When A Genetic Algorithm Outperforms Hill-Climbing

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

12th March 2004
(First submitted July 2001)

Abstract

A toy optimisation problem is introduced which consists of a fitness gradient broken up by a series of hurdles. The performance of a hill-climber and a stochastic hill-climber are computed. These are compared with the empirically observed performance of a genetic algorithm (GA) with and without. The hill-climber with a sufficiently large neighbourhood outperforms the stochastic hill-climber, but is outperformed by a GA both with and without crossover. The GA with crossover substantially outperforms all the other heuristics considered here. The relevance of this result to real world problems is discussed.

1 Introduction

Genetic algorithms have become an increasingly popular tool for solving optimisation problems in real world applications. Their popularity has grown despite a lack of an established theory explaining why they work (and why on occasions they don't). This failure of theory has left practitioners with little guidance in determining how to represent the problem, what operators to use or how to adjust parameters such as the population size or mutation rate. A further consequence has been that the whole field is often viewed with suspicion. The dearth of a theory directly useful to the practitioner has two main causes. The first is a problem for all optimisation techniques, namely the variety and complexity of the fitness landscapes for different problems precludes the possibility of an overarching theory relevant to all problem. The second, unique to evolutionary algorithms, is the intrinsic difficulty in describing the evolution of a population.

One contribution of theory, and the objective of this paper, is to provide an intuition for how GAs work by studying how they perform on various 'toy problems'. In this regard, a goal which has attracted a considerable amount of research has been to find types of problems where GAs perform well. One prominent proposal was the Royal Road function, motivated by John Holland's schema-theory. Unfortunately, GAs were not successful, as described in the paper with the lamentable title *When Will A Genetic Algorithm Beat Hill-Climbing?* [1]. Nevertheless, the problem generated considerable interest and was subsequently modelled in some detail [2,3]. A toy problem were GAs do better has been proposed in the preprint *Where Genetic Algorithms Excel* by Baum *et al.* (as far as I know the full paper has never been published [4]). Here the problem was to find a string of length L consisting of characters from an alphabet of size K. The fitness measure was the number of sites with the same character as the target sequence. For sufficiently large alphabets a (specially designed) GA was shown to outperform a hill-climber. Baum *et al.* also showed that the effect of noise added to the fitness function was to slightly slow down the GA, but not to dramatically change its evolution. This observation has important consequences in understanding the problem described in this paper.

Neither of the above problems contained local optima, so hill-climbing was a fast alternative. A toy model involving a local optimum as well as a global optimum is the *basin with a barrier* problem [5]. Here, both a GA and a local search method took exponential time, but the exponent for the GA was substantially smaller than for the local search method, so that as the problem becomes sufficiently large it became infeasible for a local search algorithm, but still feasible for a GA. If the problem becomes too large finding the optimal solution becomes infeasible for both algorithms. By carefully tuning the problem it is possible to set up a situation where a GA finds the global optimum in polynomial time, whereas a local search method takes exponential time [6]. The key component is the crossover operator which focuses the search. This same effect is seen in the problem analysed in this paper and will be discussed in more detail later. The *basin with a barrier* has been studied analytically in [7,8].

In the current paper, we consider a new family of toy problems where the fitness depends on the Hamming distance from a global optimum, but to reach the optimum requires overcoming a series of hurdles. The difficulty of the problem is determine by the hurdle widths, $n \in \{2, 3, \ldots\}$. The problems are defined over binary string of length L (i.e. $\{0,1\}^L$). Without loss of generality, we can take the global optimum to be the state of all zeros. Thus the Hamming distance is given by the number of ones. The fitness depends on the Hamming distance, H, according to

$$F_n(H) = -\left\lceil \frac{H}{n} \right\rceil - \frac{\mathrm{rem}(H,n)}{n} \tag{1}$$

where $\lceil x \rceil$ denotes the ceiling function (i.e. the least integer greater than x) and $\mathrm{rem}(H,n)$ the remainder of H modulo n. The task is to find the string which maximises the fitness, $F_n(H)$. In this paper we will mainly restrict our attention to the cases $n = 2$ and $n = 3$. The fitness functions are shown in figure 1.

They can be regarded as a roughened ones-counting problem, i.e. $F_1(H) = -H$ (the ones-counting problem can be viewed as the $n = 1$ limit in this family of toy problems). We note that there is an overall fitness gradient with many local optima to single bit mutations. Figure 1 does not show the strong entropy gradient. The number of states with Hamming distance H is $\binom{L}{H}$, which for $H \ll L$ grows approximately as $L^H/H!$.

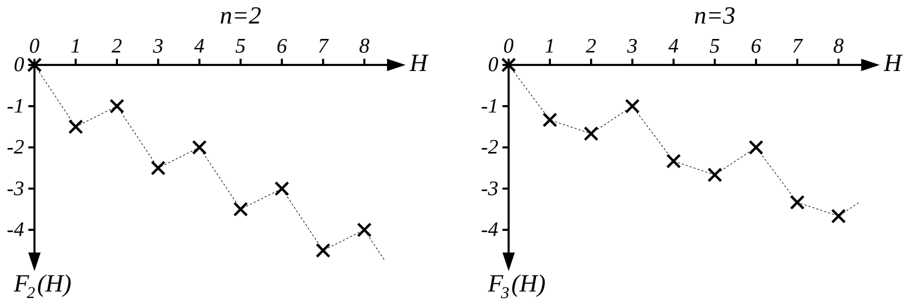

Figure 1: Fitness function close to the global optimum for the hurdle problem with $n = 2$ and $n = 3$.

We shall consider three types of search techniques on this problem, hill-climbing, stochastic hill-climbing (Metropolis) and a GA. Our major concern will be the scaling behaviour, i.e. how the typical first-passage time to reach the global optimum scales with the string length L. For n of order 1, we will see that the hill-climber takes of order L^n steps (where n is the model type) and a stochastic hill-climber takes $\Theta(L^{n+1})$ steps. A genetic algorithm without crossover appears to take $\Theta(L^n)$; similar to the hill-climber, although it requires less function evaluations. Finally a GA with crossover appears to find the global optimum in close to linear time, significantly outperforming the other search algorithms, although, this performance advantage is unlikely to persist as $L \to \infty$.

In the next section, we obtain approximations for the first-passage time for the three algorithms. We conclude in section 3 by drawing some more general lessons from these results and relate them to more realistic problems. Finally, we include an appendix where we give some basic results concerning first-passage times for Markov chains.

