transient behaviour after the weights have changed. This could be due to any combination of three simplifications – the use of a weak selection limit, the assumption of self-averaging and the assumption of maximum entropy. The weak selection approximation should hold here, as search parameters were chosen in a region which is usually well described by this limit. Similar small discrepancies in the transients were found for a range of selection strengths and population sizes. Notice that the correlation results (in the right hand part of figure 7.3) show that the averaged results are somewhat 'flatter' than the theoretical predictions, which is what one might expect if differences are due to non-self-averaging. Nevertheless, the theory provides a very close approximation to the averaged results in a wide range of situations. Errors due to lack of self-averaging should become smaller under increases in problem size and population size.

## 7.6 Conclusion

In this chapter the statistical mechanics formalism was applied to a simple diploid GA and a haploid GA with a temporally varying fitness measure. The theory compared well to simulation results in both cases. Although these were highly idealized and rather simple systems, it may be possible to use the methods developed here in order to describe the more involved and interesting situations described below.

### 7.6.1 Adaptive dominance

If the fitness measure is temporally varying, then it may be beneficial to let the dominance map evolve. One simple way to do this is to include a dominance map with each gamete and allow this to evolve in the same way as the rest of the genotype. Unfortunately, for binary genotypes this will lead to situations where two gametes disagree on the dominance at a site, leading to a possible ambiguity. Although any ambiguity could be resolved by making a random choice, this does not always give satisfactory performance. Holstein, Holland, and later Goldberg and Smith, chose a definite bias towards one choice of allele in cases where there was any ambiguity [22, 32, 33, 65]. This allowed the genotype to be represented by a triallelic scheme, where combinations of $0$ and $1$ act as if $1$ is dominant while an extra allele $1_0$ represents a $1$ over which $0$ is dominant. Although this form allows no ambiguity, the representation is now biased towards ones; an unfortunate lack of symmetry. Nevertheless, Smith and Goldberg do find a definite advantage when using this scheme on a temporally varying knapsack problem [65]. It would be possible to introduce a more symmetrical quadrallelic scheme, although it is not clear that this would be an improvement.

In order to analyse an adaptive dominance map under the present formalism, it would probably be simplest to consider a co-evolving but physically independent population of dominance maps from which dominance values at each site are chosen at random each time fitness is evaluated. The population of dominance coefficients would then correspond to a vector of probabilities associated with selecting a particular dominance value at any site. An extra set of macroscopics describing the population of dominance maps would be required in order to characterize such a GA. The generalization to closer physical proximity of dominance coefficients and expressed alleles (as in the triallelic scheme) would be difficult, but might produce similar dynamical behaviour in some cases.

### 7.6.2 Hosts and parasites

The dynamics of host-parasite interactions are of interest in artificial genetic search [31] as well as in the more familiar setting of biology [27, 39]. Hillis considered a GA for developing networks which sort sequences of numbers by rank [31]. The aim is to develop a sorting network which orders all possible sequences correctly and uses the smallest number of steps. To fully test a sorting network requires that all $2^N$ examples of a binary sequence of length $N$ are correctly ordered, but this test is very time consuming for large $N$. Using a smaller subset of training examples proves to be ineffective, as the GA soon learns how to sort most examples and they provide no information once learned, leaving the GA stuck at poor but locally optimal networks. Hillis found that by co-evolving a population of training examples as parasites he could ensure that sorting networks received a useful set of training examples each generation. The example sequences were selected by their ability to beat the sorting networks, while the networks were selected by their ability to correctly order the sequences. Adding this flexibility to the space of training examples allowed the GA to find very good solutions to the problem. Recently, however, some doubt has been cast over the stiffness of the test used by Hillis and more work is required to determine whether host-parasite interactions are really practicable for artificial genetic search [4].

A simpler co-evolution problem lends itself more readily to analysis: that of matching bit-strings. The parasite bit-string tries to match the host, while the host tries to be different from (evade) the parasite. This has analogies in biology, where interactions of this sort have received some attention [27, 39]. These studies mostly concentrate on small systems or on analysis via simulations and interesting dynamical patterns are shown to emerge even for very simple systems. An important question in biology is why sexual recombination should be so prevalent in higher organisms and there is some evidence that parasites provide one explanation for this. Under this view, sexual recombination is required to give hosts sufficient flexibility with which to deal with faster evolving parasites. This is therefore an interesting problem to study, both because it may shed light on general issues of host-parasite dynamics and also because it may have more specific implications for real biological systems.

Recall the simple time varying problem considered in section 7.5. In the simplest bit-matching problem one could treat the field of the paramagnet as the parasite, while the paramagnet would be the host (although the paramagnet fitness now changes sign – the host tries to be different from the parasite). One could then model the evolution of the parasite and host populations, which might interact in a number of ways. Many different situations can be envisioned, with varying levels of recombination within each population and varying rates of evolution. Unfortunately, the analysis in this chapter required BSC to decouple alleles at each site after crossover and the inclusion of more general forms of crossover is a formidable task. Another difficulty which emerges from the model described in section 7.5 is the observation that each time the environment changes, a new set of macroscopics are required to describe the overlap with the previous environment. This leads to an explosion in the number of macroscopics required to model the GA under continued adaptation. Overlaps with all previous environments may not be required, however, because effects would fall off with time and a truncated set of macroscopics might be sufficient to describe the dynamics accurately. Yet another possible difficulty with the present approach is that rapidly fluctuating dynamics may not be well described under an assumption of self-averaging. This would be a significant problem in host-parasite interactions, where fluctuating and chaotic behaviour has been observed in simulations [27, 39]. However, given the progress made in this chapter there is some reason to be optimistic about the prospect of deriving truly non-trivial macroscopic dynamical behaviour from a model defined in terms of microscopics.

# Chapter 8

# Conclusion and outlook

## 8.1 Thesis summary

A formalism for modelling GA dynamics using methods from statistical mechanics, originally developed by Prügel-Bennett and Shapiro [53, 54], has been reviewed and improved upon in order to describe the GA in a wider range of applications. The effect of selection on the distribution of phenotypes within the population is problem independent, and previous results for this operator were generalized to a larger class of selection schemes. The averaged dynamical trajectory of a simple finite population GA was then accurately modelled for a number of optimization problems. Although the problems for which the formalism proved most successful were rather simple or idealized, they were often sufficiently involved to capture interesting non-trivial features of the search. An attempt was also made to describe a strong NP-hard problem and although the analysis was unsuccessful in this case, some insight was gained into possible limitations of the formalism as it stands.

**The first class of problems considered**, and the class considered in greatest detail, consisted of problems where alleles of the genotype contribute additively to the phenotype (fitness was related to the phenotype by some arbitrary function). Results from Prügel-Bennett and Shapiro [54] were reproduced, including their calculation for determining non-trivial terms in the expressions describing crossover and mutation by maximizing entropy with constraints on the mean correlation (genotype similarity) and mean phenotype within the population. Some situations under which the maximum entropy ansatz might break down were considered. In particular, it was shown how mutation could take the population away from maximum entropy by flipping alleles at sites associated with large weights in the random-field paramagnet.

Prügel-Bennett and Shapiro assumed a simple relationship between the phenotypic variance and correlation which does not hold in general [54]. In order to move beyond this simplification, expressions were derived for evolving the mean correlation as an extra macroscopic, providing a significant improvement over the original formulation. To determine the expected mean correlation under selection, it was assumed that the distribution of correlations within the population can be well approximated by the distribution at maximum entropy. This is expected to be a good assumption as long as crossover is reasonably disruptive, but further analysis is required to determine when this assumption will break down.

