The dynamics of a Genetic Algorithm for a simple learning problem

Magnus Rattray § and Jonathan L Shapiro

Computer Science Department, University of Manchester, Oxford Road, Manchester M13 9PL, UK

To appear in *J. Phys* **A** (accepted August 15th, 1996)

Abstract.

A formalism for describing the dynamics of Genetic Algorithms (GAs) using methods from statistical mechanics is applied to the problem of generalization in a perceptron with binary weights. The dynamics are solved for the case where a new batch of training patterns is presented to each population member each generation, which considerably simplifies the calculation. The theory is shown to agree closely to simulations of a real GA averaged over many runs, accurately predicting the mean best solution found. For weak selection and large problem size the difference equations describing the dynamics can be expressed analytically and we find that the effects of noise due to the finite size of each training batch can be removed by increasing the population size appropriately. If this population resizing is used, one can deduce the most computationally efficient size of training batch each generation. For independent patterns this choice also gives the minimum total number of training patterns used. Although using independent patterns is a very inefficient use of training patterns in general, this work may also prove useful for determining the optimum batch size in the case where patterns are recycled.

1. Introduction

Genetic Algorithms (GAs) are adaptive search techniques, which can be used to find low energy states in poorly characterized, high-dimensional energy landscapes [8, 11]. They have already been successfully applied in a large range of domains [2] and a review of the literature shows that they are becoming increasingly popular. In particular, GAs have been used in a number of machine learning applications, including the design and training of artificial neural networks [7, 19, 28].

In the simple GA considered here, each population member is represented by a genotype, in this case a binary string, and an objective function assigns an energy

§ email: rattraym@cs.man.ac.uk

to each such genotype. A population of solutions evolves for a number of discrete generations under the action of genetic operators, in order to find low energy (high fitness) states. The most important operators are selection, where the population is improved through some form of preferential sampling, and crossover (or recombination), where population members are mixed, leading to non-local moves in the search space. Mutation is usually also included, allowing incremental changes to population members. GAs differ from other stochastic optimization techniques, such as simulated annealing, because a population of solutions is processed in parallel and it is hoped that this may lead to improvement through the recombination of mutually useful features from different population members.

A formalism has been developed by Prügel-Bennett, Shapiro and Rattray which describes the dynamics of a simple GA using methods from statistical mechanics [14, 15, 16, 17]. This formalism has been successfully applied to a number of simple Ising systems and has been used to determine optimal settings for some of the GA search parameters [21]. It describes problems of realistic size and includes finite population effects, which have been shown to be crucial to understanding how the GA searches. The approach can be applied to a range of problems including ones with multiple optima, and it has been shown to predict simulation results with high accuracy, although small errors can sometimes be detected.

Under the statistical mechanics formalism, the population is described by a small number of macroscopic quantities which are statistical measures of the population. Statistical mechanics techniques are used to derive deterministic difference equations which describe the average effect of each operator on these macroscopics. Since the dynamics of a GA is to be modelled by the average dynamics of an ensemble of GAs, it is important that the quantities which are used to describe the system are robust and self-averaging. The macroscopics which have been used are the cumulants of some appropriate quantity, such as the energy or the magnetization, and the mean correlation within the population, since these are robust statistics which average well over different realizations of the dynamics. There may be small systematic errors, since the difference equations for evolving these macroscopics sometimes involve nonlinear terms which may not self-average, but these corrections are generally small and will be neglected here.

The statistical mechanics theory is distinguished by the facts that a macroscopic description of the GA is used and that the averaging is done such that fluctuations can be included in a systematic way. Many other theoretical approaches are based on the intuitive idea that above average fitness building blocks are preferentially sampled by the GA, which, if they can be usefully recombined, results in highly fit individuals being produced [8, 11]. Although this may be a useful guide to the suitability of particular problems to a GA, it is difficult to make progress towards a quantitative description for realistic problems, as it is difficult to determine which are the relevant building blocks and which building blocks are actually present in a finite population. This approach has led to false predictions of problem difficulty, especially when the dynamic nature of the search is ignored [6, 9]. A rigorous approach introduced by Vose *et al* describes the population dynamics as a dynamical system in a high-dimensional Euclidean space, with each genetic operator incorporated as a transition tensor [25, 26]. This method uses a microscopic description and is difficult to apply to specific problems of realistic size due to high-dimensionality of the equations of motion. More recently, a number of results have been derived for the performance of a GA on a class of simple additive problems [1, 12, 20]. These approaches use a macroscopic description, but assume a particular form for the distribution of macroscopics which is only applicable in large populations and for a specific class of problem. It is difficult to see how to transfer the results to other problems where finite population effects cannot be ignored.

Other researchers have introduced theories based on averages. A description of GA dynamics in terms of the evolution of the parent distribution from which finite populations are sampled was produced by Vose and Wright [27]. This microscopic approach provides a description of the finite population effects which is elegant and correct. However, like other microscopic descriptions it is difficult to apply to specific realistic problems due to the enormous dimensionality of the system. Macroscopic descriptions can result in low-dimensional equations which can be more easily studied. Another formalism based on the evolution of parent distributions was developed by Peck and Dhawan [13], but they did not use the formalism to develop equations describing finite population dynamics.

The importance of choosing appropriate quantities to average is well-known in statistical physics, but does not seem to be widely appreciated in genetic algorithm theory. In particular, many authors use results based on properties of the *average* probability distribution; this is insensitive to finite-population fluctuations and only gives accurate results in the infinite population limit. Thus, many results are only accurate in the infinite population limit, even though this limit is not taken explicitly. For example, Srinivas and Patnaik [23] and Peck and Dhawan [13] both produce equations for the moments of the fitness distribution in terms of the moments of the initial distribution. These are moments of the average distribution. Consequently, the equations do not correctly describe a finite population and results presented in these papers reflect that. Other attempts to describe GAs in terms of population moments (or schema moments or average Walsh coefficients) suffer from this problem. Macroscopic descriptions of population dynamics are also widely used in quantitative genetics (see, for example, reference [5]). In this field the importance of finite-population fluctuations is more widely appreciated; the infinite population limit is usually taken explicitly. Using the statistical mechanics approach, equations for fitness moments which include finite-population fluctuations can be derived by averaging the cumulants, which are more robust statistics.

Here, the statistical mechanics formalism is applied to a simple problem from learning theory, generalization of a rule by a perceptron with binary weights. The perceptron learns from a set of training patterns produced by a teacher perceptron, also with binary weights. A new batch of training patterns are presented to each population member each generation which simplifies the analysis considerably, since there are no over-training effects and each training pattern can be considered as statistically independent. Baum *et al* have shown that this problem is similar to a paramagnet whose energy is corrupted by noise and they suggest that the GA may perform well in this case, since it is relatively robust towards noise when compared to local search methods [1]. The noise in the training energy is due to the finite size of the training set and is a feature of many machine learning problems [7].

We show that the noise in the training energy is well approximated by a Gaussian distribution for large problem size, whose mean and variance can be exactly determined and are simple functions of the overlap between pupil and teacher. This allows the dynamics to be solved, extending the statistical mechanics formalism to this simple, yet non-trivial, problem from learning theory. The theory is compared to simulations of a real GA averaged over many runs and is shown to agree well, accurately predicting the evolution of the cumulants of the overlap distribution within the population, as well as the mean correlation and mean best population member. In the limit of weak selection and large problem size the population size can be increased to remove finite training set effects and this leads to an expression for the optimal training batch size.

2. Generalization in a perceptron with binary weights

A perceptron with Ising weights $w_i \in \{-1, 1\}$ maps an Ising training pattern $\{\zeta_i^\mu\}$ onto a binary output,

$$O^\mu = \mathrm{Sgn}\left(\sum_{i=1}^{N} w_i \zeta_i^\mu\right) \qquad \mathrm{Sgn}(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ -1 & \text{for } x < 0 \end{cases} \tag{1}$$

where N is the number of weights. Let t_i be the weights of the teacher perceptron and w_i be the weights of the pupil. The stability of a pattern is a measure of how well it is stored by the perceptron and the stabilities of pattern μ for the teacher and pupil are Λ_t^μ and Λ_w^μ respectively,

$$\Lambda_t^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} t_i \zeta_i^\mu \qquad \Lambda_w^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \zeta_i^\mu \tag{2}$$

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

$$E = \sum_{\mu=1}^{\lambda N} \Theta(-\Lambda_t^\mu \Lambda_w^\mu) \qquad \Theta(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{3}$$

where λN is the number of training patterns presented and $\Theta(x)$ is the Heaviside function. In this work a new batch of training examples is presented each time the training energy is calculated.

