# Noisy Fitness Evaluation in Genetic Algorithms and the Dynamics of Learning

**Magnus Rattray** * **and Jonathan Shapiro** †
Computer Science Department,
University of Manchester,
Oxford Road,
Manchester M13 9PL, U.K.

## Abstract

A theoretical model is presented which describes selection in a genetic algorithm (GA) under a stochastic fitness measure and correctly accounts for finite population effects. Although this model describes a number of selection schemes, we only consider Boltzmann selection in detail here as results for this form of selection are particularly transparent when fitness is corrupted by additive Gaussian noise. Finite population effects are shown to be of fundamental importance in this case, as the noise has no effect in the infinite population limit. In the limit of weak selection we show how the effects of any Gaussian noise can be removed by increasing the population size appropriately. The theory is tested on two closely related problems: the one-max problem corrupted by Gaussian noise and generalization in a perceptron with binary weights. The averaged dynamics can be accurately modelled for both problems using a formalism which describes the dynamics of the GA using methods from statistical mechanics. The second problem is a simple example of a learning problem and by considering this problem we show how the accurate characterization of noise in the fitness evaluation may be relevant in machine learning. The training error (negative fitness) is the number of misclassified training examples in a batch and can be considered as a noisy version of the generalization error if an independent batch is used for each evaluation. The noise is due to the finite batch size and in the limit of large problem size and weak selection we show how the effect of this noise can be removed by increasing the population size. This allows the optimal batch size to be determined, which minimizes computation time as well as the total number of training examples required.

* Internet address: rattraym@cs.man.ac.uk.
† Internet address: jls@cs.man.ac.uk.

# 1 INTRODUCTION

Genetic algorithms (GAs) are adaptive search techniques which can be used to find good solutions to problems with poorly characterized and high-dimensional search spaces (Goldberg, 1989; Holland, 1975). They have been successfully applied in a large range of domains, including a number of machine learning problems. The GA differs from other stochastic search techniques, such as simulated annealing, because solutions evolve in parallel within a population. It is hoped that this may lead to improvement through the recombination of mutually useful features from different population members.

The relative merit of each solution within the population is usually determined through a fitness measure. The fitness evaluation may be noisy due to measurement limitations or incomplete training data and it is important to understand and predict the effects of such noise. In some machine learning and optimization applications there may be a tradeoff between improved fidelity in evaluating fitness and the increased computational cost this requires. It has been suggested that GAs are suitable in this domain, since they are relatively robust against the effects of noise (Fitzpatrick & Grefenstette, 1988). Indeed, GAs have recently been shown to deal better with noise than competing local search algorithms on a class of simple additive problems (Baum *et al*, 1995).

In (Miller & Goldberg, 1995), noise corrupted fitness was modelled in terms of its effect on the mean fitness after selection from a continuous and Gaussian distribution of fitness. This is effectively an infinite population assumption and leads to the conclusion that proportionate selection is unaffected by noise. In a finite population, the tails of the distribution will be sparsely populated and this will prove to be of fundamental importance when accounting for the effects of noise. Although Miller and Goldberg sized the population to account for increased finite population effects due to noise, their choice of population size was based on a conservative predictor rather than an exact result (Goldberg *et al*, 1992). Their calculation of the variance for the one-max domain assumes a binomial distribution of alleles within the population and this assumption is also made in a number of other predictive models (Mühlenbein & Schlierkamp-Voosen, 1995; Srinivas & Patnaik, 1995; Thierens & Goldberg, 1995). In a finite population this assumption breaks down, because the population becomes more correlated under selection than predicted by a binomial distribution and this results in a reduced variance.

In this work, a theoretical model is presented which describes selection under a general stochastic fitness measure and correctly accounts for finite population effects. Although this model can be applied to a number of selection schemes and noise distributions, Boltzmann selection is considered in greatest detail here as the results in this case are transparent. This is not the most common selection scheme used in GAs, but it seems an appropriate scheme for problems where the distribution of fitness is close to Gaussian, as it conserves the population's shape in this case. It is also easy to choose the selection strength so that the population makes continued progress under selection. For weak Boltzmann selection and Gaussian noise, it is shown how an increase in population size removes the effects of noise on selection. Noise only affects a finite population under this form of selection, which emphasizes the need for any theory to properly account for finite population effects.

The theory is applied to two problems for which the full dynamics can be solved, extending a formalism developed by Prügel-Bennett, Shapiro, and Rattray for modelling the dynamics of the GA using methods from statistical mechanics (Prügel-Bennett & Shapiro, 1994; Prügel-Bennett & Shapiro, 1995; Rattray, 1995; Rattray & Shapiro, 1996; Shapiro *et al*, 1994). This formalism does not require that the population be sufficiently large to ensure convergence to the global optimum and properly accounts for correlations accumulated under selection. Under this formalism, the population is described by a small number of macroscopic statistics and a maximum entropy assumption is used to determine anything not trivially related to these macroscopics. Difference equations are derived which determine the mean change to each macroscopic under each genetic operator and these can be iterated in sequence to simulate the averaged dynamics. A more exact approach also follows fluctuations from mean behaviour by following an ensemble of populations (Prügel-Bennett, 1996). However, mean behaviour alone is sufficient to accurately describe the problems under consideration here. The macroscopics which have proved most successful to date are cumulants of some appropriate quantity within the population and the mean correlation (closely related to the mean Hamming distance). The first two cumulants are the mean and variance respectively while higher cumulants describe deviations from a Gaussian distribution.

The first case considered is the one-max problem corrupted by Gaussian noise. To simplify the discussion, bit-simulated crossover is used (Syswerda, 1993) and this allows the dynamics to be modelled by iterating only two macroscopics: the mean fitness and correlation within the population. A maximum entropy assumption is required to determine the higher cumulants before selection and to evolve the correlation under selection. Relevant results from other studies are reproduced where necessary in order to make the discussion self-contained. Simulation results show very good correspondence to the theory for a range of noise strengths and the theory accurately predicts the evolution of each macroscopic, averaged over many runs of the GA.

The second case considered is a simple problem from learning theory, generalization by a binary perceptron. A perceptron with binary weights is trained to learn a teacher perceptron by training on examples produced by the teacher. This has previously been shown to be equivalent to a noisy version of one-max if a new batch of examples are presented each time the training error is calculated (Baum *et al*, 1995). This problem was solved under the statistical mechanics formalism in (Rattray & Shapiro, 1996) and those results are reviewed here. The training error is well approximated by a Gaussian distribution whose mean is the the generalization error and whose variance increases as the batch size is reduced. The theory is shown to agree closely with simulation results averaged over many runs of the GA. In the limit of large problem size and weak selection an increase in population size removes the effects of noise due to the finite size of each training batch and this allows the optimal batch size to be determined.

## 2 NOTATION

Notation will follow GA conventions where appropriate and therefore differs from a number of related publications which use conventions from statistical physics (Prügel-Bennett & Shapiro, 1994; Prügel-Bennett & Shapiro, 1995; Rattray, 1995; Rattray & Shapiro, 1996). The population size is $N$ and each population member, labelled $\alpha$, has two associated fitness measures. The ideal fitness $f_\alpha$ is some deterministic function of the genotype, while the noisy fitness $F_\alpha$ is related to this through a conditional probability distribution $p(F|f)$. For example, in a supervised learning problem $f_\alpha$ might be the fitness evaluated over all possible training examples and $F_\alpha$ might be the best estimate given a small training batch. When we refer to the fitness this will usually be the ideal fitness and the noisy fitness will always be referred to explicitly.

In the cases under consideration here, each population member's genotype is a string of binary alleles of length $l$. The usual convention in GA theory is to take alleles $x_i^\alpha \in \{0,1\}$ where $i$ labels the site and $\alpha$ labels the population member. Here, however, we choose alleles $S_i^\alpha \in \{-1,1\}$ which are more appropriate for the binary perceptron problem. A trivial change in variables maps one convention onto the other.