2 First-Passage Times

We are concerned with the average number of fitness evaluation required to reach the global optimum, which we shall refer to as the first passage time. The first-passage time can vary considerably between runs. We will mostly be concerned with the mean first-passage time, T_{fpt}, although we will also examine the median first-passage time, $\tilde{T}_{\mathrm{fpt}}$, which is typically less (and often substantially less) than the mean for this type of problem.

The local search techniques can be described by a Markov chain where the states are the Hamming distance from the global optimum and the transition probabilities are given by the transition matrix $\mathbf{W}$ with elements

$$W_{ij} = \mathrm{P}\big(H = j \to H = i\big).$$

The transition matrix depends on the local search algorithm. For fixed length chains, it is possible to compute the mean and median first-passage time exactly using simple matrix algebra operations (see appendix). However, as we are interested in the scaling behaviour we shall compute approximations for the first-passage time valid for all L.

2.1 Hill-Climbing

We consider a hill-climber where we start from a random string and at each time step we make a random mutation which we accept provided it does not decrease the fitness (in the terminology of evolutionary algorithms this is a (1+1)-EA), A hill-climber performing single bit mutations will become, almost immediately, trapped at a local optimum. We therefore consider a hill-climber performing multiple mutations. We concern ourselves at first with the simplest hurdle model $n = 2$. If a hill-climber is currently at a local optimum a Hamming distance H from the global optimum, then it can only make progress by mutating an even number of the misaligned genes (if the target is the string of all zeros then a misaligned gene is one of the 1's in the string). By far the mostly likely beneficial mutation will involve mutating two misaligned genes. We therefore concentrate on double mutations. The probability of making the move from H to $H - 2$ is

$$\frac{H(H-1)}{L(L-1)}.$$

This probability becomes increasingly small as the hill-climber approaches the global optimum. The dynamics of a hill-climber can be viewed as a sequence of transitions from local optimum H to a neighbouring optimum $H - 2$, each step in the sequence becoming increasingly difficult. The mean time to make a transition is equal to one over the transition probability (see appendix). Thus the mean time for the transition from $H = 2$ to $H = 0$ is $L(L-1)/2$; for the transition from $H = 4$ to $H = 2$ it is $L(L-1)/12$; and for $H = 6 \to H = 4$ it is $L(L-1)/30$. In the large L limit the mean first-passage time is the time it takes to make an infinite number of these transitions

$$\sum_{h=1}^{\infty} \frac{L(L-1)}{2h(2h-1)} = L(L-1)\log(2).$$

This provides an upper-bound on the mean first-passage time starting from some local optimum using double mutations. (We could guarantee to be at a local optimum by first making a step with a single mutation).

To make the comparison fairer we should consider using a hill-climber with a fixed mutation probability per site of, $u = U/L$ (i.e. on average we make U mutations per string). The number of double mutations is given by the binomial factor $B(2|L,u) = \binom{L}{2} u^2 (1-u)^{L-2}$, which is maximised when $U = 2$. For large L, the binomial distribution can be approximated by a Poisson distribution

$$\lim_{L\to\infty} B(2|L,2/L) = 2\mathrm{e}^{-2}.$$

Beneficial mutations involving $n > 2$ sites will occur of order L^{2-n} times less frequently than double mutations and so contribute only to the sub-dominant terms. To leading order, the mean first-passage time, assuming the optimal mutation rate of $U = 2$, is

$$T_{\mathrm{fpt}} \sim \frac{L^2 \mathrm{e}^2 \log(2)}{2} \tag{2}$$

where we have ignored corrections of order L and smaller.

It is instructive to compare this result to the ones-counting problem where the probability of a transition $H \to H-1$ is H/L. The time required to reach the global optimum starting from the most likely initial state $H = L/2$ is given by

$$\sum_{H=1}^{L/2} \frac{L}{H} = L\,\mathrm{H}_{L/2} \sim L\left(\log(L/2) + \gamma\right)$$

where H_n is the n^{th} harmonic number with $\mathrm{H}_n \sim \log(n) + \gamma + O(n^{-1})$ (where γ is Euler's gamma, $\gamma \approx 0.5772$). In ones-counting, the passage from a random state to the state $H = 1$ takes of order $\log(L/2)$ times longer than the time taken for the final transition. The dynamics is therefore dominated by the complete passage. In comparison, in the $n = 2$ problem above, the time taken to reach the state $H = 2$ from random stating position is a mere $2\log(2) - 1$ or 38.6% of the time taken to make the final transition. Thus the evolution is dominated by the final transition.

A consequence of the first-passage time being dominated by a single transition is that median first-passage time will be significantly less than the mean. The median time to make a transition occurring with probability p is $-\log(2)/\log(1-p) - 1$ (compared with a mean time of $1/p$, see appendix). Thus for small p the median time is typically a factor of $-\log(2) = 0.693$ less than the mean time. In the exact numerical results the median first-passage times are around 70% smaller than the mean first-passage times.

We can generalise these results to larger n. For $n = 3$ we require triple mutations to jump from one local optimum to the next. The probability for such a transition is

$$\frac{H(H-1)(H-2)}{L(L-1)(L-2)}.$$

To leading order in L, the time taken to make a series of transition from a random state to the global optimum is

$$\sum_{h=1}^{\infty} \frac{L^3}{3h(3h-1)(3h-2)} = \frac{L^3}{12}\left(\sqrt{3}\pi - 3\log(3)\right) \approx 0.1788 L^3.$$

On average, all transitions up to the last transition takes just 7.3% of the time it takes to make the last transition. If we assume that we randomly mutate each site with a probability $u = U/L$, then the optimal mutation rate is given by $U = 3$ and the average number of triple mutations is approximately $3^3 \mathrm{e}^{-3}/3!$, so that for $n = 3$ the optimal first-passage time is

$$T_{\mathrm{fpt}} \sim 0.80 L^3.$$

For arbitrary n, the first-passage time is

$$T_{\mathrm{fpt}} \sim \frac{1}{(n!)^2} \left(\frac{\mathrm{e}L}{n}\right)^n c(n) \tag{3}$$

where

$$c(n) = n! \sum_{h=1}^{\infty} \frac{\big(n(h-1)\big)!}{(n\,h)!}$$

converges exponentially fast to 1 with increasing n.

2.2 Stochastic Hill-Climber

We consider a stochastic hill-climber performing a Monte-Carlo, Metropolis dynamics [9]. That is the 'walker' randomly selects a step direction and takes the step if it increases its fitness. If the step decreases the walker's fitness by an amount ΔF (< 0) then the walker makes the move with a probability $\exp(\beta\,\Delta F)$, otherwise it stays put. The parameter β (the inverse temperature) controls the likelihood of making a move that increases the fitness. At zero temperature ($\beta = \infty$) the walker never makes a move that decreases the fitness, while at infinite temperature ($\beta = 0$) the walker performs a random walk.