For large N it is possible to calculate the entropy of solutions compatible with the total training set and there is a first-order transition to perfect generalization as the size of training set is increased [10, 22]. This transition occurs for $O(N)$ patterns and beyond the transition the weights of the teacher are the only weights compatible with the training set. In this case there is no problem with over-training to that particular set, although a search algorithm might still fail to find these weights. The GA considered here will typically require more than $O(N)$ patterns, since it requires an independent batch for each energy evaluation, so avoiding any possibility of over-training.

Define R to be the overlap between pupil and teacher,

$$R = \frac{1}{N}\sum_{i=1}^{N} w_i t_i \tag{4}$$

We choose $t_i = 1$ at every site without loss of generality. If a statistically independent pattern is presented to a perceptron, then for large N the stabilities of the teacher and pupil are Gaussian variables each with zero mean and unit variance, and with covariance R,

$$p(\Lambda_t, \Lambda_w) = \frac{1}{2\pi\sqrt{1-R^2}} \exp\left(\frac{-(\Lambda_t^2 - 2R\Lambda_t\Lambda_w + \Lambda_w^2)}{2(1-R^2)}\right) \tag{5}$$

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

$$p(E|R) = \left\langle \delta\left(E - \sum_{\mu=1}^{\lambda N} \Theta(-\Lambda_t^\mu \Lambda_w^\mu)\right)\right\rangle_{\{\Lambda_t^\mu, \Lambda_w^\mu\}} \tag{6}$$

where the brackets denote an average over stabilities distributed according to the joint distribution in equation (5). The logarithm of the Fourier transform generates the cumulants of the distribution,

$$\begin{aligned}
\hat{\rho}(-\mathrm{i}t|R) &= \int_{-\infty}^{\infty} \mathrm{d}E\, p(E|R)\, \mathrm{e}^{tE} \\
&= \left\langle \prod_{\mu=1}^{\lambda N} \exp\left[t\Theta(-\Lambda_t^\mu \Lambda_w^\mu)\right]\right\rangle \\
&= \left(1 + \frac{1}{\pi}(\mathrm{e}^t - 1)\cos^{-1}(R)\right)^{\lambda N} \qquad (7)
\end{aligned}$$

The logarithm of this quantity can be expanded in t, with the cumulants of the distribution given by the coefficients of the expansion. The higher cumulants are $O(\lambda N)$ and it turns out that the shape of the distribution is not critical as long as λ is $O(1)$.

A Gaussian distribution will be a good approximation in this case,

$$p(E|R) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(E - E_{\rm g}(R))^2}{2\sigma^2} \right) \tag{8}$$

where the mean and variance are,

$$E_{\rm g}(R) = \frac{\lambda N}{\pi} \cos^{-1}(R) \tag{9}$$

$$\sigma^2 = \frac{\lambda N}{\pi} \cos^{-1}(R) \left(1 - \frac{1}{\pi} \cos^{-1}(R) \right) \tag{10}$$

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 for convenience). The variance expresses the fact that there is noise in the energy evaluation due to the finite size of the training batch.

3. Modelling the Genetic Algorithm

3.1. The Genetic Algorithm

Initially, a random population of solutions is created, in this case Ising weights of the form $\{w_1, w_2 \ldots, w_N\}$ where the alleles w_i are the weights of a perceptron. The size of the population is P and will usually remain fixed, although a dynamical resizing of the population is discussed in section 7. Under selection, new population members are chosen from the present population with replacement, with a probability proportional to their Boltzmann weight. The selection strength β is analogous to the inverse temperature and determines the intensity of selection, with larger β leading to a higher variance of selection probabilities [3, 15]. Under standard uniform crossover, the population is divided into pairs at random and the new population is produced by swapping weights at each site within a pair with some fixed probability. Here, bit-simulated crossover is used, with new population members created by selecting weights at each site from any population member in the original population with equal probability [24]. In practice, the weights at every site are completely shuffled within the population and this brings the population straight to the fixed point of standard crossover. This special form of crossover is only practicable here because crossover does not change the mean overlap between pupil and teacher within the population. Standard mutation is used, with random bits flipped throughout the population with probability $p_{\rm m}$.

Each population member receives an independent batch of λN examples from the teacher perceptron each generation, so that the relationship between the energy and the overlap between pupil and teacher is described by the conditional probability defined in equation (6). In total, $\lambda N \times PG$ training patterns are used, where G is the total number of generations and P is the population size (or the mean population size).

The population will be described in terms of a number of macroscopic variables, the cumulants of the overlap distribution within the population and the mean correlation within the population. In the following sections, difference equations will be derived for the average change of a small set of these macroscopics, due to each operator. A more exact approach considers fluctuations from mean behaviour by modelling the evolution of an ensemble of populations described by a set of order parameters [14]. Here, it is assumed that the dynamics average sufficiently well so that we can describe the dynamics in terms of deterministic equations for the average behaviour of each macroscopic. This assumption is justified by the excellent agreement between the theory and simulations of a real GA, some of which are presented in section 8. Once difference equations are derived for each macroscopic, they can be iterated in sequence in order to simulate the full dynamics.

Notice that although we follow information about the overlap between teacher and pupil, this is of course not known in general. The only feedback available when training the GA is the training energy defined in equation 3. Selection acts on this energy, and it is therefore necessary to average over the noise in selection which is due both to the stochastic nature of the training energy evaluation and of the selection procedure itself.

Finite population effects prove to be of fundamental importance when modelling the GA. A striking example of this is in selection, where an infinite population assumption leads to the conclusion that the selection strength can be set arbitrarily high in order to move the population to the desired solution. This is clearly nonsense, as selection could never move the population beyond the best existing population member. Two improvements are required to model selection accurately; the population should be finite and the distribution from which it is drawn should be modelled in terms of more than two cumulants, going beyond a Gaussian approximation [15]. The higher cumulants play a particularly important role in selection which will be described in section 5.1 [16].

The higher cumulants of the population after bit-simulated crossover are determined by assuming the population is at maximum entropy with constraints on the mean overlap and correlation within the population (see Appendix A). The effect of mutation on the mean overlap and correlation only requires the knowledge of these two macroscopics, so these are the only quantities we need to evolve in order to model the full dynamics. All other relevant properties of the population after crossover can be found from the maximum entropy ansatz. A more general method is to follow the evolution of a number of cumulants explicitly, as in references [16, 17], but this is unnecessary here because of the special form of crossover used, which is not appropriate in problems with stronger spatial interactions.

3.3. *The cumulants and correlation*

The cumulants of the overlap distribution within the population are robust statistics which are often reasonably stable to fluctuations between runs of the GA, so that they average well [16]. The first two cumulants are the mean and variance respectively, while the higher cumulants describe the deviation from a Gaussian distribution. The third and fourth cumulants are related to the skewness and kurtosis of the population respectively. A population member, labelled α, is associated with overlap R_α defined in equation (4). The cumulants of the overlap distribution within a finite population can be generated from the logarithm of a partition function,

$$Z = \sum_{\alpha=1}^{P} \exp(\gamma R_\alpha) \tag{11}$$

where P is the population size. If κ_n is the nth cumulant, then,

$$\kappa_n = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \log Z \tag{12}$$

The partition function holds all the information required to determine the cumulants of the distribution of overlaps within the population.

The correlation within the population is a measure of the microscopic similarity of population members and is important because selection correlates a finite population, sometimes leading to premature convergence to poor solutions. It is also important in calculating the effect of crossover, since this involves the interaction of different population members and a higher correlation leads to less disruption on average. The correlation between two population members, α and β, is $q_{\alpha\beta}$ and is defined by,

$$q_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} w_i^\alpha w_i^\beta \tag{13}$$

The mean correlation is q and is defined by,

$$q = \frac{2}{P(P-1)} \sum_{\alpha=1}^{P} \sum_{\beta>\alpha} q_{\alpha\beta} \tag{14}$$

In order to model a finite population we consider that P population members are randomly sampled from an infinite population, which is described by a set of infinite population cumulants, K_n [14]. The expectation values for the mean correlation and the first cumulant of a finite population are equal to the infinite population values. The higher cumulants are reduced by a factor which depends on the population size,

$$\kappa_1 = K_1 \tag{15a}$$

$$\kappa_2 = P_2 K_2 \tag{15b}$$

$$\kappa_3 = P_3 K_3 \tag{15c}$$

$$\kappa_4 = P_4 K_4 - 6 P_2 (K_2)^2 / P \tag{15d}$$

Here, P_2, P_3 and P_4 give finite population corrections to the infinite population result (see reference [16] for a derivation),

$$P_2 = 1 - \frac{1}{P} \qquad P_3 = 1 - \frac{3}{P} + \frac{2}{P^2} \qquad P_4 = 1 - \frac{7}{P} + \frac{12}{P^2} - \frac{6}{P^3} \tag{16}$$

Although we model the evolution of a finite population, it is more natural to follow the macroscopics associated with the infinite population from which the finite population is sampled [14]. The expected cumulants of a finite population can be retrieved through equations (15a) to (15d).