### 2.1 CUMULANTS

Throughout this paper the population will be described by a number of macroscopic variables, the cumulants of the ideal fitness distribution within the population and the mean correlation within the population. Cumulants are statistics which describe the population shape and are often reasonably stable to fluctuations between runs of the GA, so that they average well (Prügel-Bennett & Shapiro, 1995). The first two cumulants are the mean and variance respectively, while higher cumulants describe deviations from a Gaussian distribution. The third and fourth cumulants are related to the skewness and kurtosis of the population, respectively. The $n$th cumulant of a finite population is denoted $\kappa_n$.

If $f_\alpha$ is the fitness of population member $\alpha$ then the cumulants of fitness within a finite population are given by,

$$\kappa_n = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \ln Z; \qquad Z = \sum_{\alpha=1}^{N} \mathrm{e}^{\gamma f_\alpha}. \tag{1}$$

Here, $Z$ is called the partition function and holds all the information required to determine the population's cumulants. So, for example, the first two cumulants (the mean and variance) are,

$$\kappa_1 = \lim_{\gamma \to 0} \frac{\sum_\alpha f_\alpha \mathrm{e}^{\gamma f_\alpha}}{\sum_\alpha \mathrm{e}^{\gamma f_\alpha}} = \frac{1}{N} \sum_{\alpha=1}^{N} f_\alpha \tag{2a}$$

$$\begin{aligned}\kappa_2 &= \lim_{\gamma \to 0} \frac{\sum_\alpha f_\alpha^2 \mathrm{e}^{\gamma f_\alpha} \left(\sum_\alpha \mathrm{e}^{\gamma f_\alpha}\right) - \left(\sum_\alpha f_\alpha \mathrm{e}^{\gamma f_\alpha}\right)^2}{\left(\sum_\alpha \mathrm{e}^{\gamma f_\alpha}\right)^2} \\ &= \frac{1}{N} \sum_{\alpha=1}^{N} (f_\alpha)^2 - \left(\frac{1}{N} \sum_{\alpha=1}^{N} f_\alpha\right)^2. \end{aligned} \tag{2b}$$

In order to model selection on a finite population, $N$ population members are randomly sampled from an infinite population before selection (this procedure is described in greater detail in section 3). It is well known that the expected variance of a finite sample is reduced by a factor of $1 - 1/N$ and similar corrections occur for the higher cumulants. If $K_n$ is the $n$th cumulant of an infinite population, then expectation values for the first four cumulants of a finite sample are given by,

$$\kappa_1 = K_1 \tag{3a}$$

$$\kappa_2 = \mathcal{N}_2 K_2 \tag{3b}$$

$$\kappa_3 = \mathcal{N}_3 K_3 \tag{3c}$$

$$\kappa_4 = \mathcal{N}_4 K_4 - 6\mathcal{N}_2 (K_2)^2/N. \tag{3d}$$

Here, $\mathcal{N}_2$, $\mathcal{N}_3$ and $\mathcal{N}_4$ give the finite population corrections (Prügel-Bennett & Shapiro, 1995),

$$\mathcal{N}_2 = 1 - \frac{1}{N} \qquad \mathcal{N}_3 = 1 - \frac{3}{N} + \frac{2}{N^2} \qquad \mathcal{N}_4 = 1 - \frac{7}{N} + \frac{12}{N^2} - \frac{6}{N^3}\,. \tag{4}$$

If $p(f)$ is the distribution of fitness in an infinite population, then the infinite population cumulants can be generated from a characteristic function (analogous to the partition function)[1],

$$K_n = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \ln \rho(\gamma); \qquad \rho(\gamma) = \int \mathrm{d}f\, p(f) \mathrm{e}^{\gamma f}. \tag{5}$$

[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 can also be written in terms of a cumulant expansion,

$$\rho(\gamma) = \exp\left(\sum_{n=1}^{\infty} \frac{K_n \gamma^n}{n!}\right). \tag{6}$$

It is often useful to parameterize the fitness distribution by expanding around a Gaussian distribution. In this case we choose a Gram-Charlier expansion (see, for example, (Stuart & Ord, 1987)),

$$p(f) = \frac{1}{\sqrt{2\pi K_2}} \exp\left(\frac{-(f-K_1)^2}{2K_2}\right) \left[1 + \sum_{n=3}^{n_c} \frac{K_n}{n! K_2^{n/2}} H_n\left(\frac{f-K_1}{\sqrt{K_2}}\right)\right], \tag{7}$$

where $H_n(x) = (-1)^n \mathrm{e}^{x^2/2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2/2}$ are Hermite polynomials and $n_c$ is the number of cumulants used. Four cumulants were used in this work and the third and fourth Hermite polynomials are $H_3(x) = (x^3 - 3x)$ and $H_4(x) = (x^4 - 6x^2 + 3)$. This 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.

### 2.2 CORRELATION

The correlation is a measure of genotype similarity. The simplest measure of correlation between two population members, $\alpha$ and $\beta$, is given by,

$$q_{\alpha\beta} = \frac{1}{l} \sum_{i=1}^{l} S_i^\alpha S_i^\beta. \tag{8}$$

Recall that $S_i^\alpha \in \{-1, 1\}$ so that this quantity equals one when two population members are identical and is zero on average for two randomly generated population members. This is closely related to the Hamming distance between two binary sequences. To get the mean correlation within the population one averages this quantity over each distinct pair of population members,

$$q = \langle q_{\alpha\beta} \rangle_{\alpha \neq \beta} = \frac{2}{N(N-1)} \sum_{\alpha=1}^{N} \sum_{\beta > \alpha} q_{\alpha\beta}. \tag{9}$$

The expected correlation of a finite sample is equal to the correlation in an infinite population.

## 3 SELECTION

To describe a general selection scheme it is instructive to separate the sampling process from the weighting process. Each population member is assigned a selection weight $w_\alpha$, which is generally some non-decreasing function of fitness (this is the measured, noisy fitness $F_\alpha$). For fitness proportionate selection the selection weight is simply equal to the fitness. Selection weights can also be defined for ranking, tournament and truncation selection, and the general method described here can be applied to these cases (Rattray, 1996). We will consider Boltzmann selection in greatest detail, as this provides a transparent result for Gaussian noise (De la Maza & Tidor, 1991; Prügel-Bennett & Shapiro, 1994). For Boltzmann selection the selection weight is defined,

$$w_\alpha = \exp(\beta F_\alpha), \tag{10}$$

where $\beta$ is the selection strength which determines the relative probability of selection for each population member. By scaling the selection strength inversely with the population's standard deviation one avoids the problem of long convergence times, often cited as a problem with using fitness-proportionate forms of selection. This scaling is used in section 3.3.

To select a new population it is necessary to take a weighted sample from the population before selection. Ideally, the proportion of each population member in the new population is given by,

$$p_\alpha = \frac{w_\alpha}{\sum_\alpha w_\alpha}. \tag{11}$$

However, it is not possible to choose exactly this amount in a finite population. We will consider Roulette wheel sampling, as this provides an analytically tractable model for finite population effects. Other, less noisy forms of sampling are often preferred in practice (see, for example, (Baker, 1987)) and a challenging task would be to extend the present analysis to these cases.

Under Roulette wheel sampling, $N$ new population members are selected with replacement, with probability $p_\alpha$. Following the discussion in (Prügel-Bennett, 1996), this process can be divided into two stages,

1. Select an infinite population from a finite population, so that $p_\alpha$ is exactly the proportion of population member $\alpha$ in the infinite population after selection.
2. Randomly sample $N$ population members from the infinite population to make up the new finite population.

Mutation and crossover do not involve sampling and can therefore be carried out during the infinite population stage of the dynamics without any loss of generality. A similar sampling procedure is used in (Vose & Wright, 1994), but there they follow an exact microscopic description of the population while we only consider a small number of macroscopic statistics. This simplification makes our prescription less general, but allows us to capture a number of interesting and non-trivial features of the dynamics in a natural way.