We consider the case when each step corresponds to a single mutations. If we were to allow multiple mutations then the optimum temperature is zero and the walker would be identical to a hill-climber. For single mutations a zero temperature dynamics would lead to the hill-climber becoming stuck. Consequently, for a stochastic hill-climber performing single mutations there is some optimum temperature at which to do the search. As with the hill-climber, we can describe a stochastic hill-climber in terms of a Markov chain. For a fixed set of parameters we can find the mean and median first-passage time exactly. However, we will derive an estimate for the optimum first-passage time in the limit of large problems ($L \to \infty$). Once again, we first consider the $n = 2$ problem and then extend the results to $n = 3$.

As for the hill-climber, the first-passage time will be dominated by the transition from the state $H = 2$ to $H = 0$. The probability to go from $H = 2$ to $H = 1$ is $2\mathrm{e}^{-\beta/2}/L$ and the probability to go from $H = 1$ to $H = 0$ is $1/L$, thus the probability to make the final transition from the local optimum at $H = 2$ to the global optimum is $2\mathrm{e}^{-\beta/2}/L^2$. The probability of making this transition increases as we increase the temperature. However, in contrast to the simple hill-climber, in the stochastic hill-climber the walker will not remain in the states $H \leq 2$, but can move backwards and forwards between local optima. The lower the temperature the more time it will spend in the state $H = 2$ so the more opportunities it has to make the final transition. It is this balance between staying in a fit state (small H) and overcoming the final barrier that leads to an non-trivial optimum temperature.

To estimate this optimum temperature we will assume that the hill-climber has reached a 'quasi-equilibrium state'. That is, the fraction of time it spends in each state with $H \geq 2$ is proportional to the Boltzmann factor $\exp(\beta F)$, where F is the fitness of the site. We call this a *quasi*-equilibrium because we are assuming that we have not yet reached the global optimum. For long times, it is well known that the time spent in each state approaches the Boltzmann distribution. Since the mean time to reach the global optimum, $H = 0$, will typically be substantially longer than the time to reach the local optimum at $H = 2$ (since the transition probability, $2\mathrm{e}^{-\beta/2}/L^2$, is very small) it will be a good approximation to assume quasi-equilibrium. We shall estimate the equilibration time later on.

Assuming a quasi-equilibrium distribution, the proportion of time spent in state $H = 2$ will be given by

$$\mathrm{P}\big(H = 2\big) = \binom{L}{2}\frac{\mathrm{e}^{-\beta}}{Z} \approx \frac{L^2\mathrm{e}^{-\beta}}{2\,Z}$$

where Z is a normalisation constant (often referred to as the partition function)

$$Z = \sum_{H=2}^{L}\binom{L}{H}\mathrm{e}^{\beta\,F_2(H)}.$$

The factor $\mathrm{e}^{\beta\,F_2(H)}$ is the Boltzmann factor giving the probability of being in a state with Hamming distance H, while $\binom{L}{H}$ is the number of states with Hamming distance H. We can write the partition function into a sum over even and odd Hamming distances

$$Z = \sum_{h=1}^{L}\binom{L}{2h}\mathrm{e}^{\beta\,F_2(2h)} + \sum_{h=0}^{L}\binom{L}{(2h+1)}\mathrm{e}^{\beta\,F_2(2h+1)}.$$

The sum over odd Hamming distance gives a sub-dominant contribution to the partition function and will be ignored. For H even we have $F_2(2h) = h$. For $H \ll L$ the binomial coefficient can be approximated by, $\binom{L}{H} \approx L^H/H!$. Thus, for large L

$$Z \approx \sum_{h \geq 1} \frac{L^{2h} \mathrm{e}^{-\beta h}}{(2h)!} = \cosh(x) - 1$$

where $x = L\mathrm{e}^{-\beta/2}$. The mean time to make a transition from $H = 2$ to $H = 0$ assuming a quasi-equilibrium is thus

$$T_{\mathrm{fpt}}(x) \sim \frac{1}{\mathrm{P}\big(H = 2\big)\,\mathrm{P}\big(H = 2 \to H = 0\big)} = L^3 \left(\frac{\cosh(x) - 1}{x} \right).$$

The optimum x is found numerically to be at 2.57, which gives an optimal inverse temperature

$$\beta_{\mathrm{opt}} = 2\log(L) - 1.89.$$

Substituting this value into $T_{\mathrm{fpt}}(x)$ we find the optimum first-passage time, assuming quasi-equilibration, is

$$T_{\mathrm{fpt}} \sim 0.328 L^3.$$

To estimate the quasi-equilibration time we assume that quasi-equilibration occurs for each local optima. We assume that the quasi-equilibration time is equal to the mean first-passage time to reach the local optimum. Assuming the same temperature is used this argument gives exactly the same correction as for the hill-climber, namely a factor of $2\log(2)$. Including this correction factor we obtain the estimate

$$T_{\mathrm{fpt}} \sim 0.455 L^3. \tag{4}$$

In figure 2, we show the asymptotic estimates given above and the exact results (see appendix). Also shown in figure 2 is β_{opt} found empirically and the estimate given above.

We can extend the above analysis to the $n = 3$ problem. The probability to be in the state $H = 3$, assuming quasi-equilibrium is given by

$$\mathrm{P}\big(H = 3\big) = \binom{L}{3} \frac{\mathrm{e}^{-\beta}}{Z} \approx \frac{L^3 \mathrm{e}^{-\beta}}{3!\,Z} = \frac{x^3}{3!\,Z}$$

where now $x = L\mathrm{e}^{-\beta/3}$. The normalisation factor (partition function), Z, is given by

$$Z \sim \sum_{h=1} \frac{x^{3h}}{(3h)!} = \frac{\mathrm{e}^x}{3} + \frac{2}{3}\mathrm{e}^{-x^2/2} \cos\left(\frac{\sqrt{3}\,x}{2} \right) - 1.$$

To compute the first-passage time we must also know the transition probability $\mathrm{P}\big(H = 3 \to H = 0\big)$. This is more complicated than in the $n = 2$ problem because

