Interactions and Dependencies in Estimation of Distribution Algorithms

Roberto Santana, Pedro Larrañaga, and José A. Lozano
Intelligent System Group
Department of Computer Science and Artificial Intelligence
University of the Basque Country
P.O. Box 649, 20080 San Sebastián - Donostia, Spain
rsantana@si.ehu.es,Pedro.Larranaga@ehu.es,lozano@si.ehu.es

Abstract- In this paper, we investigate two issues related to probabilistic modeling in Estimation of Distribution Algorithms (EDAs). First, we analyze the effect of selection in the arousal of probability dependencies in EDAs for random functions. We show that, for these functions, independence relationships not represented by the function structure are likely to appear in the probability model. Second, we propose an approach to approximate probability distributions in EDAs using a subset of the dependencies that exist in the data. An EDA that employs only malign interactions is introduced. Preliminary experiments presented show how the probability approximations based solely on malign interactions, can be applied to EDAs.

1 Introduction

The advent of probabilistic models for evolutionary computation has led to the conception of more efficient optimization algorithms, able to solve some of the limitations exhibited by genetic algorithms (GAs) [1]. Estimation of distribution algorithms (EDAs) [2, 3] replace the traditional crossover and mutation operators used in GAs with the estimation and sampling from probabilistic models. These models are built at each generation. They allow the probability distribution of the selected solutions to be estimated.

The probabilistic model must be able to capture, in the form of statistical dependencies, a number of relevant interactions between the variables. Dependencies are then used to generate solutions whithin a simulation step. The generated solutions are expected to share a number of characteristics with the selected ones. In this way, the search is led to promising areas of the search space. The success of EDAs in the solution of different practical problems has been documented in the literature [2].

The introduction of distributions in the context of evolutionary computation has also determined the need to critically review some of the theoretical tools and concepts traditionally employed to analyze GAs. New phenomena and dynamics need to be explained, and the heritage from GAs is certainly useful but not sufficient. Fortunately, most of the development of probabilistic modeling used by EDAs lay on the solid foundations of graphical models. To provide the answers that arise in the study of EDAs, this theory can be used. Nevertheless, EDAs are a novel application of graphical models. Some of the scenarios that we find in these algorithms (e.g. intensive use of the probabilistic models, critical role of sampling methods, influence of the selection method in the shape of the probability distributions, etc.) have not been the focus of previous studies in graphical models.

An important issue in EDAs is the appropriate choice of the class of probability model to be used for a given function. Two questions related to this issue are: 1) how does the problem structure determine the probability dependencies, and 2) to what extent should the dependencies that exist in the selected set be captured by the probability models?

Recently, Gao and Culberson [4] studied a class of random functions in the context of EDAs. The authors relate the structure of these functions to the interaction graphs and the computational complexity of EDAs. The relationship between the function structure and the probabilistic models learned by EDAs is essential to understand the way EDAs behave. A first aspect treated in this paper is the determination of the conditions under which the function structure and the independence graph coincide, and which factors influence this relationship. Our analysis is limited to the same class of random functions and it focuses on the role of the selection methods used by EDAs.

Another relevant issue is the determination of the class of probability approximations best suited to represent the function structure. Initial research on this subject [5] has revealed that, by grouping interacting variables in the same factors, EDAs are able to avoid the deceptive behavior of GAs that use traditional crossover operators. It has been shown that the chance of converging to the optimal solutions can be improved by using higher order statistics able to capture higher order interactions determined by the objective function [6].

There is evidence that the structure of additive decomposable functions can be used to construct factorizations of the distribution. For infinite populations, it has been proved that a FDA that uses as its probabilistic model the structure of the additive function and proportional selection converges to the optimal solution [7]. However, the grouping of variables has implications for the complexity of the algorithms. An efficient EDA needs a fine balance between the expressiveness of the probability model and its complexity. The efficiency of learning and sampling the probabilistic model can deteriorate if too many variable-dependences are included [5]. In this paper, we investigate to what extent, by disregarding some of the dependencies existing in the data, it is possible to obtain an approximation of the probability distribution of the selected set useful for the search.

