Massive Multimodality, Deception, and Genetic Algorithms

David E. Goldberg, Kalyanmoy Deb, and Jeffrey Horn
Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801

IlliGAL Report No. 92005
April 1992

(Final version published in R. Männer, B. Manderick (Ed.s),
Parallel Problem Solving from Nature (PPSN), 2
Amsterdam: North-Holland, 37–46.)

Illinois Genetic Algorithms Laboratory
Department of General Engineering
University of Illinois at Urbana-Champaign
117 Transportation Building
104 South Mathews Avenue
Urbana, IL 61801

Massive Multimodality, Deception, and Genetic Algorithms

David E. Goldberg, Kalyanmoy Deb, & Jeffrey Horn
Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
117 Transportation Building
104 South Mathews Avenue
Urbana, IL 61801, USA

Abstract

This paper considers the use of genetic algorithms (GAs) for the solution of problems that are both average-sense misleading (deceptive) and massively multimodal. An archetypical multimodal-deceptive problem, here called a *bipolar deceptive problem*, is defined and two generalized constructions of such problems are reviewed, one using reflected trap functions and one using low-order Walsh coefficients; sufficient conditions for bipolar deception are also reviewed. The Walsh construction is then used to form a 30-bit, order-six bipolar-deceptive function by concatenating five, six-bit bipolar functions. This test function, with over five million local optima and 32 global optima, poses a difficult challenge to simple and niched GAs alike. Nonetheless, simulations show that a simple GA can reliably find one of the 32 global optima if appropriate signal-to-noise-ratio population sizing is adopted. Simulations also demonstrate that a niched GA can reliably and simultaneously find all 32 global solutions if the population is roughly sized for the expected niche distribution and if the function is appropriately scaled to emphasize global solutions at the expense of suboptimal ones. These results immediately recommend the application of niched GAs using appropriate population sizing and scaling. They also suggest a number of avenues for generalizing the notion of deception.

1 Introduction

Since the introduction (Goldberg, 1987) of the term *deception* to describe problems that are misleading to genetic algorithms (GAs) in an average sense, there has been an increase in activity trying to understand what it is about a problem or a function that causes GAs difficulty, and much of this work has focused on deception. For example, there have been works that have tried to better define or classify deception (Goldberg, 1987, 1989a, 1989b; Liepins & Vose, 1990; Vose & Liepins, 1991; Whitley, 1991), studies that have constructed examples of deceptive functions (Deb & Goldberg, 1991, 1992; Goldberg, 1990; Goldberg, 1989a; Whitley, 1991), and investigations that have altered genetic algorithms to try to solve deceptive problems (Goldberg, Deb, & Korb, 1990; Goldberg, Korb, & Deb, 1989; Eshelman, 1991). Despite these and many other advances, questions have been raised about what deception is about (Mitchell & Forrest, 1991) and whether it is even very meaningful (Grefenstette, 1991). Summarizing their many words somewhat too simply, these researchers seem to be saying there is more to whether GAs have trouble solving a problem than deception, and obviously they are right. But who ever promised us a single, easy answer. Genetic algorithms, despite their apparent simplicity, are highly dimensional, multi-faceted, nonlinear, stochastic complex systems that can interact with problems of infinite variety. It has been recognized (Goldberg, Deb, & Clark, 1991) that there are at least four challenges from the algorithm's side of things (building block supply, growth, mixing, and decision making), and why should we expect to identify a single, simple, silver bullet of a problem class that defeats a GA?

On the other hand, since we do know that GAs work by emphasizing the most highly fit schemata by population sampling and redistribution, thereby tending to increase those schemata at the expense of their lowly competitors, it makes a great deal of sense to worry about those functions where the most highly fit competitors contain strings much different from the best points. This helps to explain why so much time and

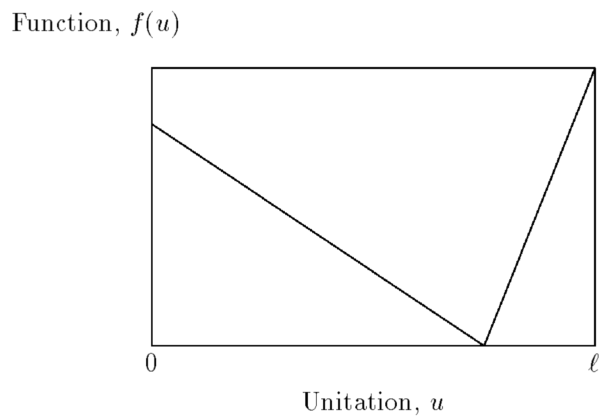

Figure 1: A typical fully deceptive function has two peaks, one at unitation $u = 0$ and one at unitation $u = \ell$.

attention has been lavished on average-sense misleading problems, and in this paper we build on that work by imagining problems that are both deceptive and *massively multimodal*. Starting with De Jong's (1975) work on crowding and continuing with the introduction of sharing functions (Goldberg & Richardson, 1987), there has been growing recognition that GAs can be used to simultaneously and stably find a large number of optima in a multimodal problem. Here we push the envelope along both of these dimensions by constructing and solving problems that are simultaneously massively multimodal and deceptive.

First, *bipolar deceptive functions* with two global optima and with a number of deceptive optima are designed. These bipolar functions are then used as building blocks to create highly deceptive functions containing millions of optima. Interestingly enough simulations of a simple GA on a 30-bit problem created as the sum of five, six-bit bipolar problems shows that the simple GA can reliably converge to one of the problem's 32 global solutions if appropriate signal-to-noise-ratio population sizing is adopted as has been suggested elsewhere (Goldberg, Deb, & Clark, 1991). When sharing-based niching is adopted, the GA is able to simultaneously find stable subpopulations of all 32 global optima if the population is properly sized along niching lines (Deb, 1989; Deb & Goldberg, 1989; Goldberg & Richardson, 1987) and when the function is suitably scaled to emphasize the global solutions at the expense of the local solutions. Not only do these results lend immediate support to the use of niched GAs to solve highly multimodal problems in practice, but the study also gives some hints how to generalize the notion of deception to characterize these and other types of difficult problems.