The theoretical results were tested on problems exhibiting directional selection (one-max and the random-field paramagnet) and stabilizing selection (the subset sum problem). The theory agreed well with averaged results from a real GA, accurately predicting the mean dynamical trajectory as long as the maximum entropy ansatz provided a good approximation. As mentioned above, moderate levels of mutation resulted in the maximum entropy ansatz breaking down for the random-field paramagnet during the later stages of the evolution, because alleles associated with high weights were flipped with significant probability.

The subset sum problem is a weakly NP-hard problem and has a strongly non-linear fitness measure. It is characterized by a stabilizing dynamics, analogous to stabilizing selection on quantitative traits in biological populations, so that the mean of the phenotype distribution centres around the optimum phenotype while the population converges. The dynamical trajectory was accurately predicted for this problem without mutation, but further work is required to determine whether the method fails when mutation is included. It was shown how the fitness of the best individual eventually degrades as the population becomes highly correlated. This effect was accurately predicted by estimating the probability of duplicates occurring within the population.

**The second class of problems considered** consisted of those whose fitness measure is a stochastic function of the phenotype (this is not mutually exclusive from the class of additive problems described above). This situation is often of interest in machine learning applications, where training data may be incomplete or noisy. A result was derived for selection on an arbitrary stochastic fitness measure and the specific case of directional Boltzmann selection on a phenotype corrupted by Gaussian noise was considered in greater detail. In the latter case, an increase in population size was shown to completely remove the detrimental effects of noise in the limit of weak selection.

A simple learning problem, generalization in a perceptron with binary weights, was shown to be closely related to a noisy version of one-max if a fresh batch of training examples were used for each training error evaluation. In this case the noise was due to the uncertainty of information contained within a finite training batch. The dynamics was solved for this problem, and in the limit of large problem size and weak selection it was shown how the population size could be chosen each generation to remove the effects of noise. When this population size was chosen, an optimal batch size was identified which minimized the computation time required for training error evaluations.

**In chapter 6** an attempt was made to model the GA on a strong NP-hard problem – storing random patterns in a binary perceptron. This differs from the other problems considered in this thesis, because the phenotype (in this case the training error) is a strongly non-linear function of the genotype. The effect of mutation was calculated under a microcanonical formulation, where perceptron configurations were assumed to be typical of configurations with a given training error. The calculation was carried out using the replica method to average over the random disorder in the training patterns and in the limit of small capacity the replica-symmetric result reduced to the much simpler annealed result. In this case it was possible to determine the cumulants of the error distribution after mutation for the step error measure, and the mean error after mutation for the summed square error measure. The higher cumulants were not calculated in the latter case because of technical difficulties.

Unfortunately, the microcanonical formulation did not describe mutation well in general and it was shown that there were at least two significant inconsistencies in the results. It was concluded that the training error did not constrain perceptron configurations sufficiently and it would be necessary to include other statistics for a better characterization. It was suggested that the mean stability of training patterns might provide useful information, although this was only conjectured and no attempt was made to model the population using extra statistics. Any analysis of crossover would be expected to introduce even greater difficulties, as this operator involves the interaction of different population members.

**In chapter 7** the formalism was extended in order to describe a class of simple diploid GAs and a haploid GA with a temporally varying fitness measure. For these problems the dynamics was solved for a GA using a limiting form of crossover, bit-simulated crossover, which completely decouples the alleles at each site (although the dynamics for a diploid GA without dominance was solved without this restriction). For a fixed binary dominance map the maximum entropy ansatz was extended to include four constraints, effectively describing the occupation and correlation at sites with each of the two different dominance values.

The simple temporally varying fitness measure considered was an Ising paramagnet, some of whose weights flip after a number of generations. In this case the maximum entropy calculation also involved four constraints after the fitness measure changed. The extra constraints described the memory of the original weight vector within the population.

The work in this chapter was incomplete and a number of possible generalizations were discussed. For example, it was shown how these results might be extended to described a diploid GA with an adaptive dominance map. This may be useful if it is not known *a priori* which dominance map to choose, or if the fitness measure changes unpredictably over time. Simple co-evolving systems were also considered, as these are of interest in natural systems as well as in artificial genetic search.

## 8.2 Strengths and weaknesses of the formalism

A statistical mechanics formalism has been shown to accurately predict the dynamical trajectory of the GA for a number of simple, but often non-trivial, problems. The expressions describing the dynamics are compact and do not depend on problem size or population size, although the assumption that trajectories self-average will probably improve with increases in both[1].

Finite population effects are accurately modelled under the formalism and provide a number of important insights. For example, Prügel-Bennett and Shapiro quantified the effect of directional selection on the higher cumulant of a finite population [53], showing how Boltzmann selection introduces skewness into an initially Gaussian distribution of phenotypes. In chapter 5 it was shown how adding Gaussian noise to the fitness only affects a finite population GA in some cases and may have no effect in the infinite population limit. This insight allowed the optimal batch size to be determined for a simple, yet by no means trivial, learning problem. It was also recognized that a finite population would correlate more rapidly under selection than would be predicted in the infinite population limit, because selection requires the duplication of population members. It was therefore necessary to quantify this duplication effect when

[1] This may not be the case if other parameters do not scale appropriately. For example, in Müller's ratchet the mutation rate is $O(1/N)$, where $N$ is the problem size, and fluctuations still dominate the dynamics of a finite population as $N$ tends to infinity [52].

modelling the dynamics.

Following the correlation as an explicit macroscopic allowed a greater number of problems to be addressed and gave improved results over the original formulation of the method, in which the correlation was deduced directly from the phenotypic variance [54]. As well as giving improved results for directional selection on an additive genotype (one-max and the random-field paramagnet) this was an essential ingredient for describing stabilizing selection (the subset sum problem). The solution to the dynamics for this problem marked significant progress, as this is an example of a weakly NP-hard problem with a strongly non-linear fitness function.

The maximum entropy ansatz often provides a powerful means of describing the distribution of alleles at each site for problems where alleles contribute additively, but inhomogeneously, to the phenotype [54]. However, there are situations when this distribution does not provide a good characterization of the population (at least with the constraints used here) and care must be taken when applying an ansatz with no *a priori* justification. The maximum entropy ansatz also provides a way to describe the distribution of correlations at each site and is therefore even necessary for modelling problems with homogeneous weights when finite population effects are important. In chapter 7 the ansatz was extended to simple diploid problems and to a temporally varying fitness measure, where four constraints were required to characterize the population. It was noted that for one-max with a binary dominance map there was a correlation measure in the expressed fitness. In this case the correlation constraints might be required even for large populations. The use of a constraint on the previous environment in the simple temporally varying problem shows that the maximum entropy result can also be used to follow history effects. This may be important when modelling more complex adaptive behaviour.

Certain limitations of the formalism were exposed in chapter 6, where an attempt was made to characterize the effect of mutation for a strong NP-hard problem. Here, the calculation was carried through under a microcanonical formulation, so that the only constraint on each population member was its training error (this is also a maximum entropy ansatz, but with a constraint on each individual rather than the whole population). However, the training error alone proved insufficient to accurately characterize configurations and the results did not even provide a reasonable approximation. Although it was suggested that other constraints might be included within the phenotype, the resulting calculations would be technically difficult and the generalization to crossover is expected to be even harder. It may also be the case that no small set of macroscopic constraints exist which accurately characterize perceptron configurations for this problem, although this conclusion seems overly pessimistic.