From a methodological point of view, we intertwine theoretical analysis with empirical evidence. Experiments are inserted in different sections to illustrate our analysis. The remainder of the paper is organized as follows. In the next

0-7803-9363-5/05/$20.00 ©2005 IEEE.

section, we present EDAs. These algorithms are analyzed in terms of the probability distributions which they determine. In Section 3, the relationship between random functions and the probability dependencies they determine is analyzed for proportional and Boltzmann selection. In Section 4, we introduce as a new alternative to deal with complex probabilistic models an EDA that considers only a subset of the probability dependencies of the data when building the model. In Section 5, the conclusions of our paper and some open questions are discussed.

2 Estimation of Distribution Algorithms

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a vector of discrete random variables. We will use $\mathbf{x} = (x_1, \ldots, x_n)$ to denote an assignment to the variables. S will denote a set of indices in $\{1, \ldots, n\}$, and $\mathbf{X}_S$ (respectively $\mathbf{x}_S$) a subset of the variables of $\mathbf{X}$ (respectively a subset of values of $\mathbf{x}$) determined by the indices in S. We will work with positive probability distributions denoted by $p(\mathbf{x})$. Similarly, $p(\mathbf{x}_S)$ will denote the marginal probability distribution for $\mathbf{X}_S$. We use $p(x_i \mid x_j)$ to denote the conditional probability distribution of X_i given $X_j = x_j$.

An undirected graph $G = (V, E)$ is defined by a nonempty set of vertices V, and a set E of unordered pairs of V called edges. Given a probability distribution $p(\mathbf{x})$, its independence graph is a graph $G = (V, E)$ that associates one vertex with each variable of $\mathbf{X}$, and where two vertices are connected if the corresponding variables are conditionally dependent given the rest of the variables.

The pseudocode of a general EDA approach is shown in Algorithm 1. The selection method employed can be any of those traditionally used by GAs. In the literature, truncation, Boltzmann and tournament selection are commonly used with EDAs. A characteristic feature of an EDA is the type of probabilitic model it uses. Models may differ in the order and number of the probabilistic dependencies that they represent. The algorithms employed to learn and sample the model also determine important differences between EDAs.

Algorithm 1: **Main scheme of the EDA approach**

```
$D_0 \leftarrow$ Generate $M$ individuals (the initial population) randomly
$l = 1$
do {
    $D^s_{l-1} \leftarrow$ Select $N \leq M$ individuals from $D_{l-1}$ according to a selection method
    $p_l(\boldsymbol{x}) = p(\boldsymbol{x}|D^s_{l-1}) \leftarrow$ Estimate the joint probability of selected individuals
    $D_l \leftarrow$ Sample $M$ individuals (the new population) from $p_l(\boldsymbol{x})$
$l \Leftarrow l + 1$
} until A stop criterion is met
```

Regarding the way learning is done in the probability graphical model, EDAs can be divided into two classes [8]. One class groups the algorithms that make only a parametrical learning of the probabilities, and the other one comprises those algorithms where in addition structural learning of the model is done. Structural learning methods can be based on probabilistic independence tests, scoring metrics or a mixture of both [8]. Relevant to our analysis are methods that learn an independence graph using independence tests.

To determine if an edge belongs to the independence graph, it is enough to make an independence test on each pair of variables given the rest. Nevertheless, from an algorithmic point of view, it is important to reduce the cost of the independence tests. Thus, some algorithms attempt to diminish the number and cost of the statistical tests [9]. The idea is to start from a complete undirected graph, and then try to remove edges by testing the conditional independence between the linked nodes, using conditioning sets that are as small as possible.

After the graph has been constructed, different alternatives can be selected to create a factorization of the probability from the graph [10]. If the graph is not chordal, one possibility is the triangulation of the graph, a process where edges are added. From a triangulated graph, a junction tree can be easily constructed, and the junction tree can be used to efficiently sample the probability distribution. The problem is that the complexity of the junction tree sampling algorithm is exponential in the size of its maximum clique. Another parameter used to measure the junction tree complexity is treewidth[1]. The role of these parameters is critical for learning, inference and sampling algorithms. Several studies have analyzed this issue. In EDAs, the question has been recently addressed [4], considering the influence of the treewidth of interactions graphs corresponding to random models on the complexity of EDAs. There are other alternatives to triangulization [10].