2 Imagining a massively multimodal and deceptive function

Rik Belew once jokingly remarked to the first author that it requires an evil mind to dream up deceptive, blocked, or otherwise "GA-yucky" functions. We are not sure that ill intent is required, but a certain amount of enthusiastic pessimism is helpful, and it does not hurt to treat your own previous work with a certain amount of (er...) flexibility.

Consider, for example, a typical fully deceptive (average sense) function in figure 1. The variable is the number of ones (or unitation) in the binary string. This function is not unlike that of Liepins and Vose (1990), and the sufficient conditions for average-sense deception have been considered elsewhere for such trap functions (Deb & Goldberg, 1991) and more generally (Deb & Goldberg, 1992). Intuitively, it is fairly easy to see that the global optimum at unitation $u = \ell$ is fairly isolated and that the local optimum at $u = 0$ is surrounded (in a Hamming sense) by highly fit individuals. In fact, these functions are usually designed so that all schemata containing the local optimum are superior to other competitors at level $\ell - 1$ and below, and only at the level ℓ does the global optimum show its stuff.

Now, let's get evil (enthusiastically pessimistic, that is). Suppose we want to have a function that is both

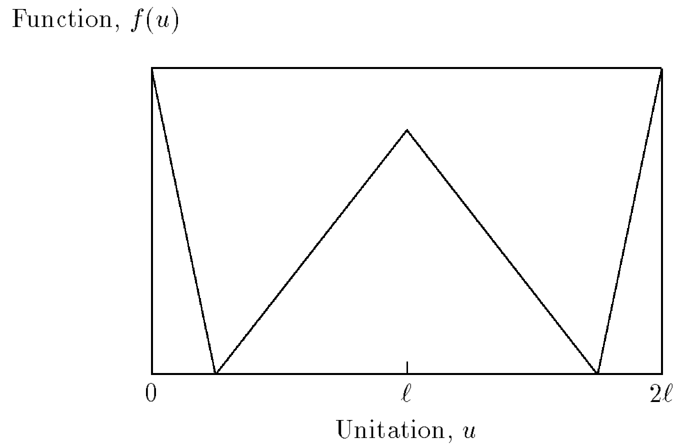

Figure 2: Intuitive construction of a bipolar deceptive function.

deceptive and multimodal. First, we notice that we already have that. Deception implies multimodality, but as far as multimodal functions go, having only two optima seems a little peak poor. One of our problems here is that we have chosen to place the global attractor maximally far from the local deceptive attractor. Moreover, we would really like to have a global solution set that has cardinality greater than one, and furthermore we would like to locate the deceptive attractor where it can have many constituents. To get a function that satisfies these requirements all we have to do is to shift the deceptive attractor to $u = \ell$ and enhance the overall function by reflecting the function about $u = \ell$ as is shown in figure 2. Now we have a function with a global set consisting of two points and a deceptive attractor at the halfway point. If we think about the deceptive attractor a little more, the shift to the midpoint increased its numbers greatly. There are $\binom{2\ell}{\ell}$ points that belong to the deceptive attractor because there are that many ways to get half ones and zeros in a length 2ℓ string. Later on we will use five subfunctions of length-six each giving us a total of $\binom{6}{3} + 2 = 20 + 2 = 22$ optima (of which two are global) per subfunction for a total of $22^5 = 5.15(10^6)$ optima of which $2^5 = 32$ are global. These are big numbers. We are not used to simultaneously searching for 32 needles in a haystack, especially when there are five million other places at which to get stuck.

In the remainder of this section, we review the definition of bipolar deception adopted here, and we review a set of sufficient conditions for deception in such functions. In the next section, we will review the construction of two examples of such functions, one modeled after the bimodal trap functions and the other modeled after deceptive functions constructed from order-limited Walsh coefficients (Goldberg, 1990).

2.1 Defining bipolar deception

In the usual deceptive functions, a schema partition is defined to be deceptive if the schema containing the deceptive optimum is better than any other competing schema in the partition. In a bipolar function, there exist two global optima and a number of deceptive optima. Every schema partition of order less than half the problem size contains a number of schemata that represent global as well as deceptive optima. For example, both schemata `00`$*\ldots*$ and `11`$*\ldots*$ contain global and deceptive optima. However, the schema `10`$*\ldots*$ contains neither global optimum, but it does contain many deceptive optima. Thus, to define deception in these problems, we must modify the usual definition. While a more rigorous definition of generalized deception is being studied by extending the notion of containment probabilistically, in this paper, we simply define a schema partition to be bipolar deceptive if the schema or schemata containing the maximum number of deceptive optima is no worse than other competing schemata. This definition of schema partition deception reduces to the usual definition in the usual case.

In a bipolar function, a schema partition of order λ contains schemata of unitation varying from zero to λ. A schema of unitation u contains $\binom{2\ell-\lambda}{\ell-u}$ deceptive optima. Thus, the schema of unitation $u = \lfloor \lambda/2 \rfloor$ contains the maximum number of deceptive optima. Thus, according to the above definition, a schema partition of order λ is deceptive, if schemata of unitation $u = \lfloor \lambda/2 \rfloor$ are better than or equal to other schemata in the schema partition. With this definition of schema partition deception, we define a *bipolar deceptive function* to be a function where every schema partition is deceptive. We recognize that both schemata in an order-one schema partition contain equal number of deceptive optima and have identical fitness. Thus, in the bipolar function, an order-one schema partition can not be strictly deceptive. Next, we find some sufficient conditions of deception in a bipolar deceptive function of unitation.