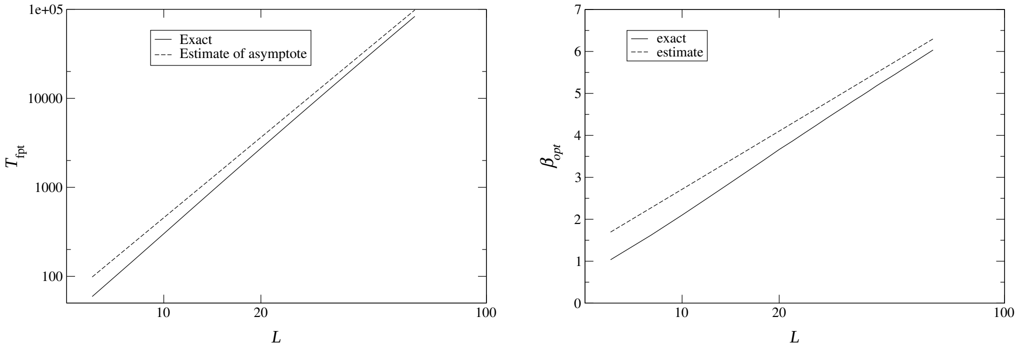

Figure 2: Estimated (dashed line) and exact results (solid line) for (a) the mean first-passage times and (b) β_{opt} versus the problem size L for a stochastic hill climber on the $n = 2$ problem. Both T_{fpt} and L are plotted on a logarithmic scale

the walker may spend some time crossing a hurdle. The probability of moving from $H = 3$ to $H = 2$ is $3\mathrm{e}^{-2\beta/3}/L$. Having reached the state $H = 2$, the walker will either move back to the $H = 3$ state with a probability of $(L-2)/L$, or it will move into the $H = 1$ state with a probability of $2/L$. Once in the $H = 1$ state the walker has a probability of $1/L$ of reaching the global optimum and a probability of $(L-1)\mathrm{e}^{-\beta/3}/L$ of moving back to the $H = 2$ state at any time. Thus the probability of it reaching the global optimum rather than returning to the $H = 2$ state is

$$\frac{1}{1 + (L-1)\mathrm{e}^{-\beta/3}}.$$

We can ignore oscillation between the $H = 2$ and $H = 1$ states as the probability of the transition from $H = 2$ to $H = 1$ is so small. We therefore find to leading order

$$\mathrm{P}\big(H = 3 \to H = 0\big) = \frac{3!\,\mathrm{e}^{-2\beta/3}}{L^2\left(1 + (L-1)\mathrm{e}^{-\beta/3}\right)} \approx \frac{3!\,x^2}{L^4(1+x)}.$$

Using the same argument as above we find the optimum inverse temperature is given by $\beta_{\mathrm{opt}} = 3\log(L) - 4.093$ and the first-passage time assuming a quasi-equilibrium state is $T_{\mathrm{fpt}} = 0.0835L^4$. Using the same estimate for the equilibration time, we find this increases the first-passage time by 7.28%, giving an estimated first-passage time

$$T_{\mathrm{fpt}} \sim 0.896L^4. \tag{5}$$

Figure 3 shows the estimated and exact first-passage and optimum inverse temperature versus the problem size, L. Following a similar argument to that given above, we can show that for problems with hurdles of size n the first-passage time will scale as L^{n+1}.

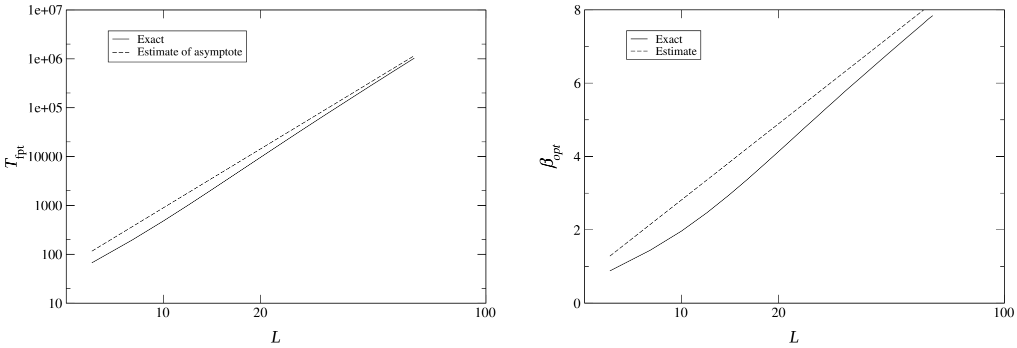

Figure 3: Estimated (dashed line) and exact results (solid line) for (a) the mean first-passage times and (b) β_{opt} versus the problem size L for a stochastic hill climber on the $n = 3$ problem. Both T_{fpt} and L are plotted on a logarithmic scale.

We can speed up the equilibration time by starting from a higher temperature and slowly reducing it as the walker approaches the global optimum. This is the strategy of simulated annealing [10]. However, as the majority of time is spent in overcoming the final hurdle, simulated annealing would only reduce the first-passage time by a small fraction, even if we used an optimal annealing schedule.

2.3 Genetic Algorithms

We conclude by considering a GA acting on this problem [11, 12]. We use a standard, generational GA with a mutation rate of $1/L$ and using uniform crossover. A Boltzmann selection scheme was used in which member α of the population is assigned a selection probability, p_α, proportional to $\exp(\beta\, F_\alpha/\sqrt{\kappa_2})$, where F_α is the fitness of individual α and κ_2 is the variance in the fitness of the population. The members of the new population are selected using stochastic universal sampling [13], where the probability of selecting an individual is given by p_α. Stochastic universal sampling reduces the deviations from the expected number of individual compared with roulette-wheel (independent) sampling. A population of size 50 and selection strength of $\beta = 2.0$ was used. These parameters values were chosen after preliminary testing. They gave good results, but are not optimal in any sense. The results are reasonably robust to different choices of the parameters. The same set of parameters were used for all sizes of problem and for both $n = 2$ and $n = 3$.

Figure 4 shows the average first-passage time found empirically for the $n = 2$ and $n = 3$ hurdle problem and the ones-counting problem. The results are averaged over 1000 runs. Error bars show the predicted error. The first-passage time is given in terms of the number of function evaluations before the global optimum is reached.

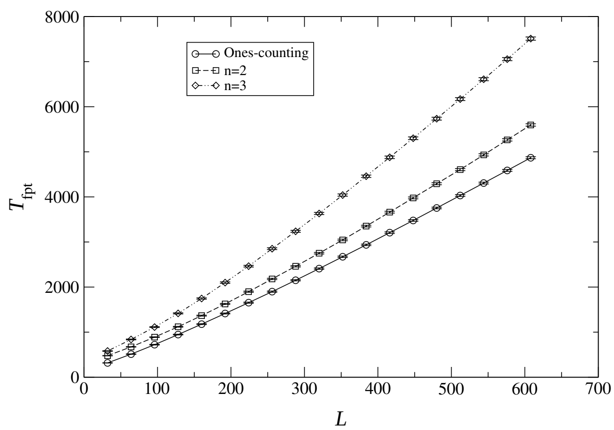

Figure 4: Mean first-passage times versus problem size L for a GA acting on the hurdle problem with $n = 2$ and $n = 3$ and on the ones-counting problem.