2.1 Multivariate probabilities and EDAs

We will work on a theoretical framework where EDAs can be analyzed in terms of the multivariate probabilities determined by their components. A schematic representation that helps to understand the way EDAs generate distributions along the evolution is shown in Figure 1. In this figure, D_l denotes the EDA population at generation l. $p^s_l(\mathbf{x})$ denotes the probability distribution after selection. $p^a_l(\mathbf{x})$ is the factorized probability distribution given by the model chosen to approximate $p^s_l(\mathbf{x})$.

In Figure 1, selection methods have been divided into two classes [11]: (1) Proportional schemes, and (2) Ordinal based schemes. Proportional schemes select an individual based on its relative fitness value compared to others. Ordinal schemes select an individual based on its ranking in the population.

In proportional schemes, selection is usually accomplished in two steps. First, the selection probabilities of the individuals in the current population are determined. Then,

[1] The treewidth of a graph G is the minimum $k \geq 0$ such that G is a subgraph of a triangulated graph H having a maximal clique of size $k + 1$. If G is triangulated, its treewidth is the size of the maximum clique in the graph minus 1.

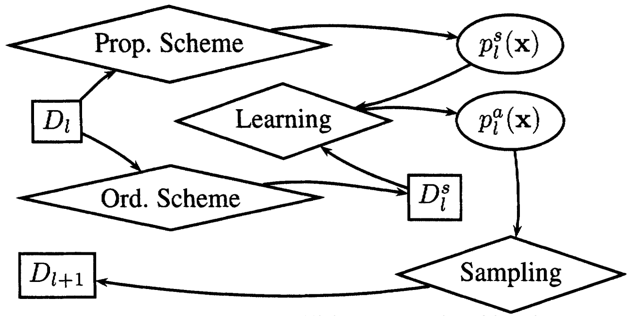

Figure 1: Multivariate probabilities determined by the components of an EDA. D_l, D_{l+1}: populations at generation l and $l+1$; $p_l^s(\mathbf{x})$, $p_l^a(\mathbf{x})$: Joint probabilities determined by selection and the probabilistic model approximations.

new individuals are sampled from these probabilities. These individuals form the selected set that will serve as a mating pool. Examples of these proportional schemes are the proportional and Boltzmann selection.

In ordinal based schemes, selection probabilities are not explicitly calculated. Instead, some procedure is used to select the individuals straight from the population. One example is tournament selection, where s individuals are randomly chosen from the population, and the best individual from this group is included in the selected set[2]. This process is repeated until the selected set has been filled.

The expression $p_l^s(\mathbf{x})$ depends on the type of selection scheme used. For ordinal schemes, it is the probability of a solution to be included in the selected set. For proportional schemes, it can be calculated straight from the fitness function. Assuming that the initial probability distribution of the points in the search space is uniform, the selection probabilities determined by proportional and Boltzmann selection are respectively given by equations (1) and (2).

Definition 1 (Proportional selection) *The probability assigned by Proportional selection to a vector* $\mathbf{x}$ *is defined as:*

$$p_{\mathcal{P}}^s(\mathbf{x}) = \frac{f(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}})} \qquad (1)$$

Definition 2 (Boltzmann selection) *The probability asigned by Boltzmann selection to a vector* $\mathbf{x}$ *is defined as:*

$$p_{\mathcal{B}}^s(\mathbf{x}) = \frac{e^{\frac{f(\mathbf{x})}{T}}}{\sum_{\tilde{\mathbf{x}}} e^{\frac{f(\tilde{\mathbf{x}})}{T}}}, \qquad (2)$$

T is a parameter of the selection called temperature.

In the next section, we will focus on distribution $p_l^s(\mathbf{x})$ determined by proportional schemes.