2.2 Sufficient condition for deception

A sufficient condition for deception may be found by imposing deception in all schema partitions. In order to do that we define the fitness of a schema of order λ and unitation u as $f(u, \lambda, 2\ell)$. This schema has u ones and $\lambda - u$ zeros in the fixed positions and $2\ell - \lambda$ don't care positions. A string would then be represented as $f(u, 2\ell, 2\ell)$, but we use $f(u)$ to denote the quantity simply. The schema fitness may be expressed in terms of the function value of strings as follows:

$$f(u, \lambda, 2\ell) = \frac{1}{2^{2\ell-\lambda}} \sum_{i=0}^{2\ell-\lambda} \binom{2\ell-\lambda}{i} f(i+u). \tag{1}$$

Since the bipolar function under consideration is symmetric about $u = \ell$, we may write that $f(u) = f(2\ell - u)$. A number of properties of a symmetric bipolar function of unitation are described elsewhere (Deb, 1992). A sufficient condition is derived using those properties and symmetry of the function. In this section, we briefly describe the outcome of that analysis. It has been found that in a symmetric bipolar function of unitation, if an odd-order schema partition is deceptive, the immediately smaller even-order schema partition is also deceptive. This suggests that in order to find deception conditions, only odd-order schema partitions need be considered. Imposing deception in all odd-order schema partitions and writing the schema fitness values in terms of the function values of strings of unitation $0 \leq u \leq \ell$, it can be found that the function is deceptive if following conditions are satisfied:

$$\begin{array}{rl} \textbf{Primary optimality condition:} & f(0) > f(\ell); \\ \textbf{Primary deception condition:} & f(\ell) > f(0) - [f(\ell-1) - f(1)]; \\ \textbf{Secondary deception conditions:} & f(i) \geq f(j) \quad \text{for } \lceil \ell/2 \rceil \leq i \leq \ell - 1 \text{ and } \ell - i \leq j < i. \end{array} \tag{2}$$

In another study (Deb & Goldberg, 1992), similar sufficient conditions were found for a uniglobal, fully deceptive function of size ℓ with the global and deceptive attractors being strings of unitation zero and ℓ respectively. This connection offers a convenient way to construct a bipolar deceptive function. In the following section, we describe two different ways to construct such functions.

3 Construction of bipolar functions

Two different ways to construct a bipolar deceptive function of unitation are described in this section. First, a bipolar function is constructed from a fully deceptive uniglobal function and then a bipolar function is constructed from low-order Walsh coefficients. Both methods satisfy the deception conditions above.

3.1 A folded deceptive function

In the previous section, we saw that a bipolar function may be constructed by shifting and reflecting a uniglobal, fully deceptive function. Defining a parameter *folded unitation*, $e = \mid u - \ell \mid$, we observe that this bipolar function expressed in folded unitation may be identically represented by the uniglobal function of unitation. Elsewhere (Deb, 1992), it is shown that the deception conditions for a uniglobal function are sufficient for deception in a bipolar function constructed from the uniglobal function. We assume that the uniglobal deceptive function is represented by $g(u)$ with unitation u varying from zero to ℓ and that the function has a global optimum at $u = 0$ and a deceptive optimum at $u = \ell$. We then construct a bipolar deceptive function $f(u)$ requiring $f(e) = g(u)$. The bipolar function $f(u)$ has two global optima one at $u = 0$ and one at $u = 2\ell$, and $\binom{2\ell}{\ell}$ deceptive optima at $u = \ell$.

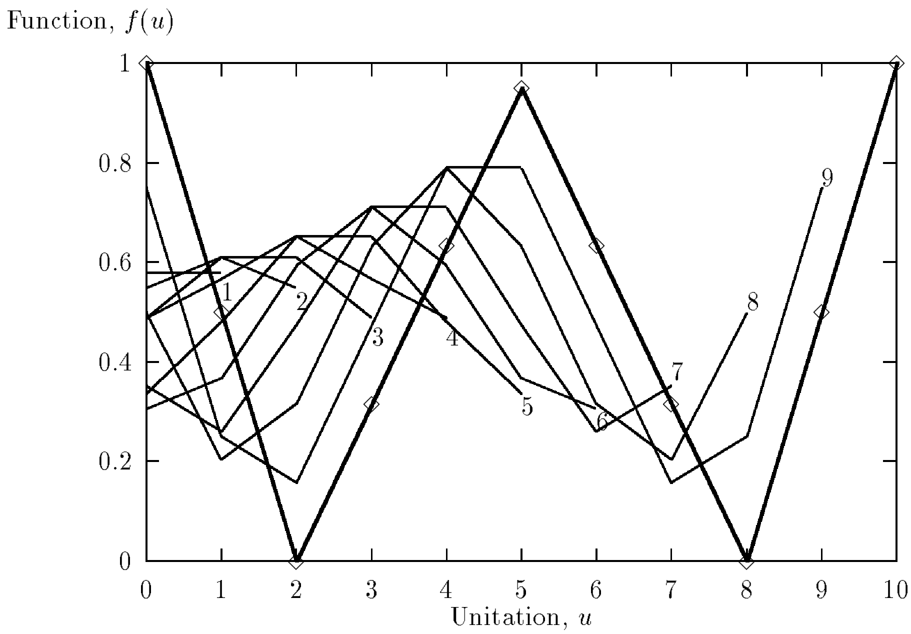

Figure 3: A 10-bit folded-trap function is plotted versus unitation. The schema average function values for schema partitions of order one to nine are also shown.