4. Crossover and mutation

The mean effects of standard crossover and mutation on the distribution of overlaps within the population are equivalent to the paramagnet results given in [16]. However, bit-simulated crossover brings the population straight to the fixed point of standard crossover, which will be assumed to be a maximum entropy distribution with the correct mean overlap and correlation, as described in Appendix A. To model this form of crossover one only requires knowledge of these two macroscopics, so these are the only two quantities we need to evolve under selection and mutation.

The mean overlap and correlation after averaging over all mutations are,

$$K_1^{\mathrm{m}} = (1 - 2p_{\mathrm{m}})K_1 \tag{17a}$$

$$q_{\mathrm{m}} = (1 - 2p_{\mathrm{m}})^2 q \tag{17b}$$

where p_{m} is the probability of flipping a bit under mutation [16]. The higher cumulants after crossover are required to determine the effects of selection, discussed in the next section. The mean overlap and correlation are unchanged by crossover and the other cumulants can be determined by noting that bit-simulated crossover completely removes the difference between site averages within and between different population members. For example, terms like $\langle w_i^\alpha w_j^\beta \rangle_{i \neq j}$ and $\langle w_i^\alpha w_j^\alpha \rangle_{i \neq j}$ are equal on average. After cancelling terms of this form one finds that the first four cumulants of an infinite population after crossover are,

$$K_1^{\mathrm{c}} = K_1 \tag{18a}$$

$$K_2^{\mathrm{c}} = \frac{1}{N}(1 - q) \tag{18b}$$

$$K_3^{\mathrm{c}} = -\frac{2}{N^2}\left(K_1 - \frac{1}{N}\sum_{i=1}^{N}\langle w_i^\alpha \rangle_\alpha^3\right) \tag{18c}$$

$$K_4^{\mathrm{c}} = -\frac{2}{N^3}\left(1 - 4q + \frac{3}{N}\sum_{i=1}^{N}\langle w_i^\alpha \rangle_\alpha^4\right) \tag{18d}$$

Here, the brackets denote population averages. The third and fourth order terms in the expressions for the third and fourth cumulants are calculated in Appendix A by making a maximum entropy ansatz. The expected cumulants of a finite population after crossover are determined from equations ($15a$) to ($15d$).

5. The cumulants after selection

Under selection, P new population members are chosen from the present population with replacement. Following Prügel-Bennett we split this operation into two stages [14]. First we randomly sample P population members from an infinite population in order to create a finite population. Then an infinite population is generated from this finite population by selection. The proportion of each population member represented in the infinite population after selection is equal to its probability of being selected, which is defined below. The sampling procedure can be averaged out in order to calculate the expectation values for the cumulants of the overlap distribution within an infinite population after selection, in terms of the infinite population cumulants before selection.

The probability of selecting population member α is p_α and for Boltzmann selection one chooses,

$$p_\alpha = \frac{\mathrm{e}^{-\beta E_\alpha}}{\sum^P \mathrm{e}^{-\beta E_\alpha}} \tag{19}$$

where β is the selection strength and the denominator ensures that the probability is correctly normalized. Here, E_α is the training energy of population member α.

One can then define a partition function for selection,

$$Z_\mathrm{s} = \sum_{\alpha=1}^{P} \exp(-\beta E_\alpha + \gamma R_\alpha) \tag{20}$$

The logarithm of this quantity generates the cumulants of the overlap distribution for an infinite population after selection,

$$K_n^\mathrm{s} = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \log Z_\mathrm{s} \tag{21}$$

One can average this quantity over the population by assuming each population member is independently selected from an infinite population with the correct cumulants,

$$\langle \log Z_\mathrm{s} \rangle = \left(\prod_{\alpha=1}^{P} \int \mathrm{d}R_\alpha \, \mathrm{d}E_\alpha \, p(R_\alpha) \, p(E_\alpha | R_\alpha) \right) \log Z_\mathrm{s} \tag{22}$$

where $p(E|R)$ determines the stochastic relationship between energy and overlap as defined in equation (6) which will be approximated by the Gaussian distribution in equation (8). Following Prügel-Bennett and Shapiro one can use Derrida's trick and express the logarithm as an integral in order to decouple the average [4, 15].

$$\begin{aligned}\langle \log Z_\mathrm{s} \rangle &= \int_0^\infty \mathrm{d}t \, \frac{\mathrm{e}^{-t} - \langle \mathrm{e}^{-tZ_\mathrm{s}} \rangle}{t} \\ &= \int_0^\infty \mathrm{d}t \, \frac{\mathrm{e}^{-t} - f^P(t, \beta, \gamma)}{t}\end{aligned} \tag{23}$$

where,

$$f(t,\beta,\gamma) = \int \mathrm{d}R\,\mathrm{d}E\,p(R)\,p(E|R)\exp\left(-t\mathrm{e}^{-\beta E+\gamma R}\right) \tag{24}$$

The distribution of overlaps within an infinite population is approximated by a cumulant expansion around a Gaussian distribution [16],

$$p(R) = \frac{1}{\sqrt{2\pi K_2}}\exp\left(\frac{-(R-K_1)^2}{2K_2}\right)\left[1+\sum_{n=3}^{n_c}\frac{K_n}{K_2^{n/2}}\,u_n\!\left(\frac{R-K_1}{\sqrt{K_2}}\right)\right] \tag{25}$$

where $u_n(x) = (-1)^n \mathrm{e}^{\frac{x^2}{2}}\frac{\mathrm{d^n}}{\mathrm{dx^n}}\mathrm{e}^{\frac{-x^2}{2}}/n!$ are scaled Hermite polynomials. Four cumulants were used for the simulations presented in section 8 and the third and fourth Hermite polynomials are $u_3(x) = (x^3-3x)/3!$ and $u_4(x) = (x^4-6x^2+3)/4!$. This function is not a well defined probability distribution since it is not necessarily positive, but it has the correct cumulants and provides a good approximation. In general, the integrals in equations (23) and (24) have to be computed numerically, as was the case for the simulations presented in section 8.

5.1. Weak selection and large N

It is instructive to expand in small β and large N, as this shows the contributions for each cumulant explicitly and gives some insight into how the size of the training set affects the dynamics. Since the variance of the population is $O(1/N)$ it is reasonable to expand the mean of $p(E|R)$, defined in equation (9), 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 gradient of the noise becomes important. Under these simplifying assumptions one finds,

$$E_{\mathrm{g}}(R) \;\simeq\; \frac{\lambda N}{\pi}\left(\cos^{-1}(K_1) - \frac{(R-K_1)}{\sqrt{1-K_1^2}}\right) \tag{26}$$

$$\sigma^2 \;\simeq\; \frac{\lambda N}{\pi}\cos^{-1}(K_1)\left(1-\frac{1}{\pi}\cos^{-1}(K_1)\right) \tag{27}$$

Following Prügel-Bennett and Shapiro [15], one can expand the integrand in equation (23) for small β (as long as λ is at least $O(1)$ so that the variance of $p(E|R)$ is $O(N)$),

$$f^P(t,\beta,\gamma) \simeq \exp(-tP\hat{\rho}_1(\beta,\gamma))\left(1+\frac{Pt^2}{2}\left(\hat{\rho}_2(\beta,\gamma)-\hat{\rho}_1^2(\beta,\gamma)\right)\right) \tag{28}$$

where,

$$\hat{\rho}_n(\beta,\gamma) = \int \mathrm{d}R\,\mathrm{d}E\,p(R)\,p(E|R)\,\mathrm{e}^{n(-\beta E+\gamma R)} \tag{29}$$

We approximate $p(E|R)$ by a Gaussian whose mean and variance are given in equations (26) and (27). Completing the integral in equation (23), one finds an expression for the cumulants of an infinite population after selection,

$$K_n^{\mathrm{s}} = \lim_{\gamma \to 0} \frac{\partial^n}{\partial \gamma^n} \left[\log(P\rho_1(k\beta,\gamma)) - \frac{\mathrm{e}^{(\beta\sigma)^2}}{2P} \left(\frac{\rho_2(k\beta,\gamma)}{\rho_1^2(k\beta,\gamma)} \right) \right] \tag{30}$$

where,

$$\begin{aligned} \rho_n(k\beta,\gamma) &= \int \mathrm{d}R \, p(R) \mathrm{e}^{nR(k\beta+\gamma)} \\ &= \exp\left(\sum_{i=1}^{\infty} \frac{n^i (k\beta+\gamma)^i K_i}{i!} \right) \end{aligned} \tag{31}$$