3 Function definition, problem structure and probabilistic models

We start by analyzing the random models for additive functions used by Gao and Culberson [4]. These authors study the influence of the fitness function on the graphical structure learned by the model, and on the complexity of the EDA. Three essential definitions [4] are: *additive fitness function*, *random model* and *interaction graph of an additive fitness function.*

A fitness function $f : \Omega_{\mathbf{x}} \rightarrow [0, \infty)$, $\Omega_{\mathbf{x}} = \{0,1\}^n$, is *additive* if it can be represented as a sum of lower dimensional functions:

$$f(\mathbf{x}) = \sum_{c \in \mathcal{C}} f_c(\mathbf{x}_c), \mathbf{x} = \{x_1, \ldots, x_n\} \in \Omega_{\mathbf{x}}, \qquad (3)$$

where $\mathcal{C}$ is a collection of subsets of $(x_1, \ldots, x_n)$. For each $c \in \mathcal{C}$, $f_c(\mathbf{x}_c)$ only depends on the variables in c, and is thus called a *local fitness function.* The parameter k of an additive fitness function f is the size of the largest variable set in $\mathcal{C}$. It can be assumed that all $\mathcal{C}$ consists of variables sets of size k.

The *random model* is defined as

$$\mathcal{F}(n,k) = \sum_{c \in \mathcal{C}} f_c(\mathbf{x}_c) \qquad (4)$$

where,

1. $\mathcal{C}$ consists of a collection of subsets of variables selected randomly according to a probability distribution from all the $\binom{n}{k}$ possible subsets of variables; and
2. The fitness values of each local fitness function are assigned randomly and independently according to a distribution on $[0, 1]$.

The *interaction graph of an additive fitness function* $f(\mathbf{x})$ is a graph $G_f(V, E)$ where the vertex set $V = \{X_1, \ldots, X_n\}$ corresponds to the set of variables, and $(X_i, X_j) \in E$ if and only if there is a subset $c \in \mathcal{C}$ such that $X_i \in c$ and $X_j \in c$.

In [4], the interaction graph of an additive fitness function is used to characterize the degree of interaction in an optimization problem. The interaction graph is calculated from the definition set of the function, and the degree of the variable interaction in an additive fitness function is measured using the treewidth of this graph. Other types of graphical representations (e.g. Bayesian networks) can be used to encode the fitness function structure in EDAs [12].

3.1 Interaction graphs and independence graphs

Intuitively, the notion of interaction can be associated with the synergy between two or more variables. A variable X_i interacts with another variable X_j if their combined effect on a third variable depends not only on their individual contributions but also on their combined effect. In GAs, the interaction between variables is usually known and analyzed under the name of epistasis [13]. Initial research [13] advanced the possibility that the amount of epistasis could be a measure of GA hardness, and that quantifying the epistasis may provide reliable *a priori* estimates of problem difficulty for GAs.

[2] Tournament selection can be done with or without replacement of the selected individual.

However, one of the difficulties of the known measures of epistasis is that they usually assume that the whole space of solutions is known. When one starts from a fitness function of a known structure, the idea of interaction can be translated into the idea of probabilistic dependency, but this requires the definition of a particular type of probability distribution. The strength, and the very existence of the dependency will depend on the way the probability distribution is defined. This is actually what occurs in EDAs, for which the characteristics of the probability distribution can be masked by the selection method chosen, particularly if it is an ordinal scheme.

3.2 Experiments

In the following experiments, we show that, in the case of random functions, the exact mapping between the structure of a random function and that of the independence graph learned from the data, arises only under very particular conditions, difficult to hold in EDAs.

The general framework of our experiment is the following. We constrain our analysis to an additive function of only one component and k binary variables. As the number of variables coincides with the size of the local function, all the variables are in the local function. Hence, for this random model, $\mathcal{F}(k,k) = f_c(\mathbf{x})$, and the fitness values of the local fitness function are assigned randomly and independently according to a distribution at $[0,1]$. We have chosen the uniform distribution at $[0,1]$.

For a function $f_c(\mathbf{x})$, all the 2^k values of the function are generated. Subsequently, the probability distribution $p(\mathbf{x})$ is defined, and it is evaluated whether its associated independence graph is complete (i.e. whether the variables that belong to c would form a clique in the independence graph). Our hypothesis is that probability distributions defined on random local functions seldom map to complete independence graphs. To evaluate this hypothesis we use the following procedure.