Whether limitations of the formalism are purely technical, or more fundamental in nature, is not yet known. An upper bound on the difficulty of problems for which the dynamics might be tractable is probably provided by thermodynamic studies, which use the powerful concept of thermal equilibrium to analyse the state space for a number of non-trivial optimization problems, including the strong NP-hard problem considered in chapter 6 [37, 43]. These studies apply the maximum entropy principle in a far more rigorous context, by considering a simulated annealing schedule which equilibrates over ergodic time-scales. In this case the dynamics is designed to approach a Boltzmann distribution. The thermodynamic formulation does not described the approach to this distribution, however, and there may be entropic barriers, or dynamic freezing transitions, which are intrinsic to the geometry of the fitness landscape and which such a study will not necessarily expose [34, 69]. In this case the thermodynamics only provides existence proofs for solutions and may say nothing about the dynamics of any search algorithm.

The formalism described in this thesis can be expected to meet with much greater technical difficulties than the thermodynamic approach, as the population is not at thermal equilibrium and it may be difficult, or impossible, to find a small set of macroscopics which accurately characterize the population. It is expected that this task will become more difficult as the mapping between genotype and phenotype becomes less direct and increasingly non-linear. In order to overcome the problem of increasing complexity, it may sometimes be possible to make simplifications to a problem without losing interesting features of the dynamics. For example, using independent training examples for each error evaluation allowed the dynamics to be solved for the generalization problem in chapter 5 and a similar simplification has recently allowed the dynamics of gradient descent to be solved for a class of multi-layer perceptrons with continuous weights [59].

Another possible problem for the statistical mechanics formalism is that much interesting detail of the dynamics may be lost through averaging. For example, the concept of punctuated equilibrium is of interest both in biological populations and in artificial genetic search [23, 46, 74]. Punctuated equilibrium describes a situation where the population is relatively stable for long periods, punctuated by short periods of rapid evolutionary change. In this case the mean dynamic trajectory over different realizations of the process does not capture important features of the dynamics and may be very difficult to compute in any case, as fluctuations will dominate the process. However, the formalism described in this thesis can be generalized to describe the evolution of an ensemble of populations, in which case large fluctuations from mean behaviour can be accurately modelled [52]. Whether this analysis can be carried out for more involved problems is not yet known.

## 8.3 Future work

The formalism described here is still under development. The predictive power of the method has been demonstrated on a number of simple examples, but it is now necessary to focus on specific issues which are of interest to the GA community, or possibly the population genetics community, for which these methods may provide novel insight. Technical improvements and generalizations of the formalism would also be of great interest. For example, it may be possible to resolve the difficulties encountered in chapter 6, or to solve the dynamics for other hard problems, by increasing the number of constraints within the phenotype. It would also be useful to examine the validity of approximations used here in greater detail, such as the weak selection expansion (chapter 3, section 3.2.2), or the use of a maximum entropy distribution for correlations (chapter 4, section 4.6.3). Some other possibilities for future research are outlined below.

### 8.3.1 Analysis of the equations of motion

Most of the work in this thesis centres around the derivation and verification of equations of motion for the GA. These expressions already provide some insight into the processes at work within the GA. For example, the higher cumulants and correlation were shown to be important in characterizing a finite population GA. The characterization of noise in the evaluation of fitness also required the accurate modelling of finite population effects and this was captured by the selection equations in a very simple and intuitive way. However, although some intuition is gained from simply looking at the equations of motion, few attempts have yet been made to analyse these equations in order to answer specific questions. Some notable exceptions are in chapter 5, where the optimal batch size was determined for a simple learning problem, and in Shapiro and Prügel-Bennett [64], where escape times are determined for a simple two-well potential. In this latter study the escape time from a local energy minima (fitness maxima) was compared to results for simulated annealing, showing that there are situations when the GA will escape more rapidly. However, more analysis is required to determine how finite population effects should be included within this analysis. It is hoped that the results described in this thesis could also provide the tools for many other studies.

### 8.3.2 Multi-layer perceptrons

In chapter 5 it was shown how the dynamics of a GA training a simple binary perceptron to generalize could be solved by describing the training error as a stochastic function of the phenotype, in this case the overlap between teacher and student. It would be interesting to attempt a generalization to multi-layer perceptrons, which are required to learn less trivial mappings (see, for example, reference [29]). In this case the phenotype would not be a single order parameter, but rather a vector of parameters describing the overlap between nodes of the student and teacher. It might then be possible to follow the joint distribution of overlaps within the population. Unlike in the simple perceptron problem, however, the search would have to break symmetry in the space of macroscopics for this problem, because the network has a number of equivalent permutations. How this symmetry breaking might occur within the population would be of great interest. It might be necessary to invoke an ensemble of populations in order to describe the many symmetrical states. This ensemble would then become multi-modal under symmetry breaking events within its constituent populations.

In order to describe the training error as a simple stochastic function of overlaps between nodes of the teacher and student network, it would be necessary to present a fresh batch of training examples for each error evaluation. This is an unfortunate idealization, as it does not capture a number of interesting features of training under more realistic scenarios, where there is a limited amount of data available to learn from. It would be most instructive to incorporate finite training set effects into the dynamics, but this would seem a formidable task as the exact characterization of such effects is difficult even in a static or thermodynamic study, where the replica method has to be used [61]. Whether an approximation exists which captures the essential features of quenched disorder without resorting to the replica method remains an open question.

### 8.3.3 Quantitative genetics

Quantitative genetics is concerned with the study of inheritable traits which can differ by degree and are mostly influenced by gene differences at many loci (see, for example, reference [12]). As described in chapter 4, section 4.7.1, the one-max problem under Boltzmann selection is equivalent to the multiplicative fitness landscape, which is one of the simplest quantitative genetics models. The dynamics of stabilizing selection and problems with inhomogeneous contributions at each site is also of some considerable interest to workers in this field[2].

Although the problems considered in this thesis are very close to those often considered relevant in quantitative genetics, there is a difference of emphasis between this work and quantitative genetics models, the resolution of which may not be straightforward. In chapter 4 it was pointed out that the correlation calculation given in section 4.6 ignores effects due to off-site terms, or linkage in the language of population genetics. This was assumed to be a good approximation as long as recombination was sufficiently disruptive. Unfortunately, in biological populations the degree of recombination is not always assumed to be high and linkage effects might become important. In this case the relevant question is: can the formalism, and in particular the maximum entropy calculation, include effects from off-site terms? The answer to this question is not yet known although it would seem a difficult problem in general, because the population would have to be constrained with both off-site and on-site averages. It would

[2] Nick Barton and Ellen Baake are currently translating some of these results into the language of population genetics.

probably be possible to include constraints on off-site averages alone, which might be sufficient in the infinite population limit, but it is unclear how relevant this limit is. It is also unclear how useful a constraint on second order off-site terms would be, because selection imposes a strong bias on third and fourth order off-site terms (related to the higher cumulants) which presumably would not be predictable from lower order terms.

The simple diploid problems considered in chapter 7 may also be of some interest in population genetics, but the diploid model outlined there was highly idealized and the analysis required an unrealistic and highly disruptive form of crossover. Whether this system is comparable to any real biological population is questionable, although it may serve as a useful solvable model in a 'fast recombination' limit. Possible extensions to an adaptive dominance map and a simple co-evolution problem were discussed in section 7.6.