### 3.1 GENERATING THE CUMULANTS AFTER SELECTION

The cumulants of an infinite population after selection can be generated from the logarithm of a selection partition function. If $K_n^s$ is the $n$th cumulant of an infinite population after selection then,

$$K_n^s = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \ln Z_s \, ; \qquad Z_s = \sum_{\alpha=1}^{N} w_\alpha e^{\gamma f_\alpha}. \tag{12}$$

For example,

$$\begin{aligned} K_1^s &= \lim_{\gamma \to 0} \frac{\sum_\alpha w_\alpha f_\alpha e^{\gamma f_\alpha}}{\sum_\alpha w_\alpha e^{\gamma f_\alpha}} \\ &= \sum_{\alpha=1}^{N} p_\alpha f_\alpha, \end{aligned} \tag{13a}$$

where we have used the definition of $p_\alpha$ in equation (11). Similarly one finds,

$$K_2^s = \sum_{\alpha=1}^{N} p_\alpha f_\alpha^2 - \left( \sum_{\alpha=1}^{N} p_\alpha f_\alpha \right)^2. \tag{13b}$$

These are exactly the cumulants of an infinite population after selection, since $p_\alpha$ is exactly the proportion of each population member in this case. The expected cumulants of a finite population after selection can be found by applying equations (3a) to (3d).

The exact fitness of each population member is not known in general, only the cumulants of the distribution from which they are sampled. The selection weight also has a stochastic component due to variance in the conditional probability distribution relating the measured fitness to the ideal fitness $p(F|f)$. Therefore, it is necessary to average over the sampling procedure and the noise in fitness evaluation in order to determine the expected cumulants after selection. Instead of averaging over the cumulants directly, it is more convenient to average over the logarithm of the partition function defined in equation (12),

$$\langle \ln Z_{\mathrm{s}} \rangle = \left( \prod_{\alpha=1}^{N} \int \mathrm{d}f_\alpha \, p(f_\alpha) \int \mathrm{d}F_\alpha \, p(F_\alpha | f_\alpha) \right) \ln Z_{\mathrm{s}}. \tag{14}$$

Following the discussion in (Prügel-Bennett & Shapiro, 1994) we use Derrida's trick to express the logarithm as an integral[2] (Derrida, 1981).

$$\langle \ln Z_{\mathrm{s}} \rangle = \int_0^\infty \mathrm{d}t \, \frac{\mathrm{e}^{-t} - \langle \mathrm{e}^{-tZ_{\mathrm{s}}} \rangle}{t}. \tag{15}$$

If the selection weight associated with population member $\alpha$ is a function of $F_\alpha$ alone then the averages on the right hand side decouple from one another,

$$\begin{aligned} \langle \mathrm{e}^{-tZ_{\mathrm{s}}} \rangle &= \left( \prod_{\alpha=1}^{N} \int \mathrm{d}f_\alpha \, p(f_\alpha) \int \mathrm{d}F_\alpha \, p(F_\alpha | f_\alpha) \right) \exp\left( -t \sum_{\alpha=1}^{N} w(F_\alpha) \mathrm{e}^{\gamma f_\alpha} \right) \\ &= \left( \int \mathrm{d}f \, p(f) \int \mathrm{d}F \, p(F|f) \exp\left( -t w(F) \mathrm{e}^{\gamma f} \right) \right)^N. \end{aligned} \tag{16}$$

### 3.2 BOLTZMANN SELECTION

Consider Boltzmann selection, in which case the selection weight is defined in equation (10). The above expression can be substituted into equation (15) and equation (12) then provides the expected cumulants of an infinite population after Boltzmann selection (the term in the integrand of equation (15) which does not involve $\gamma$ is not required for $n > 0$),

$$K_n^{\mathrm{s}} = - \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \int_0^\infty \mathrm{d}t \, \frac{g^N(t, \gamma, \beta)}{t}, \tag{17}$$

where

$$g(t, \gamma, \beta) = \int \mathrm{d}f \, p(f) \int \mathrm{d}F \, p(F|f) \exp\left( -t \mathrm{e}^{\beta F + \gamma f} \right). \tag{18}$$

Notice that although these are cumulants of an infinite population after selection, they depend on $N$ which is the population size *before* selection (see section 3). In general these integrals have to be determined numerically and for the simulation results presented in this paper the integrals were computed by Gaussian quadratures (Press *et al*, 1992). The ideal fitness distribution can be parameterized by the Gram-Charlier expansion in equation (7).

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

### 3.3 WEAK SELECTION AND GAUSSIAN NOISE

An analytically tractable case is for weak selection corrupted by additive Gaussian noise, in which case $p(F|f)$ is given by a Gaussian distribution centred around $f$,

$$p(F|f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(F-f)^2}{2\sigma^2}\right). \tag{19}$$

Here, $\sigma^2$ is the variance of the noise. As shown in (Prügel-Bennett & Shapiro, 1994), one can express the logarithm of the partition function analytically for small $\beta$. This limit is accurate for sufficiently small $\beta\sqrt{K_2+\sigma^2}$ and is instructive as it shows the relevant effects of selection for each cumulant.

For small $\beta$ and $\gamma$, $g(t,\gamma,\beta)$ which is defined in equation (18) can be expanded in $t e^{\beta F+\gamma f}$. Exponentiating this expansion one finds,

$$g^N(t,\gamma,\beta) \simeq \exp(-tN\psi(\gamma,\beta)) \left(1+\frac{Nt^2}{2}\left(\psi(2\gamma,2\beta)-\psi^2(\gamma,\beta)\right)\right), \tag{20}$$

where

$$\begin{aligned} \psi(\gamma,\beta) &= \int \mathrm{d}f\, p(f) \int \mathrm{d}F\, p(F|f)\, e^{\beta F+\gamma f} \\ &= \exp\left(\tfrac{1}{2}(\beta\sigma)^2\right) \rho(\beta+\gamma). \end{aligned}$$

Here, $\rho(\gamma)$ is the characteristic function defined in equation (5) which can be written in terms of the cumulant expansion defined in equation (6). Completing the integral in equation (17), one finds that the cumulants after selection up to $O(1/N)$ are given by,

$$K_n^s = \lim_{\gamma\to 0} \frac{\partial^n}{\partial\gamma^n}\left[\sum_{i=1}^{\infty} \frac{(\gamma+\beta)^i K_i}{i!} - \frac{e^{(\beta\sigma)^2}}{2N} \exp\left(\sum_{i=1}^{\infty} \frac{(2^i-2)(\gamma+\beta)^i K_i}{i!}\right) + O\left(\frac{1}{N}\right)\right]. \tag{21}$$

The leading term here is the infinite population result.

Expanding the first three cumulants after selection in $\beta$, for fixed $\beta\sigma$, one finds,

$$K_1^s = K_1 + \beta\left(1-\frac{e^{(\beta\sigma)^2}}{N}\right)K_2 + \frac{\beta^2}{2}\left(1-\frac{3e^{(\beta\sigma)^2}}{N}\right)K_3 + \cdots \tag{22a}$$

$$K_2^s = \left(1-\frac{e^{(\beta\sigma)^2}}{N}\right)K_2 + \beta\left(1-\frac{3e^{(\beta\sigma)^2}}{N}\right)K_3 + \cdots \tag{22b}$$

$$K_3^s = \left(1-\frac{3e^{(\beta\sigma)^2}}{N}\right)K_3 + \beta\left[\left(1-\frac{7e^{(\beta\sigma)^2}}{N}\right)K_4 - \frac{6e^{(\beta\sigma)^2}}{N}(K_2)^2\right] + \cdots. \tag{22c}$$

For zero noise ($\sigma = 0$) one retrieves the result in (Prügel-Bennett & Shapiro, 1994). As in the zero noise case, finite population effects lead to a reduced variance and a negative third cumulant[3], related to the population's skewness, which leads to an accelerated reduction in variance under further selection. Notice that a normally distributed infinite population remains Gaussian under selection and does not lose variance. This is clearly an idealization which cannot be achieved in a finite population,

[3]The third cumulant typically becomes negative even in an infinite population because of an initially negative fourth cumulant (for finite $l$) – however, finite population effects are often more significant.

where tails of the population are sparsely populated and no progress can be made beyond the best solution. The other genetic operators are required to reduce the magnitude of the higher cumulants by repopulating the tails of the population.

The noise in selection increases the magnitude of the finite population terms by reducing the accuracy of sampling, resulting in a faster loss of variance and less improvement under selection. Clearly, noise has no effect in the infinite population limit. This is because the effect of noise over-estimating and under-estimating the value of $f$ exactly cancels in this limit. This again emphasizes the need for accurate characterization of finite population effects.