All problems are solved in approximately linear time (probably $L \log(L)$). Although a GA finds the hurdle problems slightly harder than the ones-counting problem the difference is only a multiplicative factor. An intuitive argument for this is as follows. In a typical GA solving the ones-counting problem, the spread in the Hamming distance typically covers several local optima. Thus the irregularity caused by the 'hurdles' appear as little more than 'noise' in the fitness evaluation. Indeed, if we consider the noisy ones-counting problem with fitness

$$F(H) = \frac{H}{2} + \eta$$

where η is a Gaussian noise with zero mean and variance $1/4$, then we find that the first-passage time is almost identical to the n=2 case. This is illustrated in figure 5. Although not shown, the dynamics of the noisy ones-counting problem and the $n = 2$ hurdle problem is almost identical throughout the course of the dynamics (that is, it has the same average low order cumulants at each time step). The variation of the fitness in the $n = 2$ model from a ones-counting fitness function (i.e. $F_2(H) - H/2 + 1/2$ is $1/4$, which may give some explanation of this fit. Also in the $n = 3$ problem we can get very similar dynamics and first-passage time by considering a noisy ones-counting problem of the form $F(H) = H/3 + \eta$ where η is again a Gaussian random variable with zero mean and variance 0.36. This, however, is an empirical fit—there is no obvious justification for the size of this variance.

The variation in the first-passage time between runs is much smaller for the GA than either of the previous two methods. The mean and median times being

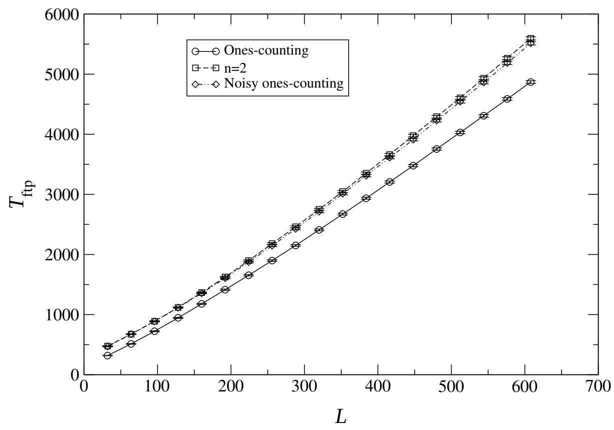

Figure 5: Mean first-passage times versus problem size L for a GA acting on: the hurdle problem with $n = 2$; the ones-counting problem; and the noisy ones-counting problem.

very similar. This is because the first-passage time for the GA is not dominated by the final transition, but has significant contribution from each stage in the dynamics. This is a similar situation to the first-passage time for ones-counting with a hill-climber.

The rapid success of the GA in finding the global optimum is mainly due to crossover. The performance of a GA without crossover scales similarly to that for a hill-walker, although it is still faster than the hill-climber. Figure 6 shows the performance of a GA with and without crossover and also the hill-climber with mutation rate $U = 2$. The mutation rate for the GA both with and without crossover was $U = 1$, which was found to better than $U = 2$. This is because the 'quasi-equilibrium' population is closer to the global optimum for the lower mutation rate. The GA overcomes the final hurdle faster than a hill-walker because it can occupy the $H = 1$ state (although this will be very suppressed), rather than having to make a double jump. In this sense it works similarly to a stochastic hill-climber. However, the stochastic hill-climber performed much worse than the hill-climber because the temperature had to be set low enough that the walker would occupy the local optimum at $H = 2$ with a reasonable probability. The selection strength in a GA plays a similar role to the temperature in the stochastic hill-climber. The effect of the population in the GA is to ensure that the local optimum at $H = 2$ is well occupied without having to make the selection rate too strong. It is thus able to explore the local environment much more easily than the stochastic hill-climber.

It is interesting to consider why crossover is such a successful search operator on this type of problem. Notice that, on average, crossover does not change the Hamming distance. That is, if we were to cross two individuals with Hamming

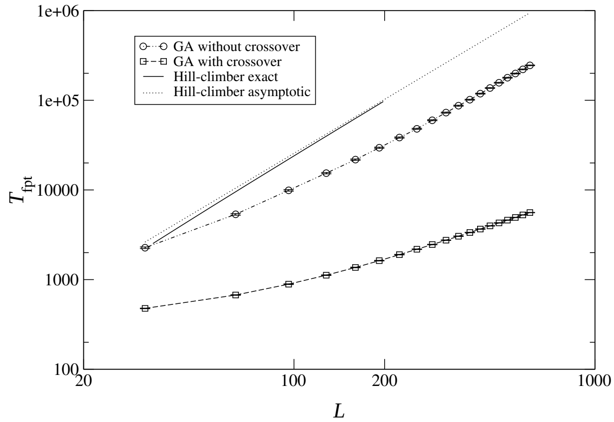

Figure 6: The mean first-passage time is shown versus the problem size, L, for the hurdle problem $n = 2$. The results are for a GA with and without crossover and for a hill-climber with mutation rate $U = 2$. Both axes are plotted on a logarithmic scale.

distance H and H' then, on average, the child would have a Hamming distance $(H + H')/2$. This contrasts strongly with the situation for mutation close to the global optimum. If we mutate a single site the probability of moving towards the global optimum, $H \to H-1$, is H/L while the probability of the mutation moving away from the global optimum is $(L - H)/L$, thus nearly all mutations increase the Hamming distance when $H \ll L$. In the $n = 2$ problem, the probability of moving from the local minimum at $H = 2$ to the global optimum is $2/L(L - 1)$, while the probability of doing this by crossing two strings with 1's at different positions is $1/16$—this is independent of the length of the strings! We can view a GA with crossover as focusing the search since in crossover since we only change the strings on sites where the parent strings take different values.

In a real GA the strings are usually not independent as they share common ancestors. If we were to cross two strings with some 1's at the same site we could never reach the global optimum. To prevent the population from becoming too correlated we allow some mutation. The situation described here is close to that of the *basin with a barrier* problem [5]. In this problem it was shown that sole role of mutation was in maintaining diversity. As a consequence, the optimum mutation rate depends on the population size [7, 8].