### 8.3.4 Truly hard problems ?

The formalism described here requires that one can determine the essential features of genotypes within the population by averaging over a small number of macroscopic statistics. Clearly, this will not always be possible, as these statistics will not always constrain the population sufficiently well for the average to be representative (or the averaging procedure may be too difficult). This was shown in chapter 6 for the problem of storing random patterns in a binary perceptron, when configurations were only constrained by their training error. Hard optimization problems such as this are generally characterized by complex and non-linear mappings from genotype to fitness, so that the fitness provides less direct information about the genotype. There might also be strong spatial interactions between alleles within the genotype which would also make any analysis very difficult. For the subset sum problem, and problems with noise corrupted fitness, it was shown how the existence of a phenotype with a simpler relationship to the genotype can make analysis easier. For very hard problems one might include more degrees of freedom within the phenotype in order to constrain the genotype better, so that averaging the phenotype is more representative. Which degrees of freedom to include within the phenotype will typically not be obvious, although looking for the order parameters in a thermodynamic study might provide some insight. Whether the approach described in this thesis can be applied to a truly hard problem is an open question and this provides a stiff challenge to the formalism.

# Appendix A

# Maximum entropy calculation for the correlation after selection

The second term in the expression for the correlation after selection given in equation (4.28) will be calculated by determining the distribution of correlations at maximum entropy. Rewriting equation (4.29),

$$\begin{aligned} q_\infty &= \int \mathrm{d}q_{\alpha\beta}\,\mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_\mathrm{s}(R_\alpha)p_\mathrm{s}(R_\beta)\,p(q_{\alpha\beta}|R_\alpha,R_\beta)\,q_{\alpha\beta} \\ &= \lim_{t\to0}\frac{\partial}{\partial t}\log\left(\int \mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_\mathrm{s}(R_\alpha)\,p_\mathrm{s}(R_\beta)\,\rho(t\,|R_\alpha,R_\beta)\right) \end{aligned} \tag{A.1}$$

where $\rho(t\,|R_\alpha,R_\beta)$ is the characteristic function of $p(q_{\alpha\beta}|R_\alpha,R_\beta)$ (see equation (2.7)),

$$\rho(t\,|R_\alpha,R_\beta) \;=\; \int \mathrm{d}q_{\alpha\beta}\,p(q_{\alpha\beta}|R_\alpha,R_\beta)\mathrm{e}^{tq_{\alpha\beta}} \tag{A.2}$$

A conditional probability for correlations $p(q_{\alpha\beta}|R_\alpha,R_\beta)$ can be defined if alleles are assumed to come from the maximum entropy distribution described in section 4.5. In this case one has,

$$\begin{aligned} p(q_{\alpha\beta}|R_\alpha,R_\beta) &= \frac{p(q_{\alpha\beta},R_\alpha,R_\beta)}{p(R_\alpha,R_\beta)} \\ &= \frac{\langle\delta(q_{\alpha\beta}-\frac{1}{N}\sum_i S_i^\alpha S_i^\beta)\,\delta(R_\alpha-\sum_i J_iS_i^\alpha)\,\delta(R_\beta-\sum_i J_iS_i^\beta)\rangle}{\langle\delta(R_\alpha-\sum_i J_iS_i^\alpha)\,\delta(R_\beta-\sum_i J_iS_i^\beta)\rangle} \end{aligned} \tag{A.3}$$

where $\delta(x)$ is the Dirac delta function and the angled brackets denote averages over configurations of $S_i^\alpha$ and $S_i^\beta$. The alleles at each site are distributed according to,

$$p(S_i) = \left(\frac{1+\tau_i}{2}\right)\delta(S_i-1) + \left(\frac{1-\tau_i}{2}\right)\delta(S_i+1) \tag{A.4}$$

Here, $\tau_i$ is the mean allele per site at maximum entropy, which is defined in equation (4.18).

Consider the characteristic function of $p(q_{\alpha\beta}|R_\alpha, R_\beta)$, as this appears in the appropriate generating function,

$$\rho(Nt\,|R_\alpha, R_\beta) = \frac{\rho(Nt, R_\alpha, R_\beta)}{\rho(0, R_\alpha, R_\beta)} \tag{A.5}$$

where the factor of $N$ is included so that $t$ is scaled appropriately. The numerator of this expression is the characteristic function of the joint distribution for correlations and phenotypes,

$$\begin{aligned}\rho(Nt, R_\alpha, R_\beta) &= \int \mathrm{d}q_{\alpha\beta} \big\langle \delta(q_{\alpha\beta} - \tfrac{1}{N}\textstyle\sum_i S_i^\alpha S_i^\beta)\,\delta(R_\alpha - \sum_i J_i S_i^\alpha)\,\delta(R_\beta - \sum_i J_i S_i^\beta)\big\rangle \mathrm{e}^{Ntq_{\alpha\beta}} \\ &= \Big\langle \delta\big(R_\alpha - \sum_i J_i S_i^\alpha\big)\,\delta\big(R_\beta - \sum_i J_i S_i^\beta\big)\exp\big(t\sum_i S_i^\alpha S_i^\beta\big)\Big\rangle_{\{S_i^\alpha, S_i^\beta\}}\end{aligned} \tag{A.6}$$

The delta functions in this expression can then be written by their Fourier representation,

$$\delta(x) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}y}{2\pi\mathrm{i}} \mathrm{e}^{-yx} \tag{A.7}$$

so that equation (A.6) becomes,

$$\rho(Nt, R_\alpha, R_\beta) = \Big\langle \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}y_\alpha \mathrm{d}y_\beta}{-4\pi^2} \exp\big(-y_\alpha R_\alpha - y_\beta R_\beta + \sum_{i=1}^{N}(y_\alpha J_i S_i^\alpha + y_\beta J_i S_i^\beta + t S_i^\alpha S_i^\beta)\big)\Big\rangle \tag{A.8}$$

Each site decouples and the average over sites can be taken by integrating over the allele distribution defined in equation (A.4). The resulting integral can be computed for large $N$ by the saddle point method since the exponent of the integrand is $O(N)$ [40].

Eventually one finds (ignoring irrelevant multiplicative constants),

$$\rho(Nt, R_\alpha, R_\beta) = \exp\big[G(t, R_\alpha, R_\beta)\big] \tag{A.9}$$

where,

$$\begin{aligned}G(t, R_\alpha, R_\beta) = -y_\alpha R_\alpha - y_\beta R_\beta + \sum_{i=1}^{N} \log\,\big[&(1+\tau_i)^2 \mathrm{e}^{t + y_\alpha J_i + y_\beta J_i} \\ &+ 2(1-\tau_i^2)\mathrm{e}^{-t}\cosh(y_\alpha J_i - y_\beta J_i) + (1-\tau_i)^2 \mathrm{e}^{t - y_\alpha J_i - y_\beta J_i}\big]\end{aligned}$$

The saddle point equations fix $y_\alpha$ and $y_\beta$ as implicit functions of $R_\alpha$, $R_\beta$ and $t$,

$$\frac{\partial G}{\partial y_\alpha} = 0 \qquad \frac{\partial G}{\partial y_\beta} = 0 \tag{A.10}$$