To illustrate a concrete example, a *folded-trap* function is constructed from a fully deceptive trap function (Deb & Goldberg, 1991). A trap function is a function of unitation with one global optimum and one deceptive optimum located maximally far from the global optimum:

$$g(u) = \begin{cases} \frac{b}{z}(z-u), & \text{if } u \leq z; \\ \frac{a}{\ell - z}(u-z), & \text{otherwise.} \end{cases} \tag{3}$$

In this function, the global optimum is the string with zero unitation (all zeros) and has a function value b and the deceptive optimum is the string with unitation ℓ (all ones) and has a function value a. The function value is zero for strings with unitation z. We construct a 2ℓ-bit, symmetric bipolar function of unitation, $f(u)$ (folded-trap function) by using the above function, $g(u)$. Elsewhere (Deb, & Goldberg, 1992), a sufficient condition for deception of this trap function has been found. According to the above discussion, the same condition apply for the bipolar deception in the folded-trap function. Thus, the sufficient conditions for the 2ℓ-bit, bipolar deceptive function, $f(u)$, is given by the following:

$$\frac{a}{b} \geq \frac{2 - 1/z}{2 - 1/(\ell - z)}. \tag{4}$$

Figure 3 shows a 10-bit folded-trap function with $a = 0.95$, $b = 1.00$, and $z = 2$. These values satisfy the condition above. Thus the folded-trap function constructed from the trap function, $g(u)$, is deceptive. The average function values of schemata of order one to nine are also shown. The figure depicts that schemata of unitation $\lfloor \lambda/2 \rfloor$ are better than or equal to other competing schemata in any partition. Thus, all schema partitions are deceptive, which makes the function bipolar deceptive.

3.2 Using low-order Walsh coefficients

We construct a bipolar deceptive function of unitation from low-order Walsh coefficients using a procedure similar to that used elsewhere to construct a uniglobal deceptive function (Goldberg, 1990). Constructing a function from Walsh coefficients makes it easier to calculate schema fitness values, thereby simplifying the deception analysis. A detailed analysis is described elsewhere (Deb, 1992). In this subsection, we briefly describe the steps used to construct a bipolar deceptive function from low-order Walsh coefficients and find the relations among those Walsh coefficients in order to make the function bipolar deceptive.

We assume that all Walsh coefficients of a particular order are identical. The schema average fitness value of a schema of order λ and unitation u is written as follows (Goldberg, 1990):

$$f(u, \lambda, 2\ell) = \sum_{i=0}^{\lambda} w_i \psi_i'(u, \lambda), \tag{5}$$

where w_i is the ith order Walsh coefficient and the function ψ_i' represents ith order Walsh function multiplied by the total number of ith order Walsh coefficients:

$$\psi_i'(u, \lambda) = \sum_{j=0}^{i} (-1)^j \binom{u}{j} \binom{\lambda - u}{i - j}. \tag{6}$$

The values of a function of size 2ℓ may be obtained from equation 5 by substituting $\lambda = 2\ell$. Since, the bipolar function is symmetric, it can be proved that all odd-order Walsh coefficients are zero. Using order-zero, order-two, and order-four Walsh coefficients only, we construct a symmetric bipolar function of unitation:

$$f(u, \lambda, 2\ell) = w_0 + w_2 \left[\binom{\lambda-u}{2} - u(\lambda - u) + \binom{u}{2}\right] + w_4 \left[\binom{\lambda-u}{4} - u\binom{\lambda-u}{3} + \binom{u}{2}\binom{\lambda-u}{2} - \binom{u}{3}(\lambda - u) + \binom{u}{4}\right]. \tag{7}$$

In order to construct a bipolar deceptive function, the above function must satisfy ℓ optimality conditions: $f(0) > f(u)$ for $1 \le u \le \ell$, and a number of deception conditions: $f(k, 2k+1, 2\ell) > f(u, 2k+1, 2\ell)$ for $1 \le k \le \ell - 1$ and $0 \le u \le k - 1$. Using these conditions and finding the critical condition in each case, we obtain the following conditions for bipolar deception in a function of unitation:

$$-(\ell - 1)(\ell - 2)/3 < w_2/w_4 < -(\ell - 2)^2/3. \tag{8}$$

In the above condition, w_4 is positive. In figure 4, we have plotted a 10-bit bipolar function of unitation. In addition to showing the function values, fitness averages for all schema partitions of order one to nine are shown. For a 10-bit bipolar deceptive function, the lower and upper limit of the ratio w_2/w_4 are -4 and -3 respectively. The function is constructed with $w_0 = 0.4350960$, $w_2 = -0.0248397$, and $w_4 = 0.0080128$, so that the ratio w_2/w_4 is -3.1 and all function values are non-negative. The figure depicts that schemata of unitation $u = \lfloor \lambda/2 \rfloor$ are better than other schemata in a schema partition of order λ. However, both schemata in the order-one schema partition have the same fitness value.

It has been shown elsewhere (Deb, 1992) that the above condition may also be obtained from the sufficient conditions for bipolar deception found in the previous section. In the following, we construct a multimodal deceptive function from bipolar functions and apply genetic algorithms to solve the function.

4 Simulation results

In this section, we report some simulation results using a simple genetic algorithm and a niched genetic algorithm on a massively multimodal deceptive problem constructed using one of the functions of the last section.

A 30-bit problem is constructed by concatenating five identical, six-bit subfunctions. Each subfunction is a six-bit bipolar deceptive function constructed from low-order Walsh coefficients ($w_0 = 0.40036, w_2 = -0.020048, w_4 = 0.060024$). Each subfunction therefore has two global optima and $\binom{6}{3} = 20$ deceptive optima. Thus, the overall 30-bit function has $(20 + 2)^5 = 5,153,632$ optima, of which only $2^5 = 32$ are global. Moreover, since the subfunctions are deceptive, low-order schemata lead the search towards the optima with lowest fitness. Thus, apparently the problem is difficult to solve.