Here, a cumulant expansion has been used. The parameter k is the constant of proportionality relating the generalization error to the overlap in equation (26) (constant terms are irrelevant, as Boltzmann selection is invariant under the addition of a constant to the energy).

$$k = \frac{\lambda N}{\pi \sqrt{1 - K_1^2}} \tag{32}$$

For the first few cumulants of an infinite population after selection one finds,

$$K_1^{\mathrm{s}} = K_1 + \left(1 - \frac{\mathrm{e}^{(\beta\sigma)^2}}{P}\right) k\beta K_2 + O(\beta^2) \tag{33a}$$

$$K_2^{\mathrm{s}} = \left(1 - \frac{\mathrm{e}^{(\beta\sigma)^2}}{P}\right) K_2 + \left(1 - \frac{3\mathrm{e}^{(\beta\sigma)^2}}{P}\right) k\beta K_3 + O(\beta^2) \tag{33b}$$

$$K_3^{\mathrm{s}} = \left(1 - \frac{3\mathrm{e}^{(\beta\sigma)^2}}{P}\right) K_3 - \frac{6\mathrm{e}^{(\beta\sigma)^2}}{P} k\beta K_2^2 + O(\beta^2) \tag{33c}$$

The expected cumulants of a finite population after selection are retrieved through equations (15a) to (15d). For the zero noise case ($\sigma = 0$) this is equivalent to selecting directly on overlaps (with energy $-R$), with selection strength $k\beta$. We will therefore call $k\beta$ the effective selection strength. It has previously been shown that this parameter should be scaled inversely with the standard deviation of the population in order to make continued progress under selection, without converging too quickly [16]. As in the problems considered in reference [16], the finite population effects lead to a reduced variance and an increase in the magnitude of the third cumulant, related to the skewness of the population. This leads to an accelerated reduction in variance under further selection. The noise due to the finite training set increases the size of the finite population effects. The other genetic operators, especially crossover, reduce the magnitude of the higher cumulants to allow further progress under selection.

6. The correlation after selection

To model the full dynamics, it is necessary to evolve the mean correlation within the population under selection. This is rather tricky, as it requires knowledge of the relationship between overlaps and correlations within the population. To make the problem tractable, it is assumed that before selection the population is at maximum entropy with constraints on the mean overlap and correlation within the population, as discussed in Appendix A. The calculation presented here is similar to that presented elsewhere [17], except for a minor refinement which seems to be important when considering problems with noise under selection.

The correlation of an infinite population after selection from a finite population is given by,

$$\begin{aligned} q_{\mathrm{s}} &= \sum_{\alpha=1}^{P} p_\alpha^2 (1 - q_{\alpha\alpha}) + \sum_{\alpha=1}^{P} \sum_{\beta=1}^{P} p_\alpha p_\beta q_{\alpha\beta} \\ &= \Delta q_{\mathrm{d}} \ + \ q_\infty \end{aligned} \tag{34}$$

where p_α is the probability of selection, defined in equation (19). The first term is due to the duplication of population members under selection, while the second term is due to the natural increase in correlation as the population moves into a region of lower entropy. The second term gives the increase in the correlation in the infinite population limit, where the duplication term becomes negligible. An extra set of variables $q_{\alpha\alpha}$ are assumed to come from the same statistics as the distribution of correlations within the population. Recall that the expectation value for the correlation of a finite population is equal to the correlation of the infinite parent population from which it is sampled.

6.1. *Natural increase term*

We estimate the conditional probability distribution for correlations given overlaps before selection $p(q_{\alpha\beta}|R_\alpha, R_\beta)$ by assuming the weights within the population are distributed according to the maximum entropy distribution described in Appendix A. Then q_∞ is simply the correlation averaged over this distribution and the distribution of overlaps after selection, $p_{\mathrm{s}}(R)$.

$$q_\infty = \int \mathrm{d}q_{\alpha\beta}\, \mathrm{d}R_\alpha\, \mathrm{d}R_\beta\, p_{\mathrm{s}}(R_\alpha) p_{\mathrm{s}}(R_\beta) p(q_{\alpha\beta}|R_\alpha, R_\beta)\, q_{\alpha\beta} \tag{35}$$

This integral can be calculated for large N by the saddle point method and we find that in this limit the result only depends on the mean overlap after selection (see Appendix B).

$$q_\infty(y) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{W_i + \tanh(y)}{1 + W_i \tanh(y)} \right)^2 \tag{36}$$

where,

$$K_1^{\mathrm{s}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i + \tanh(y)}{1 + W_i \tanh(y)} \tag{37}$$

The natural increase contribution to the correlation q_∞ is an implicit function of K_1^{s} through y, which is related to K_1^{s} by equation (37). Here, W_i is the mean weight at site i before selection (recall that we have chosen the teacher's weights to be $t_i = 1$ at every site, without loss of generality) and for a distribution at maximum entropy one has,

$$W_i = \tanh(z + x\eta_i) \tag{38}$$

The Lagrange multipliers, z and x, are chosen to enforce constraints on the mean overlap and correlation within the population before selection and η_i is drawn from a Gaussian distribution with zero mean and unit variance (see Appendix A).

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

$$K_1^{\mathrm{s}} = K_1^{\mathrm{c}} + y(NK_2^{\mathrm{c}}) + \frac{y^2}{2}(N^2K_3^{\mathrm{c}}) + \cdots \tag{39}$$

$$q_\infty(y) = q - y(N^2K_3^{\mathrm{c}}) - \frac{y^2}{2}(N^3K_4^{\mathrm{c}}) + \cdots \tag{40}$$

where K_n^{c} are the infinite population expressions for the cumulants after bit-simulated crossover, when the population is assumed to be at maximum entropy (defined in equations (18a) to (18d) up to the fourth cumulant). Here, y plays the role of the effective selection strength in the associated infinite population problem, so for an infinite population one could simply set $y = k\beta/N$, where k is defined in equation (32). To calculate the correlation after selection, we solve equation (37) for y and then substitute this value into the equation (36) to calculate q_∞. In general this must be done numerically, although the weak selection expansion can be used to obtain an analytical result which gives a very good approximation in many cases. Notice that the third cumulant in equation (40) will be negative for $K_1 > 0$ because of the negative entropy gradient and this will accelerate the increased correlation under selection.

6.2. *Duplication term*

The duplication term Δq_{d} is defined in equation (34). As in the partition function calculation presented in section 5, population members are independently averaged over a distribution with the correct cumulants,

$$\begin{aligned}\Delta q_{\mathrm{d}} &= P\left(\prod_{\alpha=1}^{P}\int \mathrm{d}R_\alpha\,\mathrm{d}E_\alpha\,\mathrm{d}q_{\alpha\alpha}p(R_\alpha)\,p(E_\alpha|R_\alpha)\,p(q_{\alpha\alpha}|R_\alpha,R_\alpha)\right)\frac{(1-q_{\alpha\alpha})\mathrm{e}^{-2\beta E_\alpha}}{(\sum_\alpha \mathrm{e}^{-\beta E_\alpha})^2} \\ &= P\left(\prod_{\alpha=1}^{P}\int \mathrm{d}R_\alpha\cdots\right)(1-q_{\alpha\alpha})\exp(-2\beta E_\alpha)\int_0^\infty \mathrm{d}t\,t\,\exp\left(-t\sum_\alpha \mathrm{e}^{-\beta E_\alpha}\right)\end{aligned} \tag{41}$$

Here, $q_{\alpha\alpha}$ is a construct which comes from the same statistics as the correlations between distinct population members. The integral in t removes the square in the denominator and decouples the average,

$$\Delta q_{\rm d} \;=\; P\!\int_0^\infty\!\mathrm{d}t\, t\, f(t)\, g^{P-1}(t) \tag{42}$$

where,

$$f(t) = \int\!\mathrm{d}R\,\mathrm{d}E\,\mathrm{d}q\, p(R)\, p(E|R)\, p(q|R,R)\;(1-q)\exp(-2\beta E - t\mathrm{e}^{-\beta E}) \tag{43}$$

$$g(t) = \int\!\mathrm{d}R\,\mathrm{d}E\, p(R)p(E|R)\exp(-t\mathrm{e}^{-\beta E}) \tag{44}$$

The overlap distribution $p(R)$ will be approximated by the cumulant expansion in equation (25) and $p(q|R,R)$ by the distribution derived in Appendix B. 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.