Define $\rho(Nt)$, whose logarithm is the generating function for $q_\infty$ (see equation (A.1)),

$$\begin{aligned} \rho(Nt) &= \int \mathrm{d}R_\alpha \, \mathrm{d}R_\beta \, p_{\mathrm{s}}(R_\alpha) \, p_{\mathrm{s}}(R_\beta) \, \rho(Nt \, | R_\alpha, R_\beta) \\ &= \int \mathrm{d}R_\alpha \, \mathrm{d}R_\beta \, p_{\mathrm{s}}(R_\alpha) \, p_{\mathrm{s}}(R_\beta) \exp[G(t, R_\alpha, R_\beta) - G(0, R_\alpha, R_\beta)] \end{aligned} \tag{A.11}$$

The overlap distributions are expressed by their Fourier transformed cumulant expansions,

$$p_{\mathrm{s}}(R_\alpha) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}a}{2\pi\mathrm{i}} \exp\left(\sum \frac{a^n}{n!} K_n^{\mathrm{s}} - aR_\alpha\right) \tag{A.12}$$

$$p_{\mathrm{s}}(R_\beta) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}b}{2\pi\mathrm{i}} \exp\left(\sum \frac{b^n}{n!} K_n^{\mathrm{s}} - bR_\beta\right) \tag{A.13}$$

Now $\rho(Nt)$ is an integral over $a$, $b$, $R_\alpha$ and $R_\beta$ which can again be computed by the saddle point method. One finds that as $t \to 0$ the saddle point equations are satisfied by,

$$y_\alpha = y_\beta = y \tag{A.14}$$

$$R_\alpha = R_\beta = K_1^{\mathrm{s}} \tag{A.15}$$

These are related through an implicit function for $y$ in terms of mean overlap after selection,

$$K_1^{\mathrm{s}} = \sum_{i=1}^{N} J_i \left( \frac{\tau_i + \tanh(yJ_i)}{1 + \tau_i \tanh(yJ_i)} \right) \tag{A.16}$$

Then $q_\infty$ is generated from the logarithm of $\rho(Nt)$,

$$\begin{aligned} q_\infty &= \frac{1}{N} \lim_{t \to 0} \frac{\partial}{\partial t} \log \rho(Nt) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left( \frac{\tau_i + \tanh(yJ_i)}{1 + \tau_i \tanh(yJ_i)} \right)^2 \end{aligned} \tag{A.17}$$

# Appendix B

# Replica calculation for mutation in the binary perceptron

## B.1 Replica calculation for a general training error

To make the calculation simpler, the number of spins flipped by mutation is fixed and is equal to $\gamma N$. In general, $\gamma$ will fluctuate around the mutation probability $p_m$ and these fluctuations should be averaged out. Here, it will be assumed that $\gamma = p_m$ is a good approximation. This is reasonable for large $N$ if $p_m$ is of order unity, which is a necessary condition for the saddle point approximation used here in any case. Unfortunately, GAs often use a mutation probability of order $1/N$, in which case this approximation may break down. It has not been determined whether the following method gives a good approximation in this case.

Choose the first $\gamma N$ sites to be flipped, with no loss of generality.

$$M_i = \begin{cases} -1 & \text{for } i = 1, 2, \dots, \gamma N \\ 1 & \text{for } i = \gamma N + 1, \dots, N-1, N \end{cases} \tag{B.1}$$

One can rewrite equation (6.16), fixing the stabilities with delta functions,

$$\begin{aligned} \rho^n(t,E) = \Bigg\langle \prod_{\alpha=1}^{n} \Bigg\langle \Bigg[ \prod_{\mu} \int \mathrm{d}\Lambda_{\mathrm{m}}^{\mu\alpha} \mathrm{d}\Lambda^{\mu\alpha} \delta\big(\Lambda_{\mathrm{m}}^{\mu\alpha} - \tfrac{1}{\sqrt{N}} \sum_i M_i S_i^\alpha \xi_i^\mu\big) \delta\big(\Lambda^{\mu\alpha} - \tfrac{1}{\sqrt{N}} \sum_i S_i^\alpha \xi_i^\mu\big) \Bigg] \\ \times \delta\big(E - \sum_\mu u_l(\mathcal{T} - \Lambda^{\mu\alpha})\big) \exp\big(t \sum_\mu u_l(\mathcal{T} - \Lambda_{\mathrm{m}}^{\mu\alpha})\big) \Bigg\rangle_{\{S_i^\alpha\}} \Bigg\rangle_{\{\xi_i^\mu\}} \end{aligned} \tag{B.2}$$

where $\alpha$ labels replicas and $\mu$ labels patterns. The inner average is over all weight configurations, while the outer average is over the quenched patterns. The delta functions constraining the stabilities can be given their Fourier representation (see equation (A.7)), with $x_\alpha^\mu$ and $z_\alpha^\mu$ conjugate to $\Lambda^{\mu\alpha}$ and $\Lambda_{\mathrm{m}}^{\mu\alpha}$ respectively,

$$\begin{aligned}
&\prod_{\mu,\alpha}\delta\big(\Lambda_{\mathrm{m}}^{\mu\alpha}-\tfrac{1}{\sqrt{N}}\sum_i M_i S_i^\alpha\xi_i^\mu\big)\delta\big(\Lambda^{\mu\alpha}-\tfrac{1}{\sqrt{N}}\sum_i S_i^\alpha\xi_i^\mu\big)\\
&\quad=\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\prod_{\mu,\alpha}\left(\frac{\mathrm{d}x_\alpha^\mu\mathrm{d}z_\alpha^\mu}{-4\pi}\right)\prod_\mu\exp\Big[\sum_\alpha\big(x_\alpha^\mu\Lambda^{\mu\alpha}+z_\alpha^\mu\Lambda_{\mathrm{m}}^{\mu\alpha}-\tfrac{1}{\sqrt{N}}\sum_i(x_\alpha^\mu+z_\alpha^\mu M_i)S_i^\alpha\xi_i^\mu\big)\Big]
\end{aligned} \tag{B.3}$$

It is now possible to average the right hand side of this expression over patterns,

$$\begin{aligned}
&\left\langle\exp\Big[\sum_\alpha\big(-\tfrac{1}{\sqrt{N}}\sum_i(x_\alpha^\mu+z_\alpha^\mu M_i)S_i^\alpha\xi_i^\mu\big)\Big]\right\rangle_{\{\xi_i^\mu\}}\\
&\quad=\quad\prod_i\left\langle\exp\Big[-\tfrac{1}{\sqrt{N}}\xi_i^\mu\sum_\alpha(x_\alpha^\mu+z_\alpha^\mu M_i)S_i^\alpha\Big]\right\rangle_{\{\xi_i^\mu\}}\\
&\quad=\quad\exp\Big[\sum_i\log\cosh\Big(\tfrac{1}{\sqrt{N}}\sum_\alpha(x_\alpha^\mu+z_\alpha^\mu M_i)S_i^\alpha\Big)\Big]\\
&\quad=\quad\exp\Big[\tfrac{1}{2N}\sum_i\Big(\sum_\alpha(x_\alpha^\mu+z_\alpha^\mu M_i)S_i^\alpha\Big)^2+O\big(\tfrac{1}{N}\big)\Big]\\
&\quad=\quad\exp\Big[\tfrac{1}{2}\sum_\alpha\big((x_\alpha^\mu)^2+(z_\alpha^\mu)^2+2\Gamma x_\alpha^\mu z_\alpha^\mu\big)+\sum_{\beta>\alpha}q_{\alpha\beta}\big(x_\alpha^\mu x_\beta^\mu+z_\alpha^\mu z_\beta^\mu+\Gamma(x_\alpha^\mu z_\beta^\mu+x_\beta^\mu z_\alpha^\mu)\big)\Big]
\end{aligned} \tag{B.4}$$

where $q_{\alpha\beta}$ is the correlation between replicas and $\Gamma$ is the mean mutation variable,

$$q_{\alpha\beta}=\tfrac{1}{N}\sum_{i=1}^N S_i^\alpha S_i^\beta \qquad \Gamma=1-2\gamma=\tfrac{1}{N}\sum_{i=1}^N M_i \tag{B.5}$$