It can be seen from equation (21) that in the weak selection limit the effects of Gaussian noise can be removed by increasing the population size appropriately. If $N_0$ is the population size for zero noise, then the effects of any Gaussian noise which is introduced can be removed by setting,

$$N = N_0 \exp\left((\beta\sigma)^2\right). \tag{23}$$

It is remarkable that the effects of noise on selection can be removed for *every* cumulant by this simple increase in population size. This population resizing proves to be particularly of interest in the context of the binary perceptron problem discussed in section 5.

Notice that this population resizing only holds if selection strength is independent of the noise variance, so that only finite population terms in equations (21) involve the noise variance. For example, this is not the case if selection strength is scaled according to statistics from the measured, noise corrupted fitness distribution (although the equations describing the dynamics would still hold). Here, we scale selection strength inversely to the standard deviation of the ideal fitness distribution $\beta = \beta_s/\sqrt{\kappa_2}$, which ensures a constant selection pressure. This is a rather artificial choice, as ideal fitness statistics would not be known in a real noise corrupted problem. However, the results derived here describe a GA with any fixed schedule for determining the selection strength each generation. The scaling used here is equivalent (on average) to an appropriate schedule for the associated noiseless problem.

## 4 ONE-MAX WITH GAUSSIAN NOISE

The dynamics for the one-max problem can be modelled using a statistical mechanics formalism developed in (Prügel-Bennett & Shapiro, 1995; Rattray, 1995; Rattray & Shapiro, 1996). This discussion will follow that presented in (Rattray & Shapiro, 1996) most closely. To simplify matters bit-simulated crossover is used, where the population is completely shuffled during crossover so that a child's alleles come from any population member with equal probability (Syswerda, 1993). This brings the population straight to the fixed point of standard uniform crossover (without selection) and allows the population to be accurately described by only two macroscopics: the mean fitness and correlation. Under more general forms of crossover it is necessary to follow the evolution of the higher cumulants, as described in (Prügel-Bennett & Shapiro, 1995; Rattray, 1995). Here, we only wish to consider the simplest GA (from a theoretical perspective) compatible with the problems under consideration.

The formalism used here differs from the models described in (Mühlenbein & Schlierkamp-Voosen, 1995; Srinivas & Patnaik, 1995; Thierens & Goldberg, 1995) by the inclusion of a constraint on the mean correlation within the population. In these models the population was considered to be binomially distributed, and this assumption breaks down when a finite population correlates under selection. This is especially important here, as noise has no effect in the infinite population limit. Unfortunately, the inclusion of an extra constraint means that the dynamic trajectory for the macroscopics can no longer be described analytically. However, the description is still compact in the sense that there are few degrees of freedom and any numerical computation which is required does not depend on population size or genotype length.

In the following sections difference equations are derived for the change in mean fitness and correlation within the population under the action of each genetic operator. To describe the population before selection it is necessary to determine terms which are not trivially related to these two macroscopics. In order to calculate these terms a maximum entropy calculation is introduced, which is described in the appendix. Finally, the theory is compared to simulation results averaged over many runs, showing excellent agreement and accurately predicting the averaged evolution of each macroscopic.

### 4.1 THE MACROSCOPICS

The ideal fitness for one-max is given by,

$$f_\alpha = \sum_{i=1}^{l} S_i^\alpha. \tag{24}$$

Here, the alleles are $S_i^\alpha \in \{-1, 1\}$, which is most convenient for the binary perceptron problem considered in section 5. This can easily be converted to the standard binary convention under a linear transformation. The mean and variance of an infinite population are,

$$K_1 = \sum_{i=1}^{l} \langle S_i^\alpha \rangle_\alpha \tag{25a}$$

$$\begin{aligned} K_2 &= \left\langle \left( \sum_{i=1}^{l} S_i^\alpha \right)^2 \right\rangle_\alpha - \left( \sum_{i=1}^{l} \langle S_i^\alpha \rangle_\alpha \right)^2 \\ &= l(1-q) + \sum_{i=1}^{l} \sum_{j \neq i} \langle S_i^\alpha S_j^\alpha \rangle_\alpha - \langle S_i^\alpha \rangle_\alpha \langle S_j^\alpha \rangle_\alpha, \end{aligned} \tag{25b}$$

where the angled brackets denote population averages and we have used an infinite population expression for the correlation,

$$q = \frac{1}{l} \sum_{i=1}^{l} \langle S_i^\alpha S_i^\beta \rangle_{\alpha \neq \beta} \overset{N \to \infty}{=} \frac{1}{l} \sum_{i=1}^{l} \langle S_i^\alpha \rangle_\alpha^2. \tag{26}$$

The finite population correction to the second cumulant is given in equation (3b).

Equation (25b) shows how an increase in correlation results in a reduced variance, all other terms being equal. The $i \neq j$ term in this expression is related to the linkage disequilibrium in population genetics (Ewens, 1979) and disappears after bit-simulated crossover. In this case the correlation can be deduced directly from the variance after crossover.

### 4.2 MUTATION

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

$$S_i^\alpha \to M_i^\alpha S_i^\alpha; \qquad M_i^\alpha = \begin{cases} 1 & \text{with probability } 1 - p_\mathrm{m} \\ -1 & \text{with probability } p_\mathrm{m}. \end{cases} \tag{27}$$

So, for example, the mean fitness of an infinite population after mutation is,

$$K_1^\mathrm{m} = \sum_{i=1}^{l} \langle M_i^\alpha S_i^\alpha \rangle_\alpha \tag{28}$$

and averaging over all mutations gives the expectation value for the mean after mutation,

$$\langle K_1^\mathrm{m} \rangle = (1 - 2p_\mathrm{m}) K_1. \tag{29}$$

This calculation can be generalized to the higher cumulants (Prügel-Bennett & Shapiro, 1995). The correlation after mutation is similarly found to be,

$$q_\mathrm{m} = (1 - 2p_\mathrm{m})^2 q. \tag{30}$$

## 4.3 CROSSOVER

Under bit-simulated crossover, the population is brought straight to the fixed point of standard uniform crossover (without selection). Notice that averages between and within population members are equal on average after this form of crossover; so, for example, terms like $\langle S_i^\alpha S_j^\beta \rangle_{i \neq j}$ and $\langle S_i^\alpha S_j^\alpha \rangle_{i \neq j}$ are equal (where brackets now denote site averages) and the second term in equation (25b) disappears. Similar cancellations are possible in the higher cumulants, as described in (Prügel-Bennett & Shapiro, 1995). To accurately model selection we describe the population by four cumulants after crossover,

$$K_1^\mathrm{c} = K_1 \tag{31a}$$

$$K_2^\mathrm{c} = l(1 - q) \tag{31b}$$

$$K_3^\mathrm{c} = -2K_1 + 2\sum_{i=1}^{l} \langle S_i^\alpha \rangle_\alpha^3 \tag{31c}$$

$$K_4^\mathrm{c} = -2l(1 - 4q) - 6\sum_{i=1}^{l} \langle S_i^\alpha \rangle_\alpha^4. \tag{31d}$$

The terms in the expressions for the third and fourth cumulants which are not trivially related to known macroscopics are calculated through a maximum entropy assumption, as described in the appendix. The correlation does not change under crossover, since the mean number of alleles at each site is conserved. In (Prügel-Bennett & Shapiro, 1995) it is shown how the cumulants relax towards this fixed point under more standard crossover schemes.

## 4.4 SELECTION

The cumulants after Boltzmann selection are given in equation (17). It only remains to calculate the correlation after selection. This is a difficult task in general, as it requires some knowledge of the mapping between genotype and fitness and we will again make use of the maximum entropy calculation described in the appendix.

It is instructive to divide the correlation after selection into two contributions: a duplication term and a natural increase term. The duplication term gives the increased correlation due to the duplication of existing population members required in a finite population. The natural increase term is due to the natural increase in correlation as the population moves into a region of higher fitness. The following results were derived in full in (Rattray, 1995; Rattray & Shapiro, 1996) and here we only provide an outline of the derivation.