For a couple of variables X_i,X_j, and a parameter N, we set as the null hypothesis that the variables are independent in a population of N points that follow the probability distribution $p(\mathbf{x})$. Pearson's chi-square statistic is calculated:

$$\chi^2_{i,j} = \sum_{x_i,x_j} \frac{(Np(x_i,x_j) - Np(x_i)p(x_j))^2}{Np(x_i)p(x_j)} \qquad (5)$$

Then, the level of significance $_{i,j}$ corresponding to $\chi^2_{i,j}$ with one degree of freedom is found. $_{i,j}$ measures to what extent the hypothesis of marginal independence between X_i and X_j can be rejected. A large $_{i,j}$ means that the difference between $p(x_i,x_j)$ and $p(x_i)p(x_j)$ is very likely due to random fluctations in the data. A low value of $_{i,j}$ means that X_i and X_j are likely to be independent and therefore, the set of variables c could be divided into two sets. An exhaustive search for possible decompositions of c would include marginal probability tests of higher order as well as conditional independence probability tests.

In our experiment, we constrain the search for decompositions to bivariate marginal independence tests and take ¯ $= min_{i,j}$ $_{i,j}$ as a metric to measure the irreducibility of c. It should be noted that focusing only on bivariate interactions between single variables will not allow modeling the dependence of multiple variables. However, many EDAs only consider pairwise computations to decide the dependences to be represented. We emphasize that ¯ is only a crude lower bound of the decomposability. However, it is accurate enough for our purpose.

Given N as input, the steps of the experiment are the following:

1. Generate $f_c(\mathbf{x})$
2. Calculate probability distribution $p(\mathbf{x})$
3. Calculate and output ¯

Two classes of probability distributions are used: proportional distribution (1), and the Boltzmann distribution (2). Whenever ¯ is to be determined, we conduct $10,000$ experiments like the one described above, and estimate ¯ as the average of these experiments.

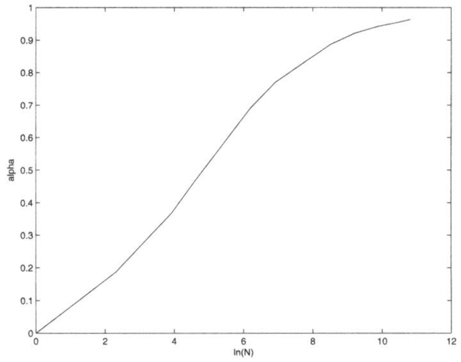

Figure 2: Values of ¯ for a random model $\mathcal{F}(3,3)$ when N increases.

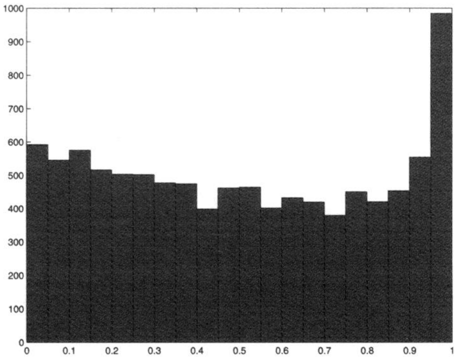

Figure 3: Histogram of the ¯ values corresponding to $10,000$ instances of random model $\mathcal{F}(4,4)$, when $N = 1,000$.

The first experiments are conducted using proportional selection. Figure 2 shows the ¯ values for a random model $\mathcal{F}(3,3)$ when N increases. As the population size N increases, the probability that the observed differences are not due to random fluctuations also increases. However, the population size needed to recover the model is not feasible. For instance, to obtain an average ¯ over 0.95, N must be higher than $20,000$.

The problem can be appreciated in Figure 3. In this figure, the histogram of the ¯ values corresponding to $10,000$ instances of the random model $\mathcal{F}(4,4)$ are shown. These values were calculated with $N = 1,000$. For most of the functions, ¯ < 0.95. Therefore, the hypothesis of independence cannot be rejected at a significance level of 0.05.