Below we present the results of several GA runs on this problem. In the first set of runs, a simple GA successfully converged to one of the 32 global optima when the population was sized correctly. Using a GA with niching to find all 32 global optima simultaneously proved more difficult, and has led to some interesting results. To concentrate the analysis on deception, we used a tight coding of these subfunctions. To keep

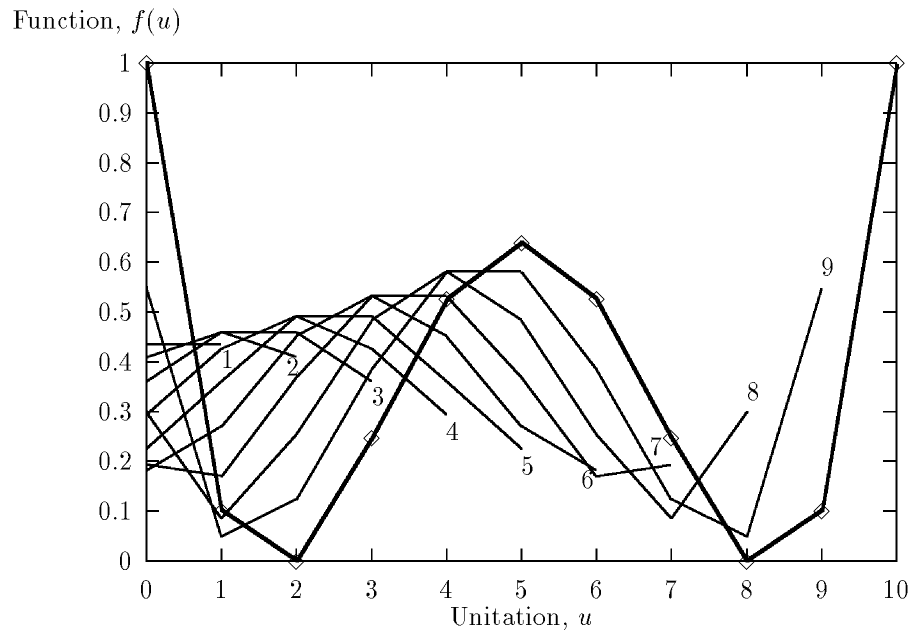

Figure 4: A 10-bit symmetric bipolar function is plotted with unitation. This function is constructed with $w_0 = 0.4350960$, $w_2 = -0.0248397$, and $w_4 = 0.0080128$. All schema fitness values in schema partitions of order one to nine are shown.

mixing high, we used a single-point crossover with a crossover rate of 0.9. A mutation rate of zero was chosen to prevent the restoration of lost diversity. Selection was accomplished by binary tournament without replacement using nonoverlapping populations. The population size was varied as described below.

4.1 Simple GAs without niching

A simple GA is able to overcome deception and converge to one of the 32 global optima if the population is of sufficient size. For the 30-bit problem described above, the population-sizing equation recommended in (Goldberg, Deb, & Clark, 1991) yields a conservative lower bound of about 391 individuals. This bound is easy to calculate, since we know all function values. We can calculate the exact value of the fitness variance of a subfunction: $\sigma^2 = 0.0600721$. The signal to be detected is the difference between the subfunction's global maximum and deceptive optima, thus $d = 1.0 - 0.640576 = 0.359424$, the number of subfunctions is $m = 5$, and the length of each subfunction is $k = 6$. Using the sizing equation: $n = 2c(m-1)2^k\sigma^2/d^2$, we obtain $n = 238c$. Sizing the population for $\zeta = 90\%$ confidence limit (c = 1.64354), we obtain a population size $n = 391.3$.

Figure 5 shows convergence results for different population sizes. Convergence is measured here as the percentage of subfunctions globally optimized by the GA in the final population. Note that both `111111` and `000000` are optimal substrings for each subfunction. The same set of ten, randomly generated, initial populations was used for all population sizes plotted. A particular run was determined to have converged if the difference between the maximum and average fitness values in the current generation was negligible. For the population sizes shown here, convergence usually occurred in less than 50 generations.

As explained elsewhere (Goldberg, Deb, & Clark, 1991), with a confidence factor of ζ, the expected lower bound (marked "Expected LB" in figure 5) of the convergence is also ζ. Thus, the simulation results support the assertion that the population sizing estimate is conservative. As figure 5 shows, the simple GA is able to overcome deception and converge to a global optimum if the population is large enough. However, for population sizes considerably less than the conservative lower bound of 391, the GA was sometimes deceived.

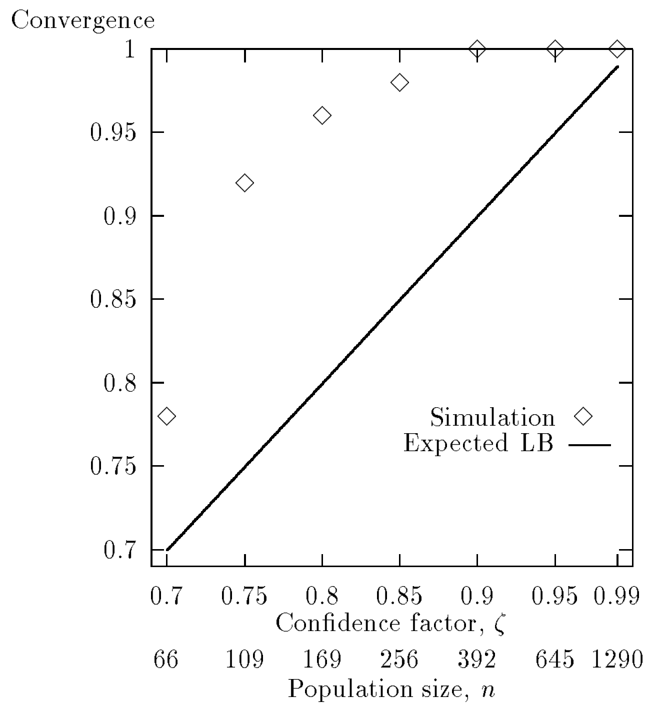

Figure 5: Simulation results for a simple GA on the 30-bit bipolar deceptive problem are shown on a graph of convergence measured by the percentage of subfunctions optimized versus confidence and population size.

When mistakes occurred, one or more of the subfunctions converged to solutions with three `1`'s and three `0`'s. These runs indicate that the problem is indeed deceptive. Even a population of 256 individuals is often lead away from all 32 global optima, converging to one of the five million deceptive optima. But with proper, conservative population sizing, good decision-making is promoted, enabling the simple GA to find one of the global optima with high probability.