The correlation in an infinite population after selection is,

$$\begin{aligned} q_{\mathrm{s}} &= \sum_{\alpha=1}^{N} p_{\alpha}^{2}(1-q_{\alpha\alpha}) + \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N} p_{\alpha}p_{\beta}q_{\alpha\beta} \\ &= \Delta q_{\mathrm{d}} + q_{\infty}, \end{aligned} \tag{32}$$

where $q_{\alpha\alpha}$ are dummy variables which are assumed to come from the same statistics as $q_{\alpha\beta}$. Thus, $q_{\alpha\alpha}$ is the expected correlation between two distinct population members both with fitness $f_{\alpha}$. The first term here is arrived at by noting that duplicates have a correlation of unity and replace a pair in the matrix of correlations which would otherwise have expected correlation $q_{\alpha\alpha}$. The second term is the natural increase in correlation as fitness increases (and entropy lowers) and is the sole contribution in the infinite population limit (these definitions differ slightly from those used in (Rattray, 1995)).

### 4.4.1 Natural Increase Term

We estimate the conditional probability distribution for correlation given two fitness values before selection $p(q_{\alpha\beta}|f_{\alpha}, f_{\beta})$ by assuming the alleles within the population are distributed according to the maximum entropy distribution described in the appendix. Then $q_{\infty}$ is simply the correlation averaged over this distribution and the distribution of fitness after selection, $p_{\mathrm{s}}(f)$.

$$q_{\infty} = \int \mathrm{d}q_{\alpha\beta}\, \mathrm{d}f_{\alpha}\, \mathrm{d}f_{\beta}\, p_{\mathrm{s}}(f_{\alpha}) p_{\mathrm{s}}(f_{\beta}) p(q_{\alpha\beta}|f_{\alpha}, f_{\beta})\, q_{\alpha\beta}. \tag{33}$$

This integral can be calculated for large $l$ by the saddle point method[4] and we find that in this limit the result depends only on the mean fitness after selection (Rattray, 1995),

$$q_{\infty}(y) = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{\tau_i + \tanh(y)}{1+\tau_i \tanh(y)}\right)^2, \tag{34a}$$

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

$$K_1^{\mathrm{s}} = \sum_{i=1}^{l} \frac{\tau_i + \tanh(y)}{1+\tau_i \tanh(y)}. \tag{34b}$$

Here, $\tau_i$ is the mean allele at site $i$ before selection and for a distribution at maximum entropy one finds (see equation (59) in the appendix),

$$\tau_i = \tanh(z + x\eta_i).$$

The Lagrange multipliers, $z$ and $x$, are chosen to enforce constraints on the mean overlap and correlation within the population before selection and $\eta_i$ is drawn from a Gaussian distribution with zero mean and unit variance.

[4] For weak selection the large $l$ restriction can be dropped (Rattray, 1996).

It is instructive to expand in $y$, which is appropriate in the weak selection limit. In this case one finds,

$$K_1^{\rm s} = K_1^{\rm c} + yK_2^{\rm c} + \frac{y^2}{2}K_3^{\rm c} + \frac{y^4}{3!}K_4^{\rm c} + \cdots \tag{35a}$$

$$q_\infty(y) = q - \frac{y}{l}K_3^{\rm c} - \frac{y^2}{2l}K_4^{\rm c} + \cdots, \tag{35b}$$

where $K_n^{\rm c}$ are the cumulants after bit-simulated crossover, when the population is assumed to be at maximum entropy (defined in equations (31a) to (31d) up to the fourth cumulant). Recall the expression for the mean fitness after selection given in equation (22a). By comparing this to the above expressions, notice that $y$ plays the role of selection strength in the associated infinite population problem, so for an infinite population one could simply set $y = \beta$.

To calculate $q_\infty$ we solve equation (34b) for $y$ and then substitute this value into equation (34a). In general this must be done numerically, although the weak selection expansion gives a very good approximation in many cases. The third cumulant in equation (35b) will be negative for $K_1 > 0$ because of the negative entropy gradient and this will accelerate the increased correlation under selection.

### 4.4.2 Duplication Term

The duplication term $\Delta q_{\rm d}$ is defined in equation (32). As in the selection calculation presented in section 3.1, population members are independently averaged over a distribution with the correct cumulants to calculate the expectation value of this quantity. In general the expressions must be computed numerically, but the results can be expanded in $1/N$ for sufficiently weak selection (Rattray & Shapiro, 1996). In this case one finds,

$$\Delta q_{\rm d} = \frac{e^{(\beta\sigma)^2}[1 - q_\infty(2\beta)]\rho(2\beta)}{N\rho^2(\beta)} + O\left(\frac{1}{N^2}\right), \tag{36}$$

where $q_\infty(y)$ is defined in equation (34a) and $\rho(\beta)$ is the characteristic function defined in equation (5). Notice that the factor involving the noise here is the same as in the cumulant result presented in equation (21). The effects of noise is therefore removed by the same population size increase as described in equation (23).

It is instructive to expand in $\beta$ as this shows the contribution from each cumulant explicitly. To third order in $\beta$ for three cumulants one finds,

$$\Delta q_{\rm d} \simeq \frac{e^{(\beta\sigma)^2}}{N}\left(1 - q_\infty(2\beta)\right)\left(1 + K_2\beta^2 - K_3\beta^3 + O(\beta^4)\right). \tag{37}$$

Selection leads to a negative third cumulant (see equation (22c)), which in turn leads to an accelerated increase in correlation under further selection. Crossover reduces this effect by reducing the magnitude of the higher cumulants.

## 4.5 SIMULATIONS

The dynamics of the GA can be simulated by iterating the expressions in the preceding sections. In figure 1 the theoretical results are compared to simulation results from a GA averaged over 1000 samples for a typical choice of parameters. The trajectories are shown for the mean and variance of the fitness 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 theoretical results show excellent agreement. The noise was measured in terms of $\kappa_2$ because this provides the most natural units for measuring noise (for example, any breakdown in the theory might be expected to occur for a particular value of $\sigma^2/\kappa_2$). This may seem rather unnatural, although in many cases the noise will fall off as fitness increases. For example, this is the case in the binary perceptron problem which is considered in the next section. In view of this, a fixed noise level might be an equally artificial construction. These considerations are not of critical importance here, however, as the aim is to verify the theory and a more realistic situation is introduced in the next section.

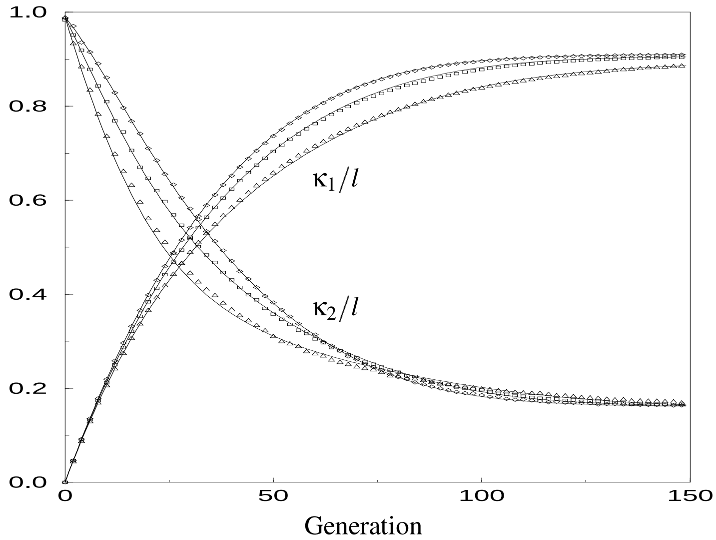


Figure 1: The theory for noisy one-max is compared to results averaged over 1000 runs of a GA. The mean ($\kappa_1$) and variance ($\kappa_2$) are shown, with solid lines showing theoretical predictions. The result for zero noise ($\diamond$) is compared to results with additive Gaussian noise of strength $\sigma^2 = 6\kappa_2$ ($\square$) and $\sigma^2 = 12\kappa_2$ ($\triangle$). The other parameters were $l = 155$, $\beta_s = 0.3$, $p_m = 0.005$, $N = 100$ and bit-simulated crossover was used.

Notice that the strength of the noise is greater than the population's standard deviation 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, probably because the 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. In this case it might be necessary to follow an ensemble of populations, as described in (Prügel-Bennett, 1996).