Instead, we expand for small β and large N as we did for the selection calculation in section 5.1 (this approximation is only used for the term involving the correlation in equation (42) for the simulations presented in section 8). In this case one finds,

$$\begin{aligned} f(t)\, g^{P-1}(t) \;\simeq\; & \hat\rho(2\beta)\exp\left[-t\left((P-1)\hat\rho(\beta)+\frac{\hat\rho(3\beta)}{\hat\rho(2\beta)}\right)\right] \\ & -\hat\rho_q(2\beta)\exp\left[-t\left((P-1)\hat\rho(\beta)+\frac{\hat\rho_q(3\beta)}{\hat\rho_q(2\beta)}\right)\right] \end{aligned} \tag{45}$$

where,

$$\hat\rho(\beta) \;=\; \int\!\mathrm{d}R\,\mathrm{d}E\, p(R)\, p(E|R)\, \mathrm{e}^{-\beta E} \tag{46}$$

$$\hat\rho_q(\beta) = \int\!\mathrm{d}R\,\mathrm{d}E\, p(R)\, p(E|R)\int\!\mathrm{d}q\, p(q|R,R)\, q\, \mathrm{e}^{-\beta E} \tag{47}$$

Completing the integral in equation (42) one finds,

$$\Delta q_{\rm d} = \frac{\hat\rho(2\beta)-\hat\rho_q(2\beta)}{P\hat\rho^2(\beta)} + O\!\left(\frac{1}{P^2}\right) \tag{48}$$

We express $\hat\rho_q(\beta)$ in terms of the Fourier transform of the distribution of correlations, which is defined in equation (B15),

$$\hat\rho_q(\beta) = \lim_{t\to 0}\frac{\partial}{\partial t}\log\left(\int\!\mathrm{d}R\,\mathrm{d}E\, p(R)\, p(E|R)\hat\rho(-\mathrm{i}t|R,R)\,\mathrm{e}^{-\beta E}\right)\hat\rho(\beta) \tag{49}$$

The integrals can be calculated by expressing $p(E|R)$ by the same approximate form as in section 5.1 and using the saddle point method to integrate over the Fourier transform as in Appendix B.

Eventually one finds,

$$\Delta q_{\rm d} \;=\; \frac{\mathrm{e}^{(\beta\sigma)^2}[1-q_\infty(2k\beta/N)]\rho_2(k\beta,0)}{P\rho_1^2(k\beta,0)} \;+\; O\Big(\frac{1}{P^2}\Big) \tag{50}$$

where $q_\infty(y)$ is defined in equation (36) and $\rho_n(k\beta,\gamma)$ is defined in equation (31).

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

$$\Delta q_{\rm d} \;\simeq\; \frac{\mathrm{e}^{(\beta\sigma)^2}}{P}\,[1-q_\infty(2k\beta/N)]\,\Big(1+K_2(k\beta)^2-K_3(k\beta)^3+O(\beta^4)\Big) \tag{51}$$

The q_∞ term has not been expanded out since it contributes terms of $O(1/N)$ less than these contributions for each cumulant. Selection leads to a negative third cumulant (see equation (33c)), 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.

7. Dynamic population resizing

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 sections 5.1 and 6.2, this can be compensated for by increasing the population size. The terms which involve noise in equations (30) and (50) can be removed by an appropriate population resizing,

$$P = P_0\exp[(\beta\sigma)^2] \tag{52}$$

Here, P_0 is the population size in the infinite training set, zero noise limit. Since these are the only terms in the expressions describing the dynamics which involve the finite population size, this effectively maps the full dynamics onto the infinite training set case.

For zero noise the selection strength should be scaled so that the effective selection strength $k\beta$ is inversely proportional to the standard deviation of the population [15],

$$\beta = \frac{\beta_{\rm s}}{k\sqrt{\kappa_2}} \tag{53}$$

Here, k is defined in equation (32) and $\beta_{\rm s}$ is the scaled selection strength and remains fixed throughout the search†. Recall that κ_2 is the expected variance of a

† This scaling of selection strength (equation (52)) requires overlap statistics which will not be known in practice. However, the results do not rely on this choice and any fixed schedule for determining β each generation could be used. This choice corresponds to an appropriate schedule for the infinite training set problem.

finite population, which is related to the variance of an infinite population through equation (15*b*). One could also include a factor of $\sqrt{\log P}$ to compensate for changes in population size, as in reference [16], but this term is neglected here. The resized population is then,

$$\begin{aligned} P &= P_0 \exp\left(\frac{(\beta_{\rm s}\sigma)^2}{k^2\kappa_2}\right) \\ &= P_0 \exp\left(\frac{\beta_{\rm s}^2(1-\kappa_1^2)\cos^{-1}(\kappa_1)(\pi-\cos^{-1}(\kappa_1))}{\lambda N\kappa_2}\right) \end{aligned} \tag{54}$$

Notice that the exponent in this expression is $O(1)$, so this population resizing does not blow up with increasing problem size. One might therefore expect this problem to scale with N in the same manner as the zero-noise, infinite training set case, as long as the batch size is $O(N)$.

Baum *et al* have shown that a closely related GA scales as $O(N\log_2^2 N)$ on this problem if the population size is sufficiently large so that weights can be assumed to come from a binomial distribution [1]. This is effectively a maximum entropy assumption with a constraint on the mean overlap alone. They use culling selection, where the best half of the population survives each generation leading to a change in the mean overlap proportional to the population's standard deviation. Our selection scaling also leads to a change in the mean of this order and the algorithms may therefore be expected to compare closely. The expressions derived here do not rely on a large population size and are therefore more general.

In the infinite population limit it is reasonable to assume $N\kappa_2 \simeq 1-\kappa_1^2$ which is the relationship between mean and variance for a binomial distribution, since in this limit the correlation of the population will not increases due to duplication under selection. In this case the above scaling results in a monotonic decrease in population size, as κ_1 increases over time. This is easy to implement by removing the appropriate number of population members before each selection.

A finite population becomes correlated under selection and the variance of the population is usually less than the value predicted by a binomial distribution. In this case the population size may have to be increased, which could be implemented by producing a larger population after selection or crossover. This is problematic, however, since increasing the population size leads to an increase in the correlation and a corresponding reduced performance. In this case the dynamics will no longer be equivalent to the infinite training set situation.

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

$$\lambda = \frac{\beta_{\rm s}^2(1-\kappa_1^2)\cos^{-1}(\kappa_1)(\pi-\cos^{-1}(\kappa_1))}{N\kappa_2\log(P/P_0)} \tag{55}$$

Figure 1 shows how choosing the batch size each generation according to equation (55) leads to the dynamics converging onto the infinite training set dynamics where the training energy is equal to the generalization error. 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. To a good approximation it seems that the population resizing in equation (54) and the corresponding batch sizing expression in equation (55) are accurate, at least as long as λ is not too small.

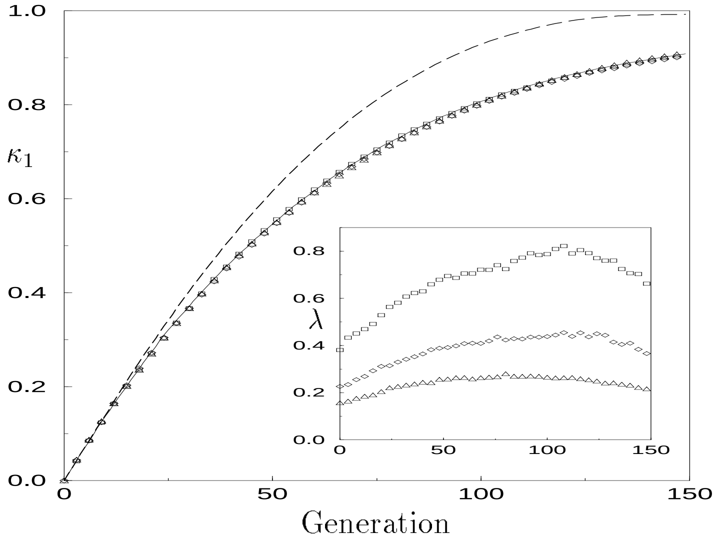

Figure 1. The mean overlap between teacher and pupil within the population is shown each generation for a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. The results were averaged over 100 runs and training batch sizes were chosen according to equation (55), leading to the trajectories converging onto the infinite training set result where $E = E_{\mathrm{g}}(R)$. The solid curve is for the infinite training set with $P_0 = 60$ and the finite training set results are for $P = 90$ ($\Box$), 120($\diamond$) and 163($\triangle$). The inset shows the mean choice of λ 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 $N = 279$, $\beta_{\mathrm{s}} = 0.25$ and $p_{\mathrm{m}} = 0.001$.