The empirical results shown in figure 6 are only for relatively small systems and do not prove that GAs solve the Hurdle problem in sub-quadratic time when L becomes very large. For very large L, we speculate that a GA even with crossover will take $\Theta(L^n)$ steps to find the global optimum. The reason for this is that as $L \to \infty$ and for H of order 1, almost all mutations are bad. To prevent the best member of the population from decreasing in fitness, it is necessary to include an elitism mechanism that ensures the best member of the population is never lost. The most likely cause for the GA to advance from a local optimum H to another local optimum $H-n$ is by crossing two strings with Hamming distance H, but which differ in at least n variables. However, this requires diversity in the population of fit members. Such diversity will be extremely difficult to maintain for large L and small H, as neutral mutations become extremely rare. The most common neutral mutation will be a double mutation which changes a one to a zero and a zero to a one. This occurs with a probability $2H(L-H)/(L(L-1))$. If $H \sim 1$ this occurs with a probability $O(1/L)$. Unfortunately, even for the case $n = 2$, a single mutation is insufficient to allow crossover to make the jump from H to $H - 2$. However, the population is likely to lose diversity through genetic drift before a second mutation occurs. For a GA to perform in $o(L^n)$ time, not only would it require fitness elitism, it would also require a 'diversity elitism' (i.e. a mechanism to prevent the loss of diversity through genetic drift). Thus, although the empirical evidence seem to suggest that a GA with crossover performs similarly to a GA acting on ones-counting, this is likely only to be true for strings below some critical length.

In this paper, we do not attempt to estimate the first-passage time for the GA. Analysis of the dynamics of similar GAs has been investigated in a number of papers, although none of these give the first-passage time for the GA used here. An analysis of the dynamics of a GA on a similar problem was first performed in [14] and [15]. The effect of noisy fitness evaluation in selection has been studied in [16]. For weak selection, noisy fitness evaluation has the effect of reducing the effective population size. Ones-counting using Boltzmann selection and assuming linkage equilibrium (i.e. plenty of crossover) can be solved exactly in the limit of weak selection and mutation [17]. With strong selection the population is no longer in linkage equilibrium. The treatment of correlations out of linkage equilibrium was first performed in [18, 19]. Strong selection has been studied in the context of tournament selection [7, 20, 21]. An analysis of stochastic universal sampling is given in [7]. Estimates for the first-passage time in the *basin with a barrier* problem has been given in [7, 8]. The role of fluctuations in determining the dynamics has been studied in [22]. The effect of using different types of crossover have been investigated in [23, 24]. Analysis of some of the differences between GAs with and without crossover on the ones-counting problem was studied in [25], although this is for weak selection. Finally, the effect of using a steady-state rather than a generational GA has been studied in [26, 27].

3 Conclusion

In this paper we have introduced a new problem in which a GA outperforms conventional optimisation techniques. The GA is able to see the global fitness gradient which is hidden by the local 'roughness' of the fitness function. Although, a GA without crossover appears to slightly outperforms a hill-climber, the GA with crossover substantially outperforms the other heuristics, at least for L not too large. The key to the GA's success is that crossover is focusing the search to parts of the string where members of the population disagree.

The question arises about what would happen on more realistic problems. Unfortunately, in many problems there is no unambiguous gradient to follow. This can be due to symmetries in the problem. A particularly interesting case is the graph colouring problem (where we try to colour a graph with K colours so that the vertices with a common edge are assigned different colours). Here there is a permutation symmetry of size $K!$. There are thus $K!$ different gradients for the members of the population to follow. Traditional GAs have performed extremely poorly on this problem. However, the permutation symmetry is an artifact of the representation of the problem. The problem is more properly a partitioning problem where there should be no edges among the vertices in the same partition. Recognising this, Galinier and Hao have constructed a crossover which ignores the colouring altogether [28]. Using this crossover in combination with a powerful mutation operator (such as Tabu search or a fast hill-climber) a GA has outperformed all other heuristics on a set of standard problems which has been competitively tested by a large number of authors [28,29]. One possible explanation of this good performance is that the GA manages to follow a general gradient, without being confused by local optima.

Unfortunately explicit symmetries are not the only types of symmetries that can appear. In many classical optimisation problems, it is known that the problems become hard at a 'replica-symmetry breaking' phase transition [30,31]. This occurs when a control parameter of the problem is adjusted so that instead of there being one, or possibly a few, global optima, there are an exponential number of global optima widely separated from each other. There is, therefore, no longer an unambiguous gradient for a GA to follow. This problem typically occurs in NP-hard optimisation problems where we known that no search method is likely to find the global optimum. However, to understand more fully how GAs perform on real world problems, we need to study their evolution in these more complex types of fitness landscape. An attempt at this has been made in [32].

Acknowledgements

I would like to thank Ingo Wegener for useful discussions and particular for pointing out the difficulty of maintaining sufficient diversity for crossover to be useful as L becomes very large.

Appendix: Markov Chains

Transition Matrices

We can view the dynamics of a hill-climber and stochastic hill-climber as a Markov chain with transitions between states $H = 0, 1, \ldots, L$. The dynamics can be described in terms of a transition matrix $\mathbf{W}$ with elements W_{ij} equal to the probability of transition from a state with Hamming distance $H = j$ to a state with Hamming distance $H = i$. The matrix $\mathbf{W}$ is a stochastic matrix (all the elements are positive and the sum of all columns equals 1) with absolute eigenvalues $|\lambda_i| \leq 1$. At least one eigenvalue is equal to 1. Provided that the matrix is ergodic (which is the case in the problems we consider) there is a unique eigenvector with eigenvalue 1. Let $\boldsymbol{p}(t)$ be the probability vector whose components $p_i(t)$ denotes the probability of being in state with Hamming distance $H = i$ at time t. By definition $\mathbf{1}^\top \boldsymbol{p}(t) = 1$, where $\mathbf{1}$ is the vector of all ones. The probability vector evolves according to

$$\boldsymbol{p}(t) = \mathbf{W}\, \boldsymbol{p}(t-1) = \mathbf{W}^t\, \boldsymbol{p}(0)$$

where $\boldsymbol{p}(0)$ is the initial probability distribution.