Finally, we investigate the same subject for the Boltzmann distribution at different values of k and for different temperatures. In the context of evolutionary algorithms, the role played by the temperature parameter of the Boltzmann distribution can be associated with the role of selection pressure, which is an important parameter that has been studied for different selection methods [14]. For the Boltzmann distribution, if the temperature increases, all points tend to have the same probability, and the shape of the probability is flattened.

Figure 4 shows the ¯ values for the Boltzmann distribution ($T \in \{0.1, 1, 10\}$), and the proportional distribution for $k = 2, \ldots, 7$, $N = 10,000$. In all the distributions, if k increases ¯ decreases. On the other hand, temperature increase makes the variables tend to be more independent. From the relationship between temperature and selection pressure, we can deduce that, in EDAs, selection pressure plays an important role to set the relationship between the function structure and the interactions captured by the independence graph.

To sum up, for random functions, the existence of a mapping between the function structure represented by the interaction graph and the independence graph will critically depend on the selection method, and the pressure of selection. For the class of random functions, it is not appreciated in general that the function structure is mapped to the independence graph. Many independence relationships seem to arise in random functions that reduce the complexity of the independence graph. The absence in the independence graph of edges between variables related by the underlying additive structure may not be harmful for EDAs because the functions defined over the subsets may not impose dependencies between the variables in these subsets at all.

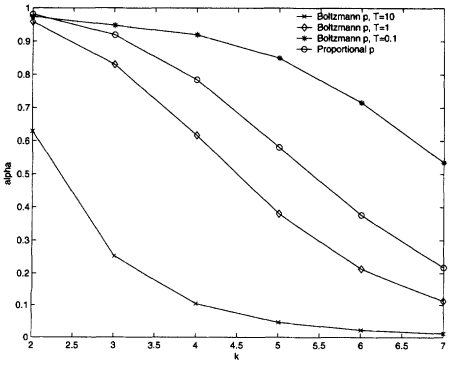

Figure 4: ¯ values for the Boltzmann distribution ($T \in \{0.1, 1, 10\}$) and the proportional distribution. In every case, $k = 2, \ldots, 7$ and $N = 10,000$.

4 Alternatives to deal with independence graph complexity

An important issue in EDAs, related to the analysis of EDA complexity, is how accurate the approximation of the distribution for the selected population must be. One of the developments of EDAs has been in the direction of using probability models able to capture a higher number of dependencies. Nevertheless, the complexity of the independence graph can make the application of exact probability models unfeasible. Yet, this does not imply that efficient EDAs cannot be conceived for these problems. The accuracy of the approximation may deteriorate and, still, the approximated model may be useful for the search. In EDAs, it is sometimes more important to distinguish good from bad solutions than to give an accurate approximation of all the points [15].

There are problems with interactions where simple EDAs that use univariate marginal distributions (e.g. UMDA [3]) can perform well. On the other hand, the class of deceptive functions [16] is an example of problems that can not be solved by simple GA or UMDA. Besides, there are problems that can be solved with simple and with more complex probabilistic models as well. However, in these cases the choice of the model also influences the number of evaluations needed. A gain in model simplicity may imply an increase in the number of function evaluations needed to solve the problem.

When the treewidth of the independence graph is too high, (situation that could be seen as a worst case scenario for EDAs), exact factorizations can be replaced by a number of alternatives like EDAs that use mixtures of distributions [17], EDAs based on the Kikuchi approximation of the distribution [10], and EDAs with probability models that use local structures [18].

4.1 An approach based on the use of malign interactions

In all the above mentioned alternatives to approximate the probability distributions, the same role is given to all dependencies that are learned by the model. However, it has been acknowledged that the existence of dependencies is not *per se* a source of difficulty for GAs. Only malign or deceptive interactions can be difficult to deal with [19]. Benign interactions are those where the interaction reinforces the message of its associated main effects. Malign interactions are those where the interaction counters the joint influence of its associated main effects. Malign interactions are what the GA community commonly calls deception [16, 19].