7.1. Optimal batch size

In the previous section it was shown how the population size could be changed to remove the effects of noise associated with a finite training set. If we use this population resizing then it is possible to define an optimal size of training set, in order to minimize the computational cost of energy evaluation. This choice will also minimize the total number of training examples presented when independent batches are used. This may be expected to provide a useful estimate of the appropriate sizing of batches in more efficient schemes, where examples are recycled, as long as the total number of examples used significantly exceeds the threshold above which over-training is impossible.

We assume that computation is mainly due to energy evaluation and note that there are P energy evaluations each generation with computation time for each scaling as λ. If the population size each generation is chosen by equation (54), then the computation time τ_c (in arbitrary units) is given by,

$$\tau_c \;=\; \lambda \exp\left(\frac{\lambda_o}{\lambda}\right) \qquad \lambda_o = \frac{\beta_{\mathrm{s}}^2(1-\kappa_1^2)\cos^{-1}(\kappa_1)(\pi - \cos^{-1}(\kappa_1))}{N\kappa_2} \tag{56}$$

The optimal choice of λ is given by the minimum of τ_c, which is at λ_o. Choosing this batch size leads to the population size being constant over the whole GA run and for optimal performance one should choose,

$$P = P_0\,\mathrm{e}^1 \;\simeq\; 2.73 P_0 \tag{57}$$

$$\lambda \;=\; \lambda_o \tag{58}$$

where P_0 is the population size used for the zero noise, infinite training set GA. Notice that it is not necessary to determine P_0 in order to choose the size of each batch, since λ_o is not a function of P_0. Since the batch size can now be determined automatically, this reduces the size of the GA's parameter space significantly.

One of the runs in figure 1 is for this choice of P and λ, showing close agreement to the infinite training set dynamics ($P = 163 \;\simeq\; P_0\mathrm{e}$). In general, the first two cumulants change in a non-trivial manner each generation and their evolution can be determined by simulating the dynamics, as described in section 8.

8. Simulating the dynamics

In sections 4, 5 and 6, difference equations were derived for the mean effect of each operator on the mean overlap and correlation within the population. The full dynamics of the GA can be simulated by iterating these equations starting from their initial values, which are zero. The equations for selection also require knowledge of the higher cumulants before selection, which are calculated by assuming a maximum entropy distribution with constraints on the two known macroscopics (see equations (18*a*) to (18*d*)). We used four cumulants and the selection expressions were calculated numerically, although for weak selection the analytical results in section 5.1 were also found to be very accurate. The largest overlap within the population was estimated by assuming population members were randomly selected from a distribution with the correct cumulants [16]. This assumption breaks down towards the end of the search, when the population is highly correlated and the higher cumulants become large, so that four cumulants may not describe the population sufficiently well.

Figures 2 and 3 show the mean, variance and largest overlap within the population each generation, averaged over 1000 runs of a GA and compared to the theory. The infinite training set case, where the training energy is the generalization error, is compared to results for two values of λ, showing how performance degrades as the batch size is reduced. Recall that λN new patterns are shown to each population member, each generation, so that the total number of patterns used is $\lambda N \times PG$, where P is population size and G is the total number of generations. The skewness and kurtosis are presented in figure 4 for one value of λ, showing that although there are larger fluctuations in the higher cumulants they seem to agree sufficiently well with the theory on average. It would probably be possible to model the dynamics accurately with only three cumulants, since the kurtosis does not seem to be particularly significant in these simulations.

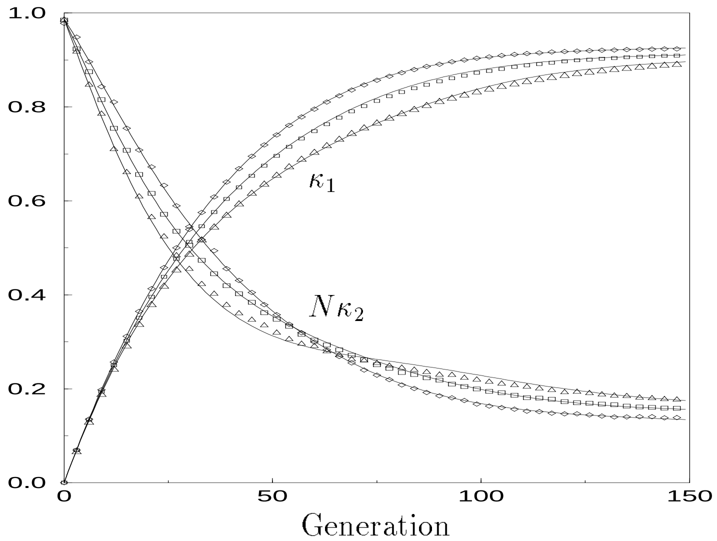

Figure 2. The theory is compared to averaged results from a GA training a binary perceptron to generalize from examples produced by a teacher perceptron. The mean and variance of the overlap distribution within the population are shown, averaged over 1000 runs, with the solid lines showing the theoretical 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 $N = 155$, $\beta_{\rm s} = 0.3$, $p_{\rm m} = 0.005$ and the population size was 80.

These results show excellent agreement with the theory, although there is a slight underestimate in the best population member for the reasons discussed above. This is typical of the theory, which has to be very accurate in order to pick up the subtle effects of noise due to the finite batch size. Unfortunately, the agreement is less accurate for low

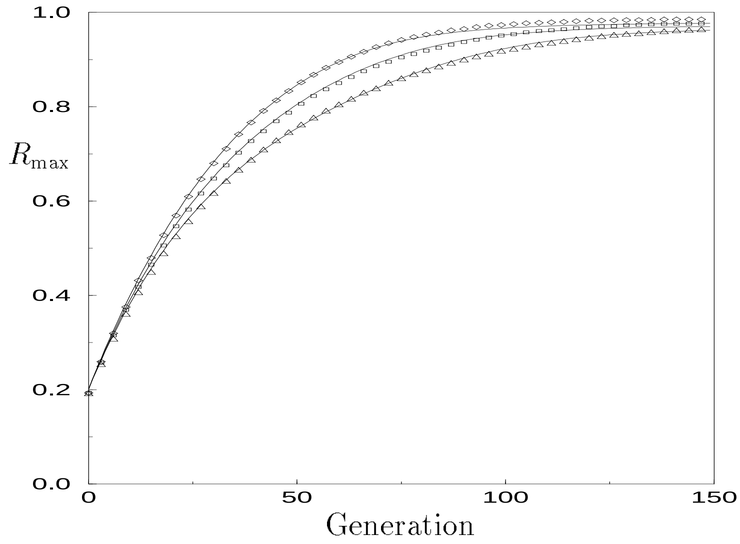

Figure 3. The maximum overlap between teacher and pupil is shown each generation, averaged over the same runs as the results presented in figure 2. The solid lines show the theoretical predictions and the symbols are as in figure 2.

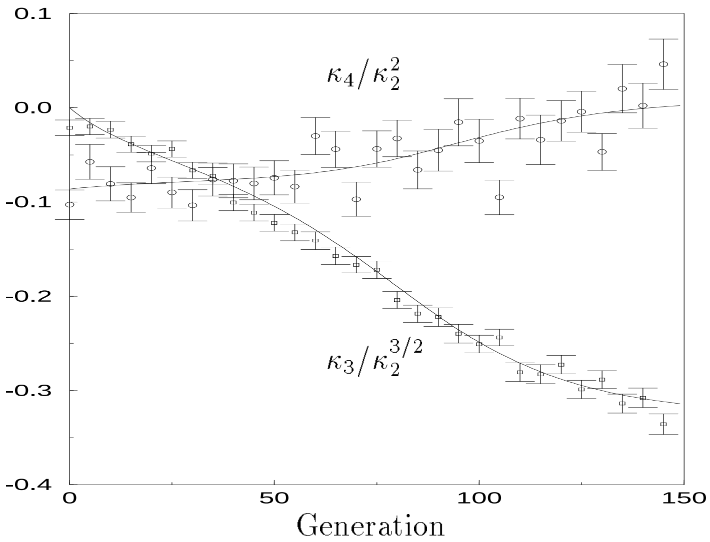

Figure 4. The skewness and kurtosis of the overlap distribution are shown averaged over the same runs as the results presented in figure 2 for $\lambda = 0.65$. Averages were taken over cumulants, rather than the ratios shown. The solid lines show the theoretical predictions for mean behaviour.

values of λ, where the noise is stronger. This may be due to two simplifications. Firstly, we use a Gaussian approximation for the noise which relies on λ being at least $O(1)$. This could be remedied by expanding the noise in terms of more than two cumulants as we have done for the overlap distribution. Secondly, the duplication term in section 6.2 uses the large N, weak selection approximation which also relies on λ being $O(1)$. The error due to this approximation is minimized by only using the approximation for the term involving the correlation in equation (42), with the other term calculated numerically. It is expected that good results for smaller values of λ would be possible for larger values of N, where the correlation calculation would be more exact.