Mean First-Passage Time

To obtain the first-passage time it is useful to consider the modified transition matrix $\hat{\mathbf{W}}$ with elements $\hat{W}_{ij} = W_{ij}$ for $j > 0$ and $\hat{W}_{i0} = 0$. $\hat{\mathbf{W}}$ describes a Markov chain with an absorbing boundary at the global optimum $H = 0$. Probability is no longer conserved in this chain and consequently all the eigenvalues of $\hat{\mathbf{W}}$ have absolute value strictly less than one. Let $\hat{\boldsymbol{p}}(t) = \hat{\mathbf{W}}\, \boldsymbol{p}(0)$, then $\hat{p}_0(t)$ is the probability of reaching the global optimum for the first time at time t and $\mathbf{1}^T \hat{\boldsymbol{p}}(t)$ is the probability of not reaching the global optimum before time t. The mean first-passage time is equal to

$$T_{\mathrm{fpt}} = \sum_{t \geq 0} t\, \hat{p}_0(t) = \sum_{t \geq 0} t\, \boldsymbol{\delta}_0^\top\, \hat{\mathbf{W}}^t\, \boldsymbol{p}(0)$$

where $\boldsymbol{\delta}_0$ is equal to the vector $(1, 0, 0, \ldots, 0)$. We can obtain a closed form expression for T_{fpt} as follows

$$T_{\mathrm{fpt}} = \boldsymbol{\delta}_0^\top \sum_{t \geq 0} (t+1)\, \hat{\mathbf{W}}^t \boldsymbol{p}(0) - \boldsymbol{\delta}_0^\top \sum_{t \geq 0} \hat{\mathbf{W}}^t \boldsymbol{p}(0).$$

The last term is

$$\boldsymbol{\delta}_0^\top \sum_{t \geq 0} \hat{\mathbf{W}}^t \boldsymbol{p}(0) = \sum_{t \geq 0} \hat{p}_0(t) = 1$$

since we will reach the global optimum eventually. We can sum the geometric series

$$\sum_{t=0}^{N-1} \hat{\mathbf{W}}^t = (\mathbf{I} - \hat{\mathbf{W}})^{-1} (\mathbf{I} - \hat{\mathbf{W}}^N)$$

which holds for any matrix (we can prove this by multiplying both sides on the left by $\mathbf{I} - \hat{\mathbf{W}}$). Since the magnitude of the eigenvalues of $\hat{\mathbf{W}}$ are all less than 1 we find

$$\lim_{N\to\infty} \hat{\mathbf{W}}^N = 0$$

thus

$$\sum_{t\geq 0} \hat{\mathbf{W}}^t = (\mathbf{I} - \hat{\mathbf{W}})^{-1}$$

and

$$\boldsymbol{\delta}_0^\top (\mathbf{I} - \hat{\mathbf{W}})^{-1} \boldsymbol{p}(0) = 1.$$

Since this holds for any vector $\boldsymbol{p}(0)$ for which $\mathbf{1}^\top \boldsymbol{p}(0) = 1$ we can deduce the identity $\boldsymbol{\delta}_0^\top (\mathbf{I} - \hat{\mathbf{W}})^{-1} = \mathbf{1}^\top$. Returning to the mean first-passage time

$$\begin{aligned}
T_{\mathrm{fpt}} &= \boldsymbol{\delta}_0^\top \sum_{t\geq 0} (t+1)\, \hat{\mathbf{W}}^t \boldsymbol{p}(0) - 1 \\
&= \boldsymbol{\delta}_0^\top \left(\frac{\partial}{\partial \hat{\mathbf{W}}} \sum_{t\geq 0} \hat{\mathbf{W}}^{t+1} \right) \boldsymbol{p}(0) - 1 \\
&= \boldsymbol{\delta}_0^\top \left(\frac{\partial}{\partial \hat{\mathbf{W}}} (\mathbf{I} - \hat{\mathbf{W}})^{-1} \right) \boldsymbol{p}(0) - 1 \\
&= \boldsymbol{\delta}_0^\top (\mathbf{I} - \hat{\mathbf{W}})^{-1} (\mathbf{I} - \hat{\mathbf{W}})^{-1} \boldsymbol{p}(0) - 1 \\
&= \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}})^{-1} \boldsymbol{p}(0) - 1.
\end{aligned}$$

Thus we can compute the mean first-passage time by inverting the matrix $\mathbf{I} - \hat{\mathbf{W}}$.

Median First-Passage Time

The median first-passage time $\tilde{T}_{\mathrm{fpt}}$ occurs when

$$\mathbf{1}^\top \tilde{\boldsymbol{p}}(\tilde{T}_{\mathrm{fpt}} - 1) > 1/2 \qquad \text{and} \qquad \mathbf{1}^\top \tilde{\boldsymbol{p}}(\tilde{T}_{\mathrm{fpt}}) < 1/2.$$

That is, when the probability of reaching the global optimum is one half. If we consider time as a continuous variable we can define the median first-passage time as

$$\mathbf{1}^\top \hat{\mathbf{W}}^{\tilde{T}_{\mathrm{fpt}}} \boldsymbol{p}(0) = \frac{1}{2}.$$

The matrix $\hat{\mathbf{W}}^{\tilde{T}_{\mathrm{fpt}}}$ can be decomposed as

$$\hat{\mathbf{W}}^{\tilde{T}_{\mathrm{fpt}}} = \sum_{i=0}^{L} \lambda_i \, \boldsymbol{w}_i \, \boldsymbol{v}_i^\top$$

where $\boldsymbol{w}_i$ is the right-hand eigenvector and $\boldsymbol{v}_i$ is the left-hand eigenvector with eigenvalue λ_i. Then the median first-passage time is given by

$$\sum_{i=0}^{L} \lambda_i^{\tilde{T}_{\text{fpt}}} c_i = \frac{1}{2}$$

where $c_i = (\mathbf{1}^\top \boldsymbol{w}_i)(\boldsymbol{v}_i^\top \boldsymbol{p}(0))$. Assuming λ_0 is the largest eigenvalue then an accurate approximation for the problems we are studying is given by $\tilde{T}_{\text{fpt}} \approx -\log(2c_0)/\log(\lambda_0)$.

Form of Transition Matrix

For the problems discussed in the main text we consider random strings as our starting condition so that $p_H(0) = 2^{-L}\binom{L}{H}$. To calculate the transition matrix for independent mutations on sites, we note that the change in Hamming distance after a mutation will be

$$H \to H' = H - \beta_{u,H} + \beta_{u,L-H}$$

where $u = U/L$ is the mutation rate and $\beta_{x,N}$ denotes a binomial deviate with probability

$$\mathrm{P}\big(\beta_{x,N} = n\big) = B(n|N,x) = \binom{N}{n} x^n (1-x)^{N-n}.$$

The probability of a transition H to H' is given by

$$\begin{aligned}
\mathrm{P}\big(H \to H'\big) &= \sum_{\Delta=\max(0,H-H')}^{\min(H,L-H')} B(\Delta|H,u/L)\, B(H'-H+\Delta|L-H,u/L) \\
&= \sum_{\Delta=\max(0,H-H')}^{\min(H,L-H')} \binom{H}{\Delta} \binom{L-H}{H'-H+\Delta} u^{H'-H+2\Delta} (1-u)^{L-H'+H-2\Delta}.
\end{aligned}$$