In writing the final line of equation (B.4), terms of $O(1/N)$ were neglected and the following approximation was used,

$$\tfrac{1}{N}\sum_{i=1}^N M_i S_i^\alpha S_i^\beta \simeq \Gamma q_{\alpha\beta} \tag{B.6}$$

This is a good approximation as long as $N$ is large and $\gamma$ is of order unity, in which case this quantity should self-average. A delta function can be used to impose the constraint on each $q_{\alpha\beta}$,

$$1=\int\mathrm{d}q_{\alpha\beta}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{N\mathrm{d}\phi_{\alpha\beta}}{2\pi\mathrm{i}}\exp\Big[\phi_{\alpha\beta}\big(Nq_{\alpha\beta}-\sum_i S_i^\alpha S_i^\beta\big)\Big] \tag{B.7}$$

Recall equation (B.2): the delta function containing $E$ can be written by its integral representation (see equation (A.7)), with $\nu_\alpha$ conjugate to $E$ for each replica. The product over $\mu$ decouples and can be written as a power. Using Gardner's notation where possible [16] one eventually finds,

$$\begin{aligned} \rho^n(t,E) &= \int \prod_{\beta>\alpha} \big(\mathrm{d}q_{\alpha\beta}\big) \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \prod_{\beta>\alpha} \left(\frac{N\mathrm{d}\phi_{\alpha\beta}}{2\pi\mathrm{i}} \frac{\mathrm{d}\nu_\alpha}{2\pi\mathrm{i}}\right) \exp(G) \\ G &= -E\sum_{\alpha} \nu_\alpha + N \sum_{\beta>\alpha} \phi_{\alpha\beta} q_{\alpha\beta} + \lambda N G_0 + N G_1 \end{aligned} \tag{B.8}$$

where $\lambda$ is the capacity. Here, $G_1$ is equivalent to Gardner's notation and $G_0$ is also equivalent to Gardner's notation in the case where $\Gamma = 0$ ($\gamma = 0.5$) and the configurations are completely randomized by mutation. In this case $\rho(t,E)$ reduces to the characteristic function of the density of states.

$$\exp(NG_1) = \left\langle \exp\Big(\sum_{\beta>\alpha} \phi_{\alpha\beta} \sum_i S_i^\alpha S_i^\beta\Big) \right\rangle_{\{S_i^\alpha\}} \tag{B.9}$$

$$\begin{aligned} \exp(G_0) &= \int \prod_\alpha \left(\mathrm{d}\Lambda^\alpha \mathrm{d}\Lambda_{\mathrm{m}}^\alpha \frac{\mathrm{d}x_\alpha \mathrm{d}z_\alpha}{-4\pi^2}\right) \exp\Big[\sum_\alpha \big(t u_l(\mathcal{T} - \Lambda_{\mathrm{m}}^\alpha) + \nu_\alpha u_l(\mathcal{T} - \Lambda^\alpha) + x_\alpha \Lambda^\alpha \\ &+ z_\alpha \Lambda_{\mathrm{m}}^\alpha + \tfrac{1}{2}(x_\alpha)^2 + \tfrac{1}{2}(z_\alpha)^2 + \Gamma x_\alpha z_\alpha\big) + \sum_{\beta>\alpha} q_{\alpha\beta}\big(x_\alpha x_\beta + z_\alpha z_\beta + \Gamma(x_\alpha z_\beta + x_\beta z_\alpha)\big)\Big] \end{aligned} \tag{B.10}$$

The integral in equation (B.8) can be computed for large $N$ by the saddle point method [40].

## B.2 The replica symmetric solution

For capacities lower than the critical capacity it is assumed that replica symmetry holds, as this is thought to be true for all temperatures in the thermodynamic treatment [43]. In this case one can make the following simplifications,

$$q_{\alpha\beta} = q \quad \alpha \neq \beta \tag{B.11}$$

$$\phi_{\alpha\beta} = \phi \quad \alpha \neq \beta \tag{B.12}$$

$$\nu_\alpha = \nu \tag{B.13}$$

The expression for $G_1$ defined in equation (B.9) can now be simplified,

$$
\begin{aligned}
G_1 &= \tfrac{1}{N}\log\Big\langle \exp\Big(\sum_{\beta>\alpha}\phi\sum_i S_i^\alpha S_i^\beta\Big)\Big\rangle_{\{S_i^\alpha\}} \\
&= \log\Big\langle \exp\Big(\tfrac{\phi}{2}\big(\sum_\alpha S^\alpha\big)^2 - \tfrac{n\phi}{2}\Big)\Big\rangle_{S^\alpha} \\
&= \log\int Du\ \exp\big[n\log\big(2\cosh(u\sqrt{\phi})\big) - \tfrac{n\phi}{2}\big] \\
&\overset{n\to 0}{=} n\int Du\ \log\big(2\cosh(u\sqrt{\phi})\big) - \frac{n\phi}{2}
\end{aligned}
\qquad \text{(B.14)}
$$

where,

$$
\int Du \ = \ \int \frac{\mathrm{d}u}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u^2}{2}} \qquad \text{(B.15)}
$$

The expression for $G_0$ can also be simplified. Consider the sum over $\beta > \alpha$ in the exponent of equation (B.10),

$$
\begin{aligned}
\sum_{\beta>\alpha} q\Big(x_\alpha x_\beta + z_\alpha z_\beta + \Gamma(x_\alpha z_\beta + x_\beta z_\alpha)\Big) \ &= \ \tfrac{1}{2}q\Gamma\Big[\Big(\sum_\alpha x_\alpha + z_\alpha\Big)^2 - \sum_\alpha (x_\alpha + z_\alpha)^2\Big] \\
&+ \tfrac{1}{2}q(1-\Gamma)\Big[\Big(\sum_\alpha x_\alpha\Big)^2 + \Big(\sum_\alpha z_\alpha\Big)^2 - \sum_\alpha (x_\alpha)^2 - \sum_\alpha (z_\alpha)^2\Big]
\end{aligned}
\qquad \text{(B.16)}
$$

The squares over sums can be removed by introducing Gaussian integrals. This allows the terms for each replica to decouple and eventually one finds,

$$
G_0 \overset{n\to 0}{=} n\int D\eta_x\, D\eta_z\, D\eta_{xz}\ \log\Big[\int \mathrm{d}\Lambda\,\mathrm{d}\Lambda_\mathrm{m}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{\mathrm{d}x\,\mathrm{d}z}{-4\pi^2}\exp\big(F(\eta,\Lambda,\Lambda_\mathrm{m},x,z)\big)\Big] \qquad \text{(B.17)}
$$

where,

$$
\begin{aligned}
F(\eta,\Lambda,\Lambda_\mathrm{m},x,z) &= tu_l(\mathcal{T}-\Lambda_m) + \nu u_l(\mathcal{T}-\Lambda) + x\Lambda + z\Lambda_m \\
&+ \tfrac{1}{2}(1-q)(x^2+z^2+2\Gamma xz) + \eta_{xz}\sqrt{q\Gamma}(x+z) + \sqrt{q(1-\Gamma)}(x\eta_x + z\eta_z)
\end{aligned}
\qquad \text{(B.18)}
$$