4.2 GAs with niching

Although it is interesting and useful to see how difficult it can be for a simple GA to find one of many global optima, a more challenging goal is to find all of the global optima, or as many as possible. In this section, we describe three different approaches to using "multimodal GAs" to find all global optima:

- conventional niche sizing,
- large, overlapping niches, and
- fitness scaling.

We explain the various limitations of each approach to the bipolar problem, and in so doing illustrate the difficulty of this particular multimodal deceptive function.

All three multimodal GA variations implement niching through the use of fitness sharing functions (Goldberg & Richardson, 1987). In a niched GA, an individual's objective function value is degraded by the presence of nearby individuals. Thus this type of sharing requires a distance metric on the phenotype or genotype of the individuals. In this study, we use the Hamming distance between the binary encodings of individuals. If two individuals, i and j, are separated by Hamming distance d_{ij}, and $d_{ij} \leq \sigma_{share}$, then we add a share value, $sh(d_{ij}) = 1 - (\frac{d_{ij}}{\sigma_{share}})^{\alpha}$ to both of their niche counts, m_i and m_j. Here, σ_{share} is the radius of our estimated niches. Individuals separated by more than σ_{share} do not degrade each other's fitness (that is, $sh(d_{ij}) = 0$ for

Table 1: Quality and Quantity of Optima in 30-bit Bipolar Problem

Number of subfunctions maximized i	5	4	3	2	1	0
Fitness f_i	5.00	4.641	4.281	3.922	3.562	3.203
Number of peaks μ_i	32	1600	32,000	320,000	1,600,000	3,200,000

$d_{ij} > \sigma_{share}$). Commonly, $\alpha = 1$, yielding the *triangular sharing function*. For an individual i, the niche count m_i is calculated as $m_i = \sum_{j=1}^{n} sh(d_{ij})$, where n is the size of the population. The shared fitness of individual i is then given by f_i/m_i.

Using shared fitness during selection discourages genetic drift, and sharing tends to spread the population out over multiple peaks (niches) in proportion to the height of the peaks. GAs with proportionate selection and fitness sharing have been successfully used in solving a variety of multimodal functions (Deb, 1989).

Elsewhere (Oei, Goldberg, & Chang, 1991), it has been shown that the naive implementation of fitness sharing does not work well with tournament selection. That study also suggests a way to implement niching where tournaments are played with two random individuals using their shared fitness as calculated in a continuously updated new population. We have implemented this niching technique and have used it in the remainder of this section.

4.2.1 Conventional niche sizing and deceptive niches

In the 30-bit multimodal problem, using current methods of estimating σ_{share}, the niched GA has difficulty maintaining even a single copy of *any* optimum. Instead, the GA searches the many deceptive niches. To understand why this is the case, we look at the quantity and quality of niches of size σ_{share} in the bipolar deceptive problem. Table 1 lists the six different types of local optima in the 30-bit problem.

We can use the information in table 1 to calculate an approximate lower bound on the population size for a niched GA as follows. In a problem with k identical, deceptive subfunctions, the set of optima can be partitioned according to the number of subfunctions maximized, resulting in $k+1$ partitions. Let the index i indicate the number of subfunctions maximized. Thus, partition k has all the subfunctions maximized, corresponding to the global optima. Partition 0 has no subfunctions maximized and corresponds to the set of minimally valued optima. Table 1 shows such a partitioning for the 30-bit problem (with $k = 5$), including the fitness and cardinality of each partition.

Suppose we would like the GA to converge to θ individuals on each of the global peaks. A niched GA will "converge" when it reaches a steady state in which the shared fitness values for all individuals are equal (Goldberg & Richardson, 1987). Let's suppose that at steady state, the average niche count for optima in partition i is m_i. Then the expected shared fitness value for an individual at an optima of type i is $\frac{f_i}{m_i}$. At steady state, $\frac{f_i}{m_i} = \frac{f_j}{m_j}$, for all i and j.

If we assume, as a first approximation, that all optima are separated from each other by a distance greater than σ_{share}, then each optima can be the center of its own niche, with no overlapping of niches. Then the niche counts m_i are equal to the expected number of individuals at each optima in partition i. Therefore, $\sum_{i=0}^{k} m_i \mu_i = n$, where n is the population size. Rearranging, we derive:

$$n = \frac{m_k}{f_k} \sum_{i=0}^{k} f_i \mu_i, \qquad (9)$$

where m_k is the expected number of individuals at each global optimum. Then to calculate the approximate lower bound on the population size required to have θ individuals at each global peak, we simply substitute θ for m_k above. For our 30-bit problem, substituting μ_i and f_i from table 1 into equation 9, we obtain

$n = 3,468,240m_5$. If we want a minimum of one copy of each global solution in the steady state population, we should use a population size of at least 3,468,240. Such a large population size is a significant fraction of the total search space: $3,468,240 \times 2^{-30} \approx 0.003$. Thus, even a run of a hundred generations means that the total computational cost will be of the same order as an enumerative search of the space. Also, such a population size is too large for most serial implementations of GAs. Several runs with populations much smaller than $3,500,000$ yielded no individuals at any of the global peaks.

For such a price we do get a proportionate distribution of individuals, indicating the relative merits of many non-global peaks. But if we are interested only in the global optima, then we are wasting valuable population real estate on uninteresting solutions.

4.2.2 Large niche size

Conventionally, we estimate σ_{share} so that at least every global peak in the space can have its own niche. For the bipolar problem, this means a $\sigma_{share} < 6$, since global optima differ from each other by as little as six bits. However, the largest set of deceptive optima (which have *no* subfunctions optimized), lie at a Hamming distance of exactly $\ell/2 = 30/2 = 15$ bits away from each global optimum. There are 3,200,000 such deceptive optima, accounting for about 62% of the total optima. With so many suboptima outside the niches of the global optima, most of the population will be absorbed by niches forming around such optima. Thus, conventional σ_{share}-sizing, without appropriate fixed-point sizing, doesn't seem promising.