## 5 GENERALIZATION IN THE BINARY PERCEPTRON

One of the key questions in learning theory is when and how one might generalize to learn a rule from a set of training examples. A simple example of this is the case where a perceptron with binary weights is trained on patterns generated from a teacher perceptron, also with binary weights. The statistical mechanics formalism was applied to this problem in (Rattray & Shapiro, 1996) and here we review these results in order to show how this work may be of relevance to problems from machine learning.

The perceptron has weights $S_i \in \{-1,1\}$ and maps a binary vector with components $\zeta_i^\mu \in \{-1,1\}$ onto a binary output,

$$O^\mu = \mathrm{sgn}\left(\sum_{i=1}^{l} S_i \zeta_i^\mu\right) ; \qquad \mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases} \tag{38}$$

Let $T_i$ be the weights of the teacher perceptron and $S_i$ be the 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,

$$\Lambda_\mathrm{t}^\mu = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} T_i \zeta_i^\mu \qquad \Lambda_\mathrm{s}^\mu = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} S_i \zeta_i^\mu . \tag{39}$$

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

$$E = \sum_{\mu=1}^{\lambda l} \Theta(-\Lambda_\mathrm{t}^\mu \Lambda_\mathrm{s}^\mu) ; \qquad \Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0, \end{cases} \tag{40}$$

where $\lambda l$ is the number of training patterns presented in a batch. To simplify the analysis a new batch of training examples is presented each time the training error is calculated.

The GA processes a population of student weight vectors and the training energy acts as a negative fitness (this is the measured, noisy fitness). Define the ideal fitness $f$ to be the overlap between the weight vectors of the teacher and the student. We choose $T_i = 1$ at every site without loss of generality, in which case the overlap associated with population member $\alpha$ is $f_\alpha$ and is defined,

$$f_\alpha = \frac{1}{l} \sum_{i=1}^{l} S_i^\alpha . \tag{41}$$

This is simply the one-max fitness measure defined in equation (24), normalized to be of order unity (the $n$th cumulant is now typically of $O(l^{1-n})$ rather than $O(l)$). Thus, this problem is equivalent to a noisy version of one-max and the only difference is in the conditional probability distribution relating training error to the overlap between teacher and student. This can be determined and if the size of each batch is $O(l)$ then $p(E|f)$ is well approximated by a Gaussian distribution (Rattray & Shapiro, 1996),

$$p(E|f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(E - E_\mathrm{g}(f))^2}{2\sigma^2}\right) , \tag{42}$$

where the mean and variance are,

$$E_\mathrm{g}(f) = \frac{\lambda l}{\pi} \cos^{-1}(f) \tag{43a}$$

$$\sigma^2(f) = E_\mathrm{g}(f)\left(1 - \frac{E_\mathrm{g}(f)}{\lambda l}\right) . \tag{43b}$$

Here, $E_\mathrm{g}(f)$ 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 $l$ here). The variance expresses the fact that there is noise in the training error (negative fitness) evaluation due to the finite size of the training set.

### 5.1 SELECTION

If the training error is considered to be a negative fitness ($E = -F$) then equation (17) generates the cumulants for the overlap distribution after selection. As before, the integrals have to be computed numerically. Notice that the mean and variance of $p(E|f)$ are non-linear functions of the overlap $f$, so this problem is not exactly equivalent to the noisy one-max problem which was considered in section 4.

For weak selection and large $l$ it is possible to apply the weak selection expansion which was introduced in section 3.3. Since the variance of overlaps within the population is $O(1/l)$ one can expand the mean of $p(E|f)$ around the mean of the population in this limit ($f \simeq K_1$). It is also assumed that the variance of $p(E|f)$ is well approximated by its leading term in this limit. Under these simplifications one finds,

$$E_{\rm g}(f) \simeq \frac{\lambda l}{\pi}\left(\cos^{-1}(K_1) - \frac{(f-K_1)}{\sqrt{1-K_1^2}}\right) \tag{44a}$$

$$\sigma^2 \simeq \frac{\lambda l}{\pi}\cos^{-1}(K_1)\left(1-\frac{1}{\pi}\cos^{-1}(K_1)\right). \tag{44b}$$

Now the mean of $p(E|f)$ is a linear function of $f$ and the problem is very similar to selection corrupted by Gaussian noise. The cumulants after selection are found to be,

$$K_n^{\rm s} = \lim_{\gamma\to 0}\frac{\partial^n}{\partial\gamma^n}\left[\sum_{i=1}^{\infty}\frac{(\gamma+k\beta)^i K_i}{i!} - \frac{{\rm e}^{(\beta\sigma)^2}}{2N}\exp\left(\sum_{i=1}^{\infty}\frac{(2^i-2)(\gamma+k\beta)^i K_i}{i!}\right)\right], \tag{45}$$

where

$$k = \frac{\lambda l}{\pi\sqrt{1-K_1^2}}. \tag{46}$$

This is equivalent to selecting on $f$ directly (see equation (21)) where $k\beta$ is the effective selection strength and $\sigma/k$ is the effective standard deviation of the noise. The correlation result can similarly be calculated by generalizing the noisy one-max result in section 4.4 and one finds that the results are equivalent under the same effective selection strength and noise. A more thorough discussion of these results is given in (Rattray & Shapiro, 1996).

### 5.2 RESIZING THE POPULATION

The noise introduced by the finite sized training set increases the magnitude of the detrimental finite population terms in selection. In the limit of weak selection and large problem size discussed in the preceding section, the effects of noise can be removed by increasing the population size according to equation (23). This maps the trajectory of the finite training set GA onto the trajectory of the GA in the zero noise, infinite training set situation. 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 (23),

$$N = N_0 \exp\left(\frac{\lambda_o}{\lambda}\right), \tag{47}$$

where,

$$\lambda_o = \lambda(\beta\sigma)^2 = \frac{(\lambda\beta)^2 l}{\pi}\cos^{-1}(K_1)\left(1-\frac{1}{\pi}\cos^{-1}(K_1)\right). \tag{48}$$

Here, $\lambda_o$ is independent of $\lambda$ because of the $\beta$ scaling described above. Choosing $N$ according to this expression removes the effects of noise due to the finite batch size and maps the dynamical trajectory onto the infinite training set dynamics (where $E = E_{\rm g}(f)$) for a GA with population size $N_0$. Typically $\beta$ is of order $1/\sqrt{l}$ and this population resizing will not blow up with increases in problem size (for fixed $\lambda$). This is consistent with the result in (Baum *et al*, 1995), although they provide a rigorous proof for the scaling of their algorithm.

Both 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[5]. 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,

$$\lambda = \frac{\lambda_o}{\log(N/N_0)}. \tag{49}$$

Figure 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 $l$ limit is not apparent here.

In this work the effective selection strength was scaled inversely to the standard deviation of overlaps ($\beta = \beta_{\rm s}/k\sqrt{\kappa_2}$). This is a rather artificial choice, as it requires information about overlap statistics which would not be known in practice. However, the population resizing in equation (47) and the corresponding batch sizing expression in equation (49) 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.

## 5.3 OPTIMAL BATCH SIZE

In the previous section it was shown how population size can 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 suggest that there is often an optimal choice of batch size (or measurement accuracy) (Fitzpatrick & Grefenstette, 1988). If the population resizing in equation (47) is used, then it is possible to identify such an 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.

[5]This is a problem for a real GA which produces a finite population after selection. The theoretical model described in section 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.