Recall the definition of $u_l(\mathcal{T}-\Lambda)$ in equation (6.6). The expression for the step function can be simplified as follows,

$$
\exp\big(t(\mathcal{T}-\Lambda)^l\Theta(\mathcal{T}-\Lambda)\big) = \Big(\int_{-\infty}^{\mathcal{T}}\mathrm{d}\epsilon_t\,\mathrm{e}^{t(\mathcal{T}-\epsilon_t)^l} + \int_{\mathcal{T}}^{\infty}\mathrm{d}\epsilon_t\Big)\delta(\epsilon_t-\Lambda) \qquad \text{(B.19)}
$$

Substituting this into the above expression for $G_0$ leads to equation (6.19) in the main text.

# Bibliography

[1] Abramowitz M., Stegun I. A. (Ed) *Handbook of Mathematical Functions* New York. Dover Publications. 1967

[2] Bäck T. *Selective Pressure in Evolutionary Algorithms: A Characterization of Selection Mechanisms* Proc. of the 1st IEEE Conf. on Evolutionary Computing (ICEC94) p 57–62, 1994

[3] Baker E. J. *Reducing Bias and Inefficiency in the Selection Algorithm* Proc. of the 2nd Int. Conf. on Genetic Algorithms Hillsdale, NJ. Lawrence Erlbaum Associates. p 14–21, 1987

[4] Baum E. B., Boneh D., Grant C. *Where Genetic Algorithms Excel* NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. (1995)

[5] Blickle T., Thiele L. *A Comparison of Selection Schemes used in Genetic Algorithms* Computer Engineering and Communication Network Lab, Swiss Federal Institute of Technology, Gloriastrasse 35, 8092 Zurich, Switzerland TIK-Report Nr.11 Version 2 (1995)

[6] Bouten M. *Replica Symmetry Instability in Perceptron Models* J. Phys. **A 27**, 6021–6023 (1994)

[7] Bulmer M. G. *The Mathematical Theory of Quantitative Genetics* Oxford. Clarendon Press. 1980

[8] Davis L. (Ed) *Handbook of Genetic Algorithms* New York. Van Nostrand Reinhold. 1991

[9] De la Maza M., Tidor B. *Increased Flexibility in Genetic Algorithms: The Use of Variable Boltzmann Selective Pressure to Control Propagation* Proc. of the ORSA CSTS Conf. - Computer Science and Operations Research: New Developments in their Interfaces p 425–440, 1991

[10] Derrida B. *Random-energy Model: an Exactly Solvable Model of Disordered Systems* Phys. Rev. **B 24**, 2613-2625 (1981)

[11] Ewens W. J. *Mathematical Population Genetics* Berlin. Springer-Verlag. 1979

[12] Falconer D. S. *Introduction to Quantitative Genetics* Burnt Mill, England. Longman Scientific and Technical. 1989

[13] Fitzpatrick J. M., Grefenstette J. J. *Genetic Algorithms in Noisy Environments* Machine Learning **3**, 101–120 (1988)

[14] Forrest S., Mitchell M. *What Makes a Problem Hard for a Genetic Algorithm? Some Anomalous Results and Their Explanation* Machine Learning **13**, 285–319 (1993)

[15] Gardner E. *Maximum Storage Capacity in Neural Networks* Europhys. Lett. **4**(4), 481-485 (1987)

[16] Gardner E., Derrida B. *Optimal Storage Properties of Neural Networks* J. Phys. **A 21**, 271-284 (1988)

[17] Garey M. R., Johnson D. S. *Computers and Intractability – A Guide to the Theory of NP-Completeness* San Francisco. W. H. Freeman and Co. 1979

[18] Goldberg D. E. *Genetic Algorithms in Search, Optimisation and Machine Learning* Massachusetts. Addison-Wesley. 1989

[19] Goldberg D. E., Deb K. *A Comparative Analysis of Selection Schemes Used in Genetic Algorithms* Foundations of Genetic Algorithms. California. Morgan Kaufmann. p 69–93, 1991

[20] Goldberg D. E., Deb K., Clark J. H. *Genetic Algorithms, Noise and the Sizing of Populations* Complex Systems **6**, 333–362 (1992)

[21] Goldberg D. E., Richardson J. *Genetic Algorithms with Sharing for Multimodal Function Optimization* Proc. of the 2nd Int. Conf. on Genetic Algorithms. Hillsdale, NJ. Lawrence Erlbaum Associates. p 41–49, 1987

[22] Goldberg D. E., Smith R. E. *Nonstationary Function Optimization using Genetic Algorithms with Dominance and Diploidy* Proc. of the 2nd Int. Conf. on Genetic Algorithms. Hillsdale, NJ. Lawrence Erlbaum Associates. p 59–68, 1987

[23] Gould S. J., Eldredge N. *Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered* Paleobiology **3**, 115–151 (1977)

[24] Gradshteyn I. S., Ryzhik I. M. *Table of Integrals, Series and Products - Corrected and Enlarged Edition* London. Academic Press Inc. 1980

[25] Grefenstette J. J. *Deception Considered Harmful* Foundations of Genetic Algorithms 2. California. Morgan Kaufmann. p 75, 1993

[26] Györgyi G. *First-order Transition to Perfect Generalization in a Neural Network with Binary Synapses* Phys. Rev. **A 41**(12), 7079–7100 (1990)

[27] Hamilton W. D. *Haploid Dynamic Polymorphism in a Host with Matching Parasites: Effects of Mutation/Subdivision, Linkage, and Patterns of Selection* J. Heredity **84**, 328–338 (1993)

[28] van Hemmen J. L., Palmer R. G. *The Replica Method and a Solvable Spin Glass Model* J. Phys. **A 12**, 563–580 (1979)

[29] Hertz J., Krogh A., Palmer R. G. *Introduction to the Theory of Neural Computation.* California. Addison-Wesley. 1991

[30] Higgs P. G. *Error Thresholds and Stationary Mutant Distributions in Multi-locus Diploid Genetic Models* Genet. Res. **63**, 63–78 (1994)

[31] Hillis W. D. *Co-evolving Parasites Improve Simulated Evolution as an Optimization Procedure* Physica D **42**, 228–234 (1990)

[32] Holland J. H. *Adaptation in Natural and Artificial Systems* The University of Michigan Press. Ann Arbor. 1975

[33] Hollstein R. B. *Artificial Genetic Adaptation in Computer Control Systems* Doctoral Dissertation, University of Michigan. 1971

[34] Horner H. *Dynamics of Learning and Generalization in Perceptrons with Constraints* Physica A **200**, 552–562 (1993)

[35] Köhler H. *Adaptive Genetic Algorithm for the Binary Perceptron Problem* J. Phys. **A 23**, L1265 (1990)

[36] Kondrashov A. S., Crow J. F. *Haploidy or Diploidy: Which is Better?* Nature **351** 314–317 (1991)

[37] Krauth W., Mézard M. *Storage Capacity of Memory Networks with Binary Couplings* J. Phys. France **50**, 3057–3066 (1989)