We analyze the interactions between variables in terms of their marginal probability distributions. We look for interactions in these marginal probability tables and use the probability values to evaluate whether the interactions are benign or malign. To illustrate this idea, an example is presented. Table 1 shows two binary functions of 3 variables ($BD3$, $MD3$). For each function, its corresponding proportional probability distribution ($p_{\mathcal{P}}^{s}$) has been calculated. Using the univariate marginal distributions calculated from each proportional probability distribution, two univariate approximations ($p^s_u(\mathbf{x}) = \prod_i p^s_{\mathcal{P}}(x_i)$) have been calculated and are shown as well in Table 1.

It can be seen that, for function $BD3$, the interactions reinforce the univariate effects. Particularly, it can be appreciated that $p^s_{\mathcal{P}}(\mathbf{x})$ and $p^s_u(\mathbf{x})$ reach their maximal values at the same configuration of the variables ($\mathbf{x} = 111$). The opposite occurs for function $MD3$; the configuration with the highest probability in the univariate model is not the same in the proportional distribution. The analysis done in Section 3 for probability dependencies is also valid for benign and malign interactions. The type of interactions will depend not only on the functions but also on the selection method, the selection pressure and other factors.

$\mathbf{x}$	$BD3$	$p^s_{\mathcal{P}}(\mathbf{x})$	$p^s_u(\mathbf{x})$	$MD3$	$p^s_{\mathcal{P}}(\mathbf{x})$	$p^s_u(\mathbf{x})$
000	0.0	0.0000	0.0688	0.9	0.2093	0.1965
001	0.8	0.1311	0.0991	0.8	0.1860	0.1415
010	0.8	0.1311	0.0991	0.8	0.1860	0.1415
011	0.9	0.1475	0.1427	0.0	0.0000	0.1019
100	0.8	0.1311	0.0991	0.8	0.1800	0.1415
101	0.9	0.1475	0.1427	0.0	0.0000	0.1019
110	0.9	0.1475	0.1427	0.0	0.0000	0.1019
111	1.0	0.1639	0.2056	1.0	0.2326	0.0734

Table 1: Functions with benign and malign interactions, associated proportional distributions and univariate approximations calculated from the proportional distributions.

Even if the importance of distinguishing between the effect of malign and benign interactions has been previously described in evolutionary computation, we are not acquainted with any previous use of malign interactions in EDAs. Now, a way to approximate probability distributions by considering only malign interactions is introduced. We propose to simplify the independence graph by removing all the edges corresponding to bivariate dependencies that are benign.

The assumption behind this approach is that benign interactions do not need to be considered by the model because their common effect can be captured by independently sampling the variables. Only those interactions that cannot be recovered by independently sampling the variables need to be stored in the model. Evidently, ignoring these dependencies will have an effect on the accuracy of the approximation, but the idea is that the proposed approach is sufficient regarding the aim of the EDA.

4.2 An EDA that discards benign interactions

To implement the idea presented in the previous section, we have chosen an EDA that uses a tree as its probabilistic model (Tree-EDA). EDAs that use the tree as its based probabilistic model were originally proposed by Baluja and Davies [20].

In the Tree-EDA, the model is found using the algorithm that calculates the maximum weight spanning tree from the matrix of mutual information between each pair of variables [20]. In the construction of the tree, a threshold is used to avoid the addition of edges between variables that are independent. In this way, the tree can have disconnected components, being in fact a forest.

The modification which we introduce in the classical Tree-EDA is to consider in the tree structure only edges that correspond to malign interactions. Once the bivariate and univariate probabilities have been estimated for all variables, we check, for each pair of variables, whether their most probable bivariate configuration is also the configuration for which the product of the respective univariate marginals is the highest. If such is the case, the bivariate dependency is benign and its mutual information is not calculated. This dependency will not be included in the final tree. Otherwise, we calculate the mutual information of this pair of variables. The step of checking the type of interaction can be inserted in the calculation of the mutual information. Algorithm 2 shows the pseudocode of Tree-EDA-M. Considering that bivariate and univariate marginal distributions have been calculated, the order of constructing the mutual information matrix is $O(n^2 \cdot L)$, where n is the number of variables, and L is the size of the largest bivariate probability table.