Another approach to this difficulty might be to increase the niche size (σ_{share}) so that the 3,200,000 deceptive optima (and all other deceptive optima, consequently), are within niches centered around a global optimum. When a global and a deceptive optima compete within the same niche, the global will have an advantage. A σ_{share} value greater than or equal to $\lfloor \ell/2 \rfloor$ ensures that all deceptive optima lie within at least one global-centered niche.

Runs with $\sigma_{share} = \lfloor \ell/2 \rfloor + 2$ did indeed result in several global optima being found and maintained with significant subpopulations[1]. For a five-subfunction problem (with 32 global optima), four global optima were found and maintained (approximately 2000 copies each), while the other 28 received inconsequential numbers of copies (0 to 200), which fluctuated widely.

Why were only four of 32 globals found? The answer lies in the fact that the niches, centered on the global optima, overlap. Given a niche radius (σ_{share}) of $\lfloor \ell/2 \rfloor$ in the problem above, we can choose at most 4 of the 32 optima such that each is at least $\sigma_{share} + 1$ bits different from any of the other three. Globals within the niche will compete with each other. Genetic drift, and other factors, will cause the niche to converge to one global. Thus, the maximum number of globals we can expect to maintain using a $\sigma_{share} \geq \lfloor \ell/2 \rfloor$ in this problem is four. Indeed, the four optima with over 1000 copies at the end of each of our large-σ_{share} runs are more than σ_{share} bits apart.

This drawback of the large-niche-size approach is a severe one. In general, for such bipolar problems of size ℓ, the majority of deceptive optima will lie at a distance of exactly $\lfloor \ell/2 \rfloor$ bits from each global optimum, while each global optimum will be at most ℓ_s bits (the length of the subfunction) from k other global optima. Thus we can expect never to get more than a small fraction of the total number of global optima using this approach.

Hence this problem is difficult not only because of deception and the quantity of deceptive attractors, but also because of the placement of these attractors among the global optima. Given the spacing of the global optima, the majority of deceptive optima are located at points maximally distant from the set of globals. This makes niche sizing extremely difficult. Forcing deceptive optima to compete with globals while keeping globals from competing with each other requires augmentation of the standard niched algorithm. In the next section, we present one augmentation that succeeds in finding all 32 optima for this problem.

4.2.3 Fitness scaling

So far, the simulations have revealed three ways in which the 30 by 6 bipolar function is difficult. The simple GA runs demonstrated the effect of deceptive schema partitions. The conventional-niche sizing calculations showed how the sheer numbers of optima (niches) can be a problem for a niched GA. Finally, analysis and runs with large niche sizes revealed the importance of *separability* of global and deceptive optima. These three

[1] In order to promote a high degree of sharing out to the edge of the niche, the exponent, α, of the power sharing function was increased to four.

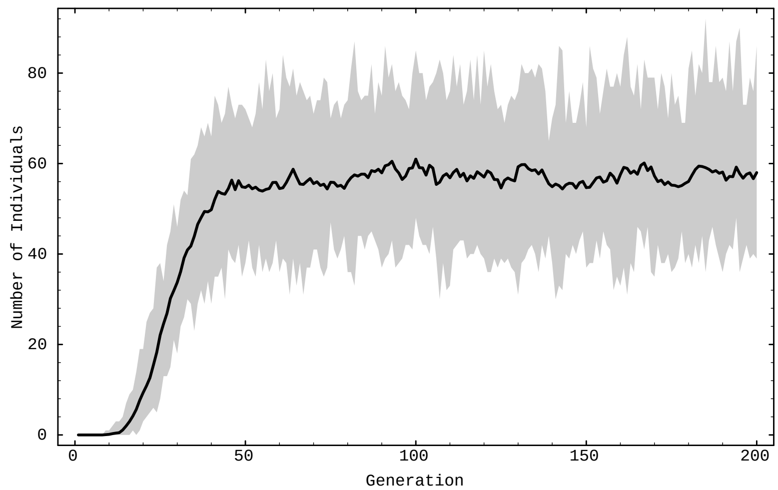

Figure 6: The range of subpopulation sizes for the 32 global optima, for each of 200 generations, is shown by the shaded area. Also plotted as a dark solid line is the mean of these 32 values at each generation.

aspects of multimodal problem difficulty are not unrelated, and it is no coincidence that the 30 by 6 bipolar function seems to maximize all three. But in this section we are concerned with finding some way to overcome these difficulties and find all 32 optima.

Many augmentations of the basic niching algorithm are possible. One simple approach is scaling of the function prior to calculating the shared fitness. By raising the objective function to a suitably large exponent, we increase the capacity of the niches centered on the global optima, relative to the niches centered on deceptive attractors. From the steady state niching equations,

$$n = m_k \sum_{i=0}^{k} \mu_i \left(\frac{f_i}{f_k}\right)^{\beta}.$$

Increasing β, while holding m_k constant, decreases the population size n needed to maintain m_k, the expected number of individuals at each global optima, in steady state.

In particular, for our 30-bit bipolar problem, we achieved substantial, stable subpopulations at each of the global optima by using $\beta = 15$ and a population size of 5000. Several runs were performed using initial populations drawn from the set of initial populations created for the simple GA runs described earlier. Figure 6 presents the results of one such run. Other runs produced very similar results when plotted in this manner. Figure 6 gives some indication of the number of individuals at each of the 32 global peaks at each generation. The shaded area is actually the range of values for the optima every generation. Thus, the upper border of the shaded area is the maximum number of copies at any of the 32 optima for that generation, while the lower border is the minimum such value. The solid line is a graph of the average number of individuals per peak at each generation. The two plots indicate significant variation from peak to peak and generation to generation, but it is also clear that subpopulations are being maintained within a range of about 40 to 80 individuals. This number seems large and stable enough to indicate that all 32 optima have been found and are being maintained in some kind of noisy steady state.