[38] Kroger B., Vornberger O. *Enumerative vs Genetic Optimization – Two Parallel Algorithms for the Bin Packing Problem* Lecture Notes in Computer Science **594**, 330–362 (1992)

[39] Lively C. M., Howard R. S., *Selection by Parasites for Clonal Diversity and Mixed Mating* Phil. Trans. R. Soc. Lond. **346**, 271–281 (1994)

[40] Marsden J. E. *Basic Complex Analysis* San Francisco. W. H. Freeman and Company. 1973

[41] McCulloch W. S., Pitt W. *A Logical Calculus of Ideas Immanent in Nervous Activity* Bull. of Math. Biophysics **5**, 115–133 (1943)

[42] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E. *Equation of State Calculations for Fast Computing Machines* J. Chem. Phys. **21**, 1087–1092 (1953)

[43] Mézard M., Parisi G. *Spin Glass Theory and Beyond* Singapore. World Scientific. 1987

[44] Miller B. L., Goldberg D. E. *Genetic Algorithms, Selection Schemes and the Varying Effects of Noise* Dept. of General Engineering, University of Illinois at Urbana-Champaign, 117 Transportation Building, Urbana, IL 61801. (IlliGAL Report No. 95009) (1995)

[45] Mühlenbein H., Schlierkamp-Voosen D. *Analysis of Selection, Mutation and Recombination in Genetic Algorithms* Lecture Notes in Computer Science **899**, 188-214 (1995)

[46] van Nimwegen E., Crutchfield J. P., Mitchell M. *Finite Populations Induce Metastability in Evolutionary Search* submitted to Phys. Lett. **A** (1996)

[47] Nix A., Vose M. D. *Modelling Genetic Algorithms with Markov Chains* Annals of Mathematics and Artificial Intelligence **5**, 79-88 (1991)

[48] Patel K. *Computational Complexity, Learning Rules and Storage Capacities: Monte Carlo Study for the Binary Perceptron* J. Phys. **B 91**, 257-266 (1993)

[49] Pettey C. C., Leuze M. R., Grefenstette J. J. *A Theoretical Investigation of a Parallel Genetic Algorithm* Proc. of the 2nd Int. Conf. on Genetic Algorithms. Hillsdale, NJ. Lawrence Erlbaum Associates. p 155–161, 1987

[50] Pitt L., Vailant L. G. *Computational Limitations on Learning from Examples* J. ACM. **35**(4), 965–984 (1988)

[51] Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T. *Numerical Recipes in C: The Art of Scientific Computing* 2nd Ed. Cambridge. Cambridge University Press, 1992

[52] Prügel-Bennett A. *Modelling Evolving Populations* NORDITA, Blegdamsvej 17, DK-2100 Copenhagen, Denmark, (submitted to J. Theor. Biol.) (1996)

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

[54] Prügel-Bennett A., Shapiro J. L. *The Dynamics of a Genetic Algorithm for the Ising Spin-Glass Chain* Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, UK (submitted to Physica D for publication) (1995)

[55] Rattray L. M. *An Analysis of a Genetic Algorithm Training the Binary Perceptron* MSc. Thesis, Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, UK. 1994

[56] Rattray L. M. *The Dynamics of a Genetic Algorithm under Stabilizing Selection* Complex Systems **9**, 213–234 (1995)

[57] Rattray L. M., Shapiro J. L. *Noisy Fitness Evaluation in Genetic Algorithms and the Dynamics of Learning* to appear in Foundations of Genetic Algorithms 4. California. Morgan Kaufmann. 1997

[58] Rattray L. M., Shapiro J. L. *The Dynamics of a Genetic Algorithm for a Simple Learning Problem* Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, UK (to appear in J. Phys. **A**) (1996)

[59] Saad D., Solla S. A. *On-Line Learning in Soft Committee Machines* Phys. Rev. **E 52**(4), 4225–4243 (1995)

[60] Schaffer J. D., Whitley D., Eshelman L. J. *Combinations of Genetic Algorithms and Neural Networks: A Survey of the State of the Art* Proc. of the Int. Conf. on Combinations of Genetic Algorithms and Neural Networks (IEEE Computer Society Press, Los Alamitos, CA) p 1–37, 1992

[61] Schwarze H. *Learning a Rule in a Multi-layer Neural-network* J. Phys. **A 26**(21), 5781–5794 (1993)

[62] Seung H. S., Sompolinsky H. *Statistical Mechanics of Learning from Examples* Phys. Rev. **A 45**(8), 6056–6091 (1992)

[63] Shapiro J. L., Prügel-Bennett A., Rattray L. M. *A Statistical Mechanical Formulation of the Dynamics of Genetic Algorithms* Lecture Notes in Computer Science **865**, 17–27 (1994)

[64] Shapiro J. L., Prügel-Bennett A. *Genetic Algorithm Dynamics in Two-well Potentials with*

*Basins and Barriers* to appear in Foundations of Genetic Algorithms 4. California. Morgan Kaufmann. 1997

[65] Smith R. E., Goldberg D. E. *Diploidy and Dominance in Artificial Genetic Search* Complex Systems **6**, 251–285 (1992)

[66] Spears W. M., De Jong K. A. *Analyzing GAs using Markov Models with Semantically Ordered and Lumped States* to appear in Foundations of Genetic Algorithms 4. California. Morgan Kaufmann. 1997

[67] Srinivas M., Patnaik L. M. *Binomially Distributed Populations for Modelling GAs* Proc. of the 5th Int. Conf. on Genetic Algorithms. Hillsdale, NJ. Lawrence Erlbaum Associates. p 138–145, 1993

[68] Srinivas M., Patnaik L. M. *Genetic Search: Analysis using Fitness Moments* IEEE Trans. on Knowledge and Data Engineering **8**(1), 120–133 (1996)

[69] Stein D. L., Newman C. M. *Broken Ergodicity and the Geometry of Rugged Landscapes* Phys. Rev. **E 51**(6A), 5228–5238 (1995)

[70] Stuart A., Ord J. K. *Kendall's Advanced Theory of Statistics, Vol 1. Distribution Theory* Ed 5. New York. Oxford University Press. 1987

[71] Syswerda G. *Simulated Crossover in Genetic Algorithms* Foundations of Genetic Algorithms 2. California. Morgan Kaufmann. 1993

[72] Thierens D., Goldberg D. E. *Convergence Models of Genetic Algorithm Selection Schemes* Parallel Problem Solving from Nature III, Lecture Notes in Computer Science **866**, 119–129 (1994)

[73] Vose M. D. *Modelling Simple Genetic Algorithms* Foundations of Genetic Algorithms 2. p 63. California. Morgan Kaufmann. 1993

[74] Vose M. D., Liepins G. E. *Punctuated Equilibria in Genetic Search* Complex Systems **5**, 31-44 (1991)

[75] Vose M. D., Wright A. H. *Simple Genetic Algorithms with Linear Fitness* Evol. Comp. **2**, 347–368 (1994)

[76] Wolfram S. *Mathematica. A System for Doing Mathematics by Computer* 2nd Edition. Massachusetts. Addison-Wesley. 1992

[77] Yukiko Y., Adachi N. *A Diploid Genetic Algorithm for Preserving Population Diversity – pseudo-Meiosis GA* Parallel Problem Solving from Nature III, Lecture notes in Computer Science **866**, 36–45 (1994)