9. Conclusion

A statistical mechanics formalism has been used to solve the dynamics of a GA for a simple problem from learning theory, generalization in a perceptron with binary weights. To make the dynamics tractable, the case where a new batch of examples was presented to each population member each generation was considered. For $O(N)$ training examples per batch the training energy was well approximated by a Gaussian distribution whose mean is the generalization error and whose variance increases as the batch size is reduced. The use of bit-simulated crossover, which takes the population straight to the fixed point of standard crossover, allowed the dynamics to be modelled in terms of only two macroscopics; the mean correlation and overlap within the population. The higher cumulants of the overlap distribution after crossover were required to calculate the effect of selection and were estimated by assuming maximum entropy with respect to the two known macroscopics. By iterating difference equations describing the average effect of each operator on the mean correlation and overlap the dynamics of the GA were simulated, showing very close agreement with averaged results from a GA.

Although the difference equations describing the effect of each operator required numerical enumeration in some cases, analytical results were derived for the weak selection, large N limit. It was shown that in this limit a dynamical resizing of the population maps the finite training set dynamics onto the infinite training set situation. Using this resizing it is possible to calculate the most computationally efficient size of population and training batch, since there is a diminishing return in improved performance as batch size is increased. For the case of independent training examples considered here this choice also gives the minimum total number of examples presented.

In future work it would be essential to look at the situation where the patterns are recycled, leading to a much more efficient use of training examples and the possibility of over-training. In this case, the distribution of overlaps between teacher and pupil would not be sufficient to describe the population, since the training energy would then be dependent on the training set. One would therefore have to include information specific to the training set, such as the mean pattern per site within the training set. This might be treated as a quenched field at each site, although it is not obvious how one could best incorporate such a field into the dynamics.

Another interesting extension of the present study would be to consider multi-layer networks, which would present a much richer dynamical behaviour than the single-layer perceptron considered here. This would bring the formalism much closer to problems of realistic difficulty. In order to describe the population in this case it would be necessary to consider the joint distribution of many order parameters within the population. It would be interesting to see how the dynamics of the GA compares to gradient methods in networks with continuous weights, for which the dynamics of generalization for a class of multi-layer architectures have recently been solved analytically in the case of on-line learning [18]. In order to generalize in multi-layer networks it is necessary for the search to break symmetry in weight space and it would be of great interest to understand how this might occur in a population of solutions, whether it would occur spontaneously over the whole population in analogy to a phase transition or whether components would be formed within the population, each exhibiting a different broken symmetry. This would again require the accurate characterization of finite population effects, since an infinite population might allow the coexistence of all possible broken symmetries, which is presumably an unrealizable situation in finite populations.

Acknowledgments

We would like to thank Adam Prügel-Bennett for many helpful discussions and for providing code for some of the numerical work 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).

Appendix A. The maximum entropy distribution

After bit-simulated crossover the population is assumed to be at maximum entropy with constraints on the mean overlap and correlation within the population. This is a special case of the result derived for the paramagnet by Prügel-Bennett and Shapiro [16] and this discussion follows theirs closely.

Let W_i be the mean weight at site i within the population,

$$W_i = \langle w_i^\alpha \rangle_\alpha = \frac{1}{P}\sum_{\alpha=1}^{P} w_i^\alpha \tag{A1}$$

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

$$zPK_1 \;=\; \frac{z}{N}\sum_{\alpha=1}^{P}\sum_{i=1}^{N} w_i^\alpha \;=\; \frac{zP}{N}\sum_{i=1}^{N} W_i \tag{A2}$$

$$\frac{(xP)^2}{2}q \;=\; \frac{x^2}{2N}\sum_{\alpha=1}^{P}\sum_{\beta=1}^{P}\sum_{i=1}^{N} w_i^\alpha w_i^\beta \;=\; \frac{(xP)^2}{2N}\sum_{i=1}^{N} W_i^2 \tag{A3}$$

Recall that we have chosen $t_i = 1$ at each site without loss of generality. The correlation expression is for large P and finite population corrections can be included retrospectively.

Without constraints, the fraction of positive weights at site i is given by a binomial coefficient,

$$\Omega(W_i) = \frac{1}{2^P}\begin{pmatrix} P \\ P(1+W_i)/2 \end{pmatrix} \tag{A4}$$

So one can define an entropy,

$$\begin{aligned} S(W_i) &= \log[\Omega(W_i)] \\ &\sim -\frac{P}{2}\log(1-W_i^2) + \frac{PW_i}{2}\log\left(\frac{1-W_i}{1+W_i}\right) \end{aligned} \tag{A5}$$

where Stirling's approximation has been used. One can then define a probability distribution for the $\{W_i\}$ configuration which decouples at each site,

$$p(\{W_i\}) = \prod_{i=1}^{N} p(W_i) = \prod_{i=1}^{N} \exp[S(W_i) + zPW_i + (xPW_i)^2/2] \tag{A6}$$

$$p(W_i) = \int \frac{\mathrm{d}\eta_i}{\sqrt{2\pi}} \exp\left(\frac{-\eta_i^2}{2} + PG(W_i, \eta_i)\right) \tag{A7}$$

where

$$G(W_i, \eta_i) = S(W_i)/P + zW_i + x\eta_i W_i \tag{A8}$$

The maximal value of G with respect to W_i gives the maximum entropy distribution for W_i at each site. This leads to the expression,

$$W_i = \tanh(z + x\eta_i) \tag{A9}$$

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

$$K_1 = \frac{1}{N}\sum_{i=1}^{N} \overline{\tanh(z + x\eta_i)} \tag{A10}$$

$$q = \frac{1}{N}\sum_{i=1}^{N} \overline{\tanh^2(z + x\eta_i)} \tag{A11}$$

The bars denote averages over the Gaussian noise which in general must be done numerically.

The third and fourth order terms in equations (18c) and (18d) can be found once the Lagrange multipliers have been determined,

$$\frac{1}{N}\sum_{i=1}^{N} \langle w_i^\alpha \rangle_\alpha^3 = \overline{\tanh^3(z + x\eta)} \tag{A12}$$

$$\frac{1}{N}\sum_{i=1}^{N} \langle w_i^\alpha \rangle_\alpha^4 = \overline{\tanh^4(z + x\eta)} \tag{A13}$$

Again, the bars denote averages over the Gaussian noise.

Appendix B. The distribution of correlations

Rewriting equation (35) we have,

$$\begin{aligned} q_\infty &= \int \mathrm{d}q_{\alpha\beta}\,\mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_{\mathrm{s}}(R_\alpha)p_{\mathrm{s}}(R_\beta)\,p(q_{\alpha\beta}|R_\alpha,R_\beta)\,q_{\alpha\beta} \\ &= \lim_{t\to 0}\frac{\partial}{\partial t}\log\left(\int \mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_{\mathrm{s}}(R_\alpha)\,p_{\mathrm{s}}(R_\beta)\,\hat{\rho}(-\mathrm{i}t|R_\alpha,R_\beta)\right) \end{aligned} \tag{B14}$$

where $\hat{\rho}(-\mathrm{i}t|R_\alpha,R_\beta)$ is the Fourier transform of $p(q_{\alpha\beta}|R_\alpha,R_\beta)$,

$$\hat{\rho}(-\mathrm{i}t|R_\alpha,R_\beta) \;=\; \int \mathrm{d}q_{\alpha\beta}\,p(q_{\alpha\beta}|R_\alpha,R_\beta)\mathrm{e}^{tq_{\alpha\beta}} \tag{B15}$$

The conditional probability for correlations $p(q_{\alpha\beta}|R_\alpha,R_\beta)$ can be defined if weights are assumed to come from the maximum entropy distribution defined in Appendix A. In this case one has,

$$\begin{aligned} p(q_{\alpha\beta}|R_\alpha,R_\beta) &= \frac{p(q_{\alpha\beta},R_\alpha,R_\beta)}{p(R_\alpha,R_\beta)} \\ &= \frac{\langle\delta(q_{\alpha\beta}-\frac{1}{N}\sum_i w_i^\alpha w_i^\beta)\delta(R_\alpha-\frac{1}{N}\sum_i w_i^\alpha)\delta(R_\beta-\frac{1}{N}\sum_i w_i^\beta)\rangle}{\langle\delta(R_\alpha-\frac{1}{N}\sum_i w_i^\alpha)\delta(R_\beta-\frac{1}{N}\sum_i w_i^\beta)\rangle} \end{aligned} \tag{B16}$$

where the angled brackets denote averages over w_i^α and w_i^β. The weights at each site are distributed according to,

$$p(w_i) = \left(\frac{1+W_i}{2}\right)\delta(w_i-1) + \left(\frac{1-W_i}{2}\right)\delta(w_i+1) \tag{B17}$$

Here, W_i is the mean weight per site, defined in equation (A9).