Two State Systems

Much of the analysis given the paper involves a transition between two states. We therefore consider the problem of starting in a state 1 and making a transition to state 0, where the probability of making the transition is p. We can calculate the mean and median first-passage time directly from the fact that the probability of make a transition at time t is given by the exponential distribution $p(1-p)^t$. However, we will use the results given above. The modified transition matrix for this problem is

$$\hat{\mathbf{W}} = \begin{pmatrix} 0 & p \\ 0 & (1-p) \end{pmatrix}.$$

The initial state is $\boldsymbol{p}(0) = (0,1)^\top$. The mean first-passage time is

$$\begin{aligned} T_{\rm fpt} &= \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}})^{-1} \boldsymbol{p}(0) - 1 \\ &= \mathbf{1}^\top \begin{pmatrix} 1 & -p \\ 0 & p \end{pmatrix}^{-1} \boldsymbol{p}(0) - 1 \\ &= (1,1) \begin{pmatrix} 1 & 1 \\ 0 & 1/p \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} - 1 = \frac{1}{p}. \end{aligned}$$

The eigenvalues of $\hat{\mathbf{W}}$ are $\lambda_0 = 1 - p$ and $\lambda_1 = 0$. The right and left-handed eigenvectors corresponding to λ_0 are $\boldsymbol{w}_0 = (p/(p-1), 1)^\top$ and $\boldsymbol{v}_0 = (0,1)^\top$. Thus $c_0 = (\mathbf{1}^\top \boldsymbol{w}_0)(\boldsymbol{v}_0^\top \boldsymbol{p}(0)) = 1/(1-p)$. The median first-passage time is thus

$$\tilde{T}_{\rm fpt} = \frac{-\log\big(2/(1-p)\big)}{\log(1-p)} = \frac{-\log(2)}{\log(1-p)} - 1.$$

For small p we find $\tilde{T}_{\rm fpt} = \log(2)/p = \log(2)\, T_{\rm fpt}$.

References

[1] M. Mitchell, J. Holland, and S. Forrest. When will a genetic algorithm outperform hill climbing? In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, pages 51–58, San Francisco, CA., 1994. Morgan Kauffman.

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

[3] Erik van Nimwegen, James P. Crutchfield, and Melanie Mitchell. Statistical dynamics of the royal road genetic algorithm. *Theoretical Computer Science*, 229(1):41–102, 1999.

[4] E. B. Baum, D. Boneh, and C. Garrett. Where genetic algorithms excel. NEC preprint, Expanded version of *On Genetic Algorithms*, COLT '95: Preceedings of the Eigth Annual Conference on Computational Learning Theory, Association for Computing Machinery, New York, (1995) pp230–239., 1996.

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

[6] T. Jansen and I. Wegener. On the analysis of evolutionary algorithms — a proof that crossover can really help. In J. Nešetřil, editor, *Proceedings of the 7th Annual European Symposium on Algorithms (ESA'99)*, pages 184–193, Berlin, 1999. Springer.

[7] A. Rogers and A. Prügel-Bennett. The dynamics of a genetic algorithm on a model hard optimization problem. *Complex Systems*, 11(6):437–464, 2000.

[8] A. Rogers and A. Prügel-Bennett. A solvable model of a hard optimization problem. In L. Kallel, B. Naudts, and A. Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing, pages 207–221, Berlin, 2001. Springer.

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations for fast computing machines. *J. Chem. Phys.*, 21:1087–92, 1953.

[10] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

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

[12] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning.* Addison-Wesley (Reading, Mass), 1989.

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

[14] A. Prügel-Bennett and J. L. Shapiro. An analysis of genetic algorithms using statistical mechanics. *Physical Review Letters*, 72(9):1305–1309, 1994.

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

[16] M. Rattray and J. L. Shapiro. Noisy fitness evaluations in genetic algorithms and the dynamics of learning. In R. K. Belew and M. D. Vose, editors, *Foundations of Genetic Algorithms 4*, pages 117–139, San Francisco, 1997. Morgan Kaufmann.

[17] M. Rattray and J. L. Shapiro. Cumulant dynamics of a population under multiplicative selection, mutation and drift. *Journal of Theoretical Population Biology*, 60(1):17–32, 2001.

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

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

[20] A. Prügel-Bennett and A. Rogers. Modelling GA dynamics. In L. Kallel, B. Naudts, and A. Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing, pages 59–85. Springer, Berlin, 1st edition, 2001.

[21] A. Prügel-Bennett. Finite population effects for ranking and tournament selection. *Complex Systems*, 12(2):183–205, 2000.

[22] A. Prügel-Bennett. Modelling evolving populations. *Journal of Theoretical Biology*, 185:81–95, 1997.

[23] A. Prügel-Bennett. Modelling crossover induced linkage in genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 5(4):376–387, 2001.

[24] A. Prügel-Bennett. The mixing rate of different crossover operators. In W. N. Martin and W. M. Spears, editors, *Foundations of Genetic Algorithms 6*, pages 261–274. Morgan Kaufmann, San Francisco, 2001.

[25] A. Prügel-Bennett. On the long string limit. In W. Banzhaf and C. Reeves, editors, *Foundations of Genetic Algorithms 5*, pages 45–56, San Francisco, 1999. Morgan Kaufmann.

[26] A. Rogers and A. Prügel-Bennett. Modelling the dynamics of steady-state genetic algorithms. In W. Banzhaf and C. Reeves, editors, *Foundations of Genetic Algorithms 5*, pages 57–68, San Francisco, 1999. Morgan Kaufmann.

[27] A. Rogers and A. Prügel-Bennett. Genetic drift in genetic algorithm selection schemes. *IEEE Transactions on Evolutionary Computation*, 3(4):298–303, 1999.

[28] P. Galinier and J. K. Hao. Hybrid evolutionary algorithms for graph coloring. *Journal of Combinatorial Optimization*, 3(4):379–397, 1999.

[29] C. A. Glass and A. Prügel-Bennett. Genetic algorithms for graph colouring: Exploration of Galinier and Hao's algorithm. *Journal of Combinatorial Optimization*, 7:229–236, 2003.

[30] H. Horner. Dynamics of learning for the binary perceptron problem. *Zeitschrift für Physik: B*, 86:291–308, 1992.

[31] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky. Determining computational complexity from characteristic 'phase transisition'. *Nature*, 400:133–137, 1999.

[32] A. Prügel-Bennett. Symmetry breaking in population-based optimization. *IEEE Transactions on Evolutionary Computation*, 8(1):63–79, 2004.