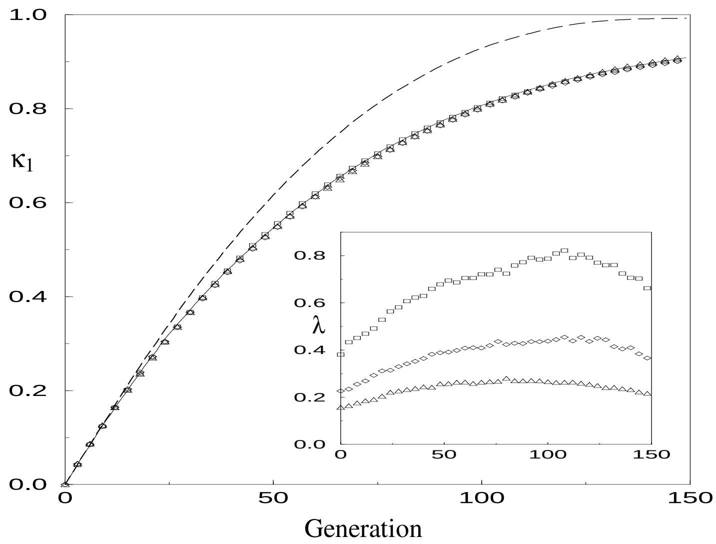


Figure 2: The mean overlap between teacher and student within the population is shown each generation, averaged over 100 runs of a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. Training batch sizes were chosen according to equation (49), leading to trajectories converging onto the infinite training set result where $E = E_{\rm g}(f)$. The solid curve is for the infinite training set result with $N_0 = 60$ and finite training set results are for $N = 90$ ($\Box$), $120(\diamond)$ and $163(\triangle)$. The inset shows the mean choice of batch parameter ($\lambda$) each generation. The dashed line is the infinite training set result for $N = 163$, showing that there is significant potential variability of trajectories under different batch sizing schemes. The other parameters were $l = 279$, $\beta_{\rm s} = 0.25$ and $p_{\rm m} = 0.001$.

It is assumed that computation is mainly due to error evaluation and that other overheads can be neglected. There are $N$ error evaluations each generation with computation time for each scaling as $\lambda$. If the population size each generation is chosen by equation (47), then the computation time $\tau_c$ is related to batch size by,

$$\tau_c(\lambda) \propto \lambda \exp\left(\frac{\lambda_o}{\lambda}\right). \tag{50}$$

The optimal choice of $\lambda$ is given by the minimum of $\tau_c$, which is at $\lambda_o$ (defined in equation (48)). Choosing this batch size leads to the population size being constant over the whole GA run and for optimal efficiency one should choose,

$$N = N_0 \mathrm{e}^1 \simeq 2.73 N_0 \tag{51a}$$
$$\lambda = \lambda_o, \tag{51b}$$

where $N_0$ 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 $N_0$ in order to choose the size of each batch, since $\lambda_o$ is not a function of $N_0$ (see equation (48)). One of the runs in figure 2 is for this choice of $N$ and $\lambda$, showing close agreement to the infinite training set result ($N = 163 \simeq N_0 \mathrm{e}$).

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 the variance of noise decreases as the mean overlap increases (see equation (44b)). Setting $K_1 = 0$ in equation (48) provides this bound,

$$\lambda_o \leq \tfrac{1}{4}(\lambda\beta)^2 l. \tag{52}$$

Recall that $\beta$ is proportional to $1/\lambda$, so that the right hand side of this expression is independent of $\lambda$. 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). The selection strength $\beta$ is typically of order $1/\sqrt{l}$ so that the optimum batch size is typically of order $l$ (recall that the batch size is $\lambda l$).

### 5.4 SIMULATIONS

The dynamics can be modelled by combining the selection results from section 5.1 with the expressions for mutation and crossover derived in section 4. Figures 3 and 4 show the trajectories of the mean and variance of the overlap distribution as well as the maximum overlap, averaged over 1000 runs of a GA for a typical choice of search parameters. The infinite training batch result, where $E = E_{\mathrm{g}}(f)$, is compared to results for two fixed batch sizes, showing how performance degrades as the batch size is reduced. The theoretical curves show excellent agreement to simulation results. The theoretical estimate for the maximum overlap was obtained by assuming population members are randomly sampled from a population with the correct cumulants (Prügel-Bennett & Shapiro, 1995).

There is a slight systematic error in the curves for the smallest batch size and as the batch size is reduced further the theory breaks down. This is probably because a weak selection, low noise approximation was required to calculate the duplication contribution to the correlation after selection, as was also the case for the noisy one-max problem. It is also possible that the Gaussian approximation for $p(E|f)$ breaks down for small $\lambda$, in which case it would be necessary to expand the noise in terms of more cumulants. Results for the higher cumulants also agree with high significance, as shown in (Rattray & Shapiro, 1996).

## 6 CONCLUSION

A theory which describes selection on a finite population under a general stochastic fitness measure has been applied to two related problems, showing excellent predictive power. The problems considered were the one-max problem corrupted by Gaussian noise and a simple learning problem, generalization by a perceptron with binary weights. This work significantly extends the scope of a statistical mechanics formalism for describing the averaged dynamics of the GA and shows how important it is to correctly account for finite population effects.

In the limit of weak Boltzmann selection, the expressions describing the effect of selection on each fitness cumulant can be expressed analytically and we find that an increased population size removes the effects of noise in this limit. This may have important implications in learning theory, where there is often noise in fitness evaluation due to incomplete training data. Indeed, it is shown how this population sizing can be used to determine the optimal batch size in the binary perceptron problem, which minimizes computation time, as well as the total number of training examples required when independent batches are used.

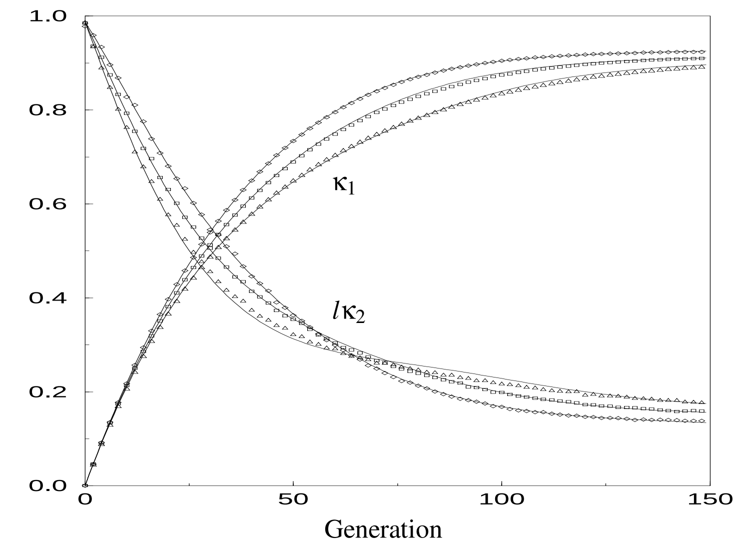


Figure 3: The theory is compared to averaged results from a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. The mean and variance of the overlap distribution are shown averaged over 1000 runs, with solid lines showing theoretical predictions. The infinite training set result ($\diamond$) is compared to results for a finite training set with $\lambda = 0.65$ ($\square$) and $\lambda = 0.39$ ($\triangle$). The other parameters were $l = 155$, $\beta_s = 0.3$, $p_m = 0.005$, $N = 80$ and bit-simulated crossover was used. Adapted from (Rattray & Shapiro, 1996).

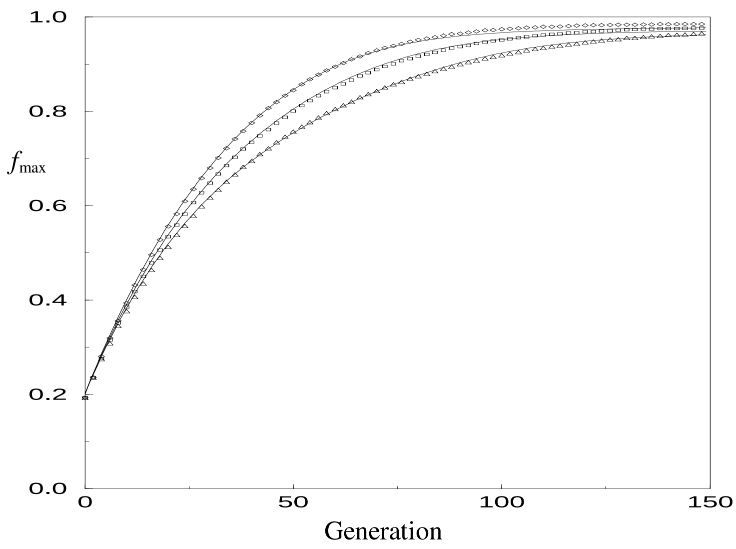


Figure 4: The maximum overlap between teacher and pupil is shown each generation, averaged over the same simulations as the results presented in figure 3. The solid lines show the theoretical predictions and symbols are as in figure 3.