The fitness-scaling approach succeeds by addressing the first two of these difficulties directly and by sidestepping the third. First, scaling reduces the number of deceptive schema partitions. Second, by scaling the fitness values, the expected numbers of the global solutions are enhanced at the expense of the deceptive solutions. With lower expected niche counts, large numbers of deceptive solutions are effectively driven to extinction, making their better separability something of a moot point.

5 Conclusions

This paper has considered the construction and optimization of problems that are average-sense misleading (deceptive) and massively multimodal. Specifically, bipolar deceptive functions have been defined with two global optima and a number of deceptive attractors. Sufficient conditions for deception in a bipolar function have been found, and specific bipolar functions have been constructed by two different methods—using a uniglobal deceptive function and the idea of folded unitation, and using low-order Walsh coefficients.

A 30-bit function has been formed by concatenating five, six-bit bipolar functions together. The function has more than five million optima, of which only 32 are global. A simple GA has been able to find one of the 32 global solutions, if populations are properly sized to detect signal from noise. GAs with sharing have found all 32 global solutions simultaneously when used with fitness scaling and appropriate fixed-point sizing.

In a practical sense, the study opens the door for the everyday solution of difficult, massively multimodal problems in a variety of applications. From a theoretical point of view, the study uncovers avenues along which the concept of deception might be generalized so it can become a more all-encompassing abstraction of what it means for a problem to be hard for a genetic algorithm. From a more elevated perspective, this paper confirms the fundamental good and common sense of intimately and iteratively tying theory and computation to the study of GAs. It may be a reasonable pastime in Algorithms 101 to sit on the sidelines and wring one's hands whether this definition or that is just so, but genetic algorithms (and all other complex systems) are no quicksort. Computation and analysis must continually interact if we are ever to understand GAs well enough to use them reliably across the problem spectrum that awaits our attention. Although asking good questions is part of that process, it seems to us that giving good intuitive answers, and backing them up with the one-two punch of theory and computational experiments is the ticket to the promised land. Even though many researchers have already taken their seats, there still is plenty of room for a good many more, and as the engineer checks to see that the track is clear and the destination is sure, the passengers on the platform must decide whether to answer the conductor's call of "all aboard!"

Acknowledgments

The authors acknowledge the support provided by the US Army under Contract DASG60-90-C-0153 and by the National Science Foundation under Grant ECS-9022007.

References

De Jong, K. A. (1975). An analysis of the behavior of a class of genetic adaptive systems. (Doctoral dissertation, University of Michigan). *Dissertation Abstracts International, 36(*10), 5140B. (University Microfilms No. 76-9381)

Deb, K. (1989). *Genetic algorithms in multimodal function optimization* (MS Thesis and TCGA Report No. 89002). Tuscaloosa: University of Alabama, The Clearinghouse for Genetic Algorithms.

Deb, K. (1992). *Multimodal deceptive functions.* (IlliGAL Report No. 92003). Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Deb, K., & Goldberg, D. E. (1991). *Analyzing deception in trap functions.* (IlliGAL Report No. 91009), Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Deb, K., & Goldberg, D. E. (1992). *Sufficient conditions for deceptive and binary functions.* (IlliGAL Report No. 92001). Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Deb, K., & Goldberg, D. E. (1989). An investigation of niche and species formation in genetic function optimization. *Proceedings of the Third International Conference on Genetic Algorithms*, 42–50.

Eshelman, L. J. (1991). The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. *Foundations of Genetic Algorithms*, 265–283.

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal deceptive problem. In L. Davis (Ed.), *Genetic algorithms and simulated annealing* (pp. 74–88). London: Pitman.

Goldberg, D. E. (1989a). Genetic algorithms and Walsh functions: Part I, a gentle introduction. *Complex Systems, 3*, 129–152.

Goldberg, D. E. (1989b). Genetic algorithms and Walsh functions: Part II, deception and its analysis. *Complex Systems, 3*, 153–171.

Goldberg, D. E. (1990). *Construction of high-order deceptive functions using low-order Walsh coefficients.*, (IlliGAL Report No. 90002). Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Goldberg, D. E., Deb, K., & Clark, J. H. (1991). *Genetic algorithms, noise, and the sizing of populations* (IlliGAL Report No. 91010). Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Goldberg, D. E., Deb, K., & Korb, B. (1990). Messy genetic algorithms revisited: Studies in mixed size and scale. *Complex Systems, 4*, 415–444.

Goldberg, D. E., Korb, B., & Deb, K. (1989). Messy genetic algorithms: Motivation, analysis, and first results. *Complex Systems, 3*, 493–530.

Goldberg, D. E., & Richardson, J. (1987). Genetic algorithms with sharing for multimodal function optimization. *Proceedings of the Second International Conference on Genetic Algorithms*, 41–49.

Grefenstette, J. J. (1991). Building block hypothesis considered harmful. *Genetic Algorithms Digest [email journal]*, Volume *5*(19).

Liepins, G. E., & Vose, M. D. (1990). Representational issues in genetic optimization. *Journal of Experimental and Theoretical Artificial Intelligence, 2*(2), 4–30.

Mitchell, M., & Forrest, S. (1991). *What is deception anyway? And what does it have to do with GAs? Some concerns inspired by the Tanese functions.* Unpublished manuscript.

Oei, C. K., Goldberg, D. E., & Chang, S. (1991). *Tournament selection, niching, and the preservation of diversity.* (IlliGAL Report No. 91011). Urbana: University of Illinois, Illinois Genetic Algorithms Laboratory.

Vose, M. D, & Liepins, G. E. (1991). Schema disruption. *Proceedings of the Fourth International Conference on Genetic Algorithms*, 237–242.

Whitley, D. (1991). Fundamental principles of deception in genetic search. *Foundations of Genetic Algorithms*, 221–241.