We consider the Fourier transform of $p(q_{\alpha\beta}|R_\alpha,R_\beta)$ since this appears in the appropriate generating function,

$$\hat{\rho}(-\mathrm{i}t|R_\alpha,R_\beta) = \frac{\hat{\rho}(-\mathrm{i}t,R_\alpha,R_\beta)}{\hat{\rho}(0,R_\alpha,R_\beta)} \tag{B18}$$

Writing the delta functions as integrals and noting that one of the integrals is removed by the Fourier transform, one finds (ignoring multiplicative constants),

$$\hat{\rho}(-\mathrm{i}t,R_\alpha,R_\beta) = \left\langle \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}y_\alpha \mathrm{d}y_\beta \exp(F)\right\rangle_{\{w_i^\alpha,w_i^\beta\}} \tag{B19}$$

$$F = -y_\alpha R_\alpha - y_\beta R_\beta + \frac{1}{N}\sum_{i=1}^{N}(y_\alpha w_i^\alpha + y_\beta w_i^\beta + t w_i^\alpha w_i^\beta)$$

Each site decouples and the average over sites can be taken by integrating over the weight distribution defined in equation (B17). The resulting integral can be computed for large N by the saddle point method since the exponent can be made extensive by appropriate rescaling. Eventually one finds (ignoring multiplicative constants),

$$\hat{\rho}(-\mathrm{i}t,R_\alpha,R_\beta) = \exp(-y_\alpha R_\alpha - y_\beta R_\beta + G) \tag{B20}$$

$$G = \frac{1}{N}\sum_{i=1}^{N}\log\left[(1+W_i)^2\mathrm{e}^{t+y_\alpha+y_\beta} + 2(1-W_i^2)\mathrm{e}^{-t}\cosh(y_\alpha - y_\beta) + (1-W_i)^2\mathrm{e}^{t-y_\alpha-y_\beta}\right]$$

The saddle point equations fix y_α and y_β as implicit functions of R_α, R_β and t,

$$R_\alpha = \frac{\partial G}{\partial y_\alpha} \qquad R_\beta = \frac{\partial G}{\partial y_\beta} \tag{B21}$$

Define $\hat\rho(-\mathrm{i}t)$, whose logarithm is the generating function for q_∞,

$$\begin{aligned}\hat\rho(-\mathrm{i}t) &= \int \mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_\mathrm{s}(R_\alpha)\,p_\mathrm{s}(R_\beta)\,\hat\rho(-\mathrm{i}t|R_\alpha,R_\beta)\\ &= \int \mathrm{d}R_\alpha\,\mathrm{d}R_\beta\,p_\mathrm{s}(R_\alpha)\,p_\mathrm{s}(R_\beta)\exp[G(t)-G(0)]\end{aligned} \tag{B22}$$

We express the overlap distributions by their Fourier transformed cumulant expansions,

$$p_\mathrm{s}(R_\alpha) = -\mathrm{i}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{\mathrm{d}a}{2\pi}\exp\left(\sum\frac{a^n}{n!}K_n^\mathrm{s} - aR_\alpha\right) \tag{B23}$$

$$p_\mathrm{s}(R_\beta) = -\mathrm{i}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{\mathrm{d}b}{2\pi}\exp\left(\sum\frac{b^n}{n!}K_n^\mathrm{s} - bR_\beta\right) \tag{B24}$$

Now $\hat\rho(-\mathrm{i}t)$ is an integral over a, b, R_α and R_β which can again be computed by the saddle point method. One finds that as $t\to 0$, the saddle point equations are satisfied by,

$$y_\alpha = y_\beta = y \tag{B25}$$

$$R_\alpha = R_\beta = K_1^\mathrm{s} \tag{B26}$$

These are related through an implicit function for y in terms of mean overlap after selection,

$$K_1^\mathrm{s} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i + \tanh(y)}{1 + W_i\tanh(y)} \tag{B27}$$

Then the natural increase contribution for the correlation after selection is given by,

$$\begin{aligned}q_\infty &= \lim_{t\to 0}\frac{\partial}{\partial t}\log\hat\rho(-\mathrm{i}t)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i+\tanh(y)}{1+W_i\tanh(y)}\right)^2\end{aligned} \tag{B28}$$

References

[1] Baum E B, Boneh D and Garret C 1995 *COLT '95: Proc. of the 8th Annual Conf. on Computational Learning Theory (New York)* p 230–239

[2] Davis L 1991 *Handbook Of Genetic Algorithms* (Van Nostrand Reinhold, New York)

[3] De la Maza M and Tidor B 1991 *Proc. of the ORSA CSTS Conf. - Computer Science and Operations Research: New Developments in their Interfaces* p 425–440

[4] Derrida B 1981 *Phys. Rev.* B **24** 2613–25

[5] Falconer D S 1989 *Introduction to Quantitative Genetics* (Longman Scientific and Technical, Burnt Mill, England)

[6] Forrest S and Mitchell M 1993 *Foundations of Genetic Algorithms 2* ed Whitley L D (Morgan Kaufmann, San Mateo, Calif) p 109–129

[7] Fitzpatrick J M and Grefenstette J J 1988 *Machine Learning* **3** 101–120

[8] Goldberg D E 1989 *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA)

[9] Grefenstette J J 1993 *Foundations of Genetic Algorithms 2* ed Whitley L D (Morgan Kaufmann, San Mateo, Calif) p 75

[10] Györgyi G 1990 *Phys. Rev.* A **41** 7097–7100

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

[12] Mühlenbein H and Schlierkamp-Voosen D 1995 *Lecture Notes in Computer Science* **899** 142–168

[13] Peck C C and Dhawan A P 1995 *Evolutionary Computation* **3** 1 39–80

[14] Prügel-Bennett A 1996 Modelling Evolving Populations, NORDITA, Blegdamsvej 17, DK-2100 Copenhagen, Denmark, (submitted for publication)

[15] Prügel-Bennett A and Shapiro J L 1994 *Phys. Rev. Lett.* **72** 1305

[16] ——1995 The Dynamics of a Genetic Algorithm for Simple Random Ising Systems, Computer Science Dept., University of Manchester, Oxford Road, Manchester M13 9PL, UK (to appear in *Physica D*)

[17] Rattray L M 1995 *Complex Systems* **9**(3) 213–234

[18] Saad D and Solla S A 1995 *Phys. Rev.* E **52** 4225

[19] Schaffer J D, Whitley D and Eshelman L J 1992 *Proc. of the Int. Conf. on Combinations of Genetic Algorithms and Neural Networks* (IEEE Computer Society Press, Los Alamitos, CA) p 1–37

[20] Thierens D and Goldberg D 1995 *Lecture Notes in Computer Science* **866** 119–129

[21] Shapiro J L, Prügel-Bennett A and Rattray L M 1994 *Lecture Notes in Computer Science* **865** 17

[22] Sompolinsky H and Tishby N 1990 *Phys. Rev. Lett.* **65** 1683–86

[23] Srinivas M and Patnaik L M 1996 *IEEE Trans. Knowledge Data Eng.* **8** 1 120–133

[24] Syswerda G 1993 *Foundations of Genetic Algorithms 2* ed Whitley L D (Morgan Kaufmann, San Mateo, CA)

[25] Vose M D and Liepins G E 1991 *Complex Systems* **5** 31

[26] Vose M D 1993 *Foundations of Genetic Algorithms 2* ed Whitley L D (Morgan Kaufmann, San Mateo, Calif.) p 63

[27] Vose M D and Wright A H 1995 *Evolutionary Computation* **2** 4 347–368

[28] Yao X 1993 *Int. J. of Neural Systems* **4**(3) 203–222