**Acknowledgements**

We would like to thank Adam Prügel-Bennett for many helpful discussions and for providing some of the numerical code used here. We would also like to thank the anonymous reviewers for making a number of useful suggestions. MR was supported by an EPSRC award (ref. 93315524).

## MAXIMUM ENTROPY DISTRIBUTION

After bit-simulated crossover the population is assumed to be at maximum entropy with constraints on the mean fitness and correlation within the population. This is a special case of the result derived in (Prügel-Bennett & Shapiro, 1995) for the paramagnet and this discussion follows theirs closely.

Let $\tau_i$ be the mean allele at site $i$ within the population,

$$\tau_i = \langle S_i^\alpha \rangle_\alpha = \frac{1}{N} \sum_{\alpha=1}^{N} S_i^\alpha. \tag{53}$$

To calculate the distribution of this quantity over sites one imposes constraints on the mean overlap and correlation with Lagrange multipliers $x$ and $z$,

$$zNK_1 = z \sum_{\alpha=1}^{N} \sum_{i=1}^{l} S_i^\alpha = zN \sum_{i=1}^{l} \tau_i \tag{54a}$$

$$\tfrac{1}{2}(xN)^2 q = \frac{x^2}{2l} \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \sum_{i=1}^{l} S_i^\alpha S_i^\beta = \frac{(xN)^2}{2l} \sum_{i=1}^{l} \tau_i^2. \tag{54b}$$

The correlation expression is for large $N$ and finite population corrections can be included retrospectively.

Without constraints, the fraction of allele configurations which are compatible with mean allele $\tau_i$ is given by a binomial coefficient,

$$\Omega(\tau_i) = \frac{1}{2^N} \binom{N}{N(1+\tau_i)/2}. \tag{55}$$

One can then define an entropy,

$$\mathcal{S}(\tau_i) = \log[\Omega(\tau_i)] \sim -\frac{N}{2} \log(1-\tau_i^2) + \frac{N\tau_i}{2} \log\left(\frac{1-\tau_i}{1+\tau_i}\right), \tag{56}$$

where Stirling's approximation has been used. The probability distribution for allele configurations decouples at each site,

$$p(\{\tau_i\}) = \prod_{i=1}^{l} p(\tau_i) = \prod_{i=1}^{l} \exp[\mathcal{S}(\tau_i) + zN\tau_i + (xN\tau_i)^2/2]. \tag{57}$$

A Gaussian integral removes the square in the exponent,

$$p(\tau_i) = \int \frac{d\eta_i}{\sqrt{2\pi}} \exp\left(\frac{-\eta_i^2}{2} + NG(\tau_i, \eta_i)\right) ; \quad G(\tau_i, \eta_i) = S(\tau_i)/N + z\tau_i + x\eta_i\tau_i. \tag{58}$$

The maximal value of $G$ with respect to $\tau_i$ gives the maximum entropy distribution for $\tau_i$ at each site,

$$\tau_i = \tanh(z + x\eta_i), \tag{59}$$

where $\eta_i$ is drawn from a Gaussian with zero mean and unit variance. The constraints can be used to obtain values for the Lagrange multipliers,

$$K_1 = \sum_{i=1}^{l} \overline{\tanh(z + x\eta_i)} \qquad q = \frac{1}{l}\sum_{i=1}^{l} \overline{\tanh^2(z + x\eta_i)}. \tag{60}$$

Bars denote averages over the Gaussian noise which in general must be done numerically (Gauss-Hermite quadrature was used here (Press *et al*, 1992)).

The third and fourth order terms in equations (31c) and (31d) can be found once the Lagrange multipliers have been determined,

$$\sum_{i=1}^{l} \langle S_i^{\alpha} \rangle_{\alpha}^3 = l\,\overline{\tanh^3(z + x\eta)} \qquad \sum_{i=1}^{l} \langle S_i^{\alpha} \rangle_{\alpha}^4 = l\,\overline{\tanh^4(z + x\eta)}. \tag{61}$$

Again, bars denote averages over the Gaussian noise.

## References


J. E. Baker (1987) “Reducing Bias and Inefficiency in the Selection Algorithm,” Proc. of the 2nd Int. Conf. on Genetic Algorithms, ed J. J. Grefenstette (Hillsdale, NJ; Lawrence Erlbaum) p 14-21.

E. B. Baum, D. Boneh and C. Garret (1995) “On Genetic Algorithms,” in COLT ’95: Proc. of the 8th Annual Conf. on Computational Learning Theory (New York; Assoc. for Computing Machinery Inc.) p 230–239.

T. Blickle, L. Thiele (1995) “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.

M. De la Maza, B. Tidor (1991) “Increased Flexibility in Genetic Algorithms: The Use of Variable Boltzmann Selective Pressure to Control Propagation,” Proc. of the ORSA CSTS Conference - Computer Science and Operations Research: New Developments in their Interfaces, p 425–440.

B. Derrida (1981) “Random-energy Model: An Exactly Solvable Model of Disordered Systems,” Phys. Rev. **B 24**, 2613–2625.

W. J. Ewens (1979) “Mathematical Population Genetics,” (Berlin; Springer-Verlag).

J. M. Fitzpatrick, J. J. Grefenstette (1988) “Genetic Algorithms in Noisy Environments,” Machine Learning **3**, 101–120.

D. E. Goldberg (1989) “Genetic Algorithms in Search, Optimization and Machine Learning,” (Reading, MA; Addison-Wesley).

D. E. Goldberg, K. Deb, J. H. Clark (1992) “Genetic Algorithms, Noise, and the Sizing of Populations,” Complex Systems **6**, 333-362.

J. H. Holland (1975) "Adaptation in Natural and Artificial Systems," (Ann Arbor; The University of Michigan Press).

B. L. Miller, D. E. Goldberg (1995) "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).

H. Mühlenbein, D Schlierkamp-Voosen (1995) "Analysis of Selection, Mutation and Recombination in Genetic Algorithms," Lecture Notes in Computer Science **899**, 188-214.

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling (1992) "Numerical Recipes in C : The Art of Scientific Computing," 2nd ed. (Cambridge; Cambridge University Press).

A. Prügel-Bennett, J. L. Shapiro (1994) "An Analysis of Genetic Algorithms using Statistical Mechanics," Phys. Rev. Lett. **72**(9), 1305.

A. Prügel-Bennett, J. L. Shapiro (1995) "The Dynamics of a Genetic Algorithm for Simple Random Ising Systems," Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, U.K. (to appear in Physica D).

A. Prügel-Bennett (1996) "Modelling Evolving Populations," NORDITA, Blegdamsvej 17, DK-2100 Copenhagen, Denmark, (submitted to J. Theor. Biol.).

L. M. Rattray (1995) "The Dynamics of a Genetic Algorithm under Stabilizing Selection," Complex Systems **9**(3), 213–234.

L. M. Rattray (1996) "Modelling the Dynamics of Genetic Algorithms using Statistical Mechanics," Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, UK (PhD. Thesis - In Preperation).

L. M. Rattray , J. L. Shapiro (1996) "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**).

J. L. Shapiro, A. Prügel-Bennett, L. M. Rattray (1994) "A Statistical Mechanical Formulation of the Dynamics of Genetic Algorithms," Lecture Notes in Computer Science **865**, 17–27.

M. Srinivas, L. M. Patnaik (1995) "Binomially Distributed Populations for Modelling GAs," Proc. of the 5th Int. Conf. on Ganetic Algorithms, ed S. Forrest (San Mateo, CA; Morgan Kaufmann) p 138–145.

A. Stuart, J. K. Ord (1987) "Kendall's Advanced Theory of Statistics, Vol 1. Distribution Theory," 5th ed. (New York; Oxford University Press).

G. Syswerda (1993) "Simulated Crossover in Genetic Algorithms," in Foundations of Genetic Algorithms 2, (San Mateo, CA; Morgan Kaufmann).

D. Thierens, D. Goldberg (1995) "Convergence Models of Genetic Algorithm Selection Schemes," Parallel Problem Solving from Nature III (in Lecture Notes in Computer Science **866**), 119–129.

M. D. Vose, A. H. Wright (1994) "Simple Genetic Algorithms with Linear Fitness," Evol. Comp. **2**, 347–368.