Algorithm 2: **Tree-EDA-M**

```
$D_0 \leftarrow$ Generate $M$ individuals (the initial population) randomly
$l = 1$
do {
    $D^s_{l-1} \leftarrow$ Select $N \leq M$ individuals from $D_{l-1}$ according to a selection method
    Compute the univariate and bivariate marginal frequencies $p^s_l(x_i)$ and $p^s_l(x_i, x_j)$ of $D^s_{l-1}$
    Detect malign interactions
    Calculate matrix of mutual information corresponding to malign interactions
    Calculate the tree structure $T$ from the matrix of mutual information
    $D_l \leftarrow$ Sample $M$ individuals (the new population) from $T$
    $l \Leftarrow l + 1$
} until A stop criterion is met
```

4.3 Experiments

The goal of our experiments is to evaluate whether the removal of benign interactions from the model has any implication in the behavior of EDAs. First, we have selected a set of eight instances of the Ising model as the function benchmark. The generalized Ising model is described by the energy functional (Hamiltonian) (6) where L is the set of sites called a lattice. Each spin variable $_i$ at site $i \in L$ either takes the value 1 or -1. One specific choice of values for the spin variables is called a configuration. The constants J_{ij} are the interaction coefficients. In our experiments, we take $h_i = 0$, $\forall i \in L$. The ground state is the configuration with minimum energy.

$$H = -\sum_{i<j\in L} J_{ij} \quad _i \quad _j - \sum_{i\in L} h_i \quad _i \qquad (6)$$

We have generated four random instances of the Ising model for each value of $n \in \{36, 64\}$. To generate a random instance where $J_{ij} \in \{-1, 1\}$, each coupling was set at -1 with probability 0.5. Otherwise, the constant was set at $+1$. We have verified the results using the Spin Glass Ground State server, provided by the group of Prof. Juenger[3].

To compare the effect of discarding the benign interactions, in our experiments we have used the normal Tree-EDA, and a Tree-EDA that only considers malign interactions (Tree-EDA-M). Additionally, and in order to also evaluate the effect of discarding benign interactions, we have implemented an EDA that only uses benign interactions (Tree-EDA-B). For each instance, we have found the population size needed by Tree-EDA to achieve a successful rate of 90 percent in 100 experiments. This critical population size has been used to compare the performance of algorithms using the same population size.

I	n	N	Tree-EDA		Tree-EDA-M		Tree-EDA-B	
			S	$\hat{g}$	S	$\hat{g}$	S	$\hat{g}$
$i1$	36	300	91	7.39	90	8.07	79	12.64
$i2$		500	93	7.87	90	8.22	65	12.87
$i3$		600	92	7.08	97	7.02	66	11.71
$i4$		700	96	7.77	98	7.10	67	12.47
$i5$	64	750	86	10.84	97	11.31	71	19.04
$i6$		1000	91	11.17	79	11.07	64	19.79
$i7$		1500	89	10.92	82	10.57	47	20.78
$i8$		1500	92	10.19	96	10.29	62	19.72

Table 2: Results of the comparison between Tree-EDAs with all, only malign and only benign interactions.

Table 2 shows the results of the algorithms for the eight instances. In the table, n refers to size of the instance, N is the critical population size, S is the rate of succes over 100 experiments, and $\hat{g}$ is the average number of generations needed to find the optimum in those experiments in which it is reached. The maximum number of generations was 30 and truncation selection with parameter of truncation 0.15 was used. Notice that, for some of the instances, the Tree-EDA has a success rate under 90. The reason is that the critical population size has been determined for a previous series of experiments.

It can be observed in the table that there are no evident differences between the performance of Tree-EDA and Tree-EDA-M. However, it can be seen that the performance of Tree-EDA-B is clearly inferior to the other two algorithms. To evaluate this difference from a statistical point of view, we have used the Kruskal-Wallis test of independence for each of the eight instances. The parameter used for the test is the best fitness value achieved in each experiment. The test intends to accept or reject the null hypothesis that the 100 samples (best solution found) corresponding to the algorithms outputs for each experiment have been taken from equal populations. The test significance level was 0.05.

For all the instances considered, no significant statistical differences were found between Tree-EDA and Tree-EDA-M. Significant statistical differences were found between Tree-EDA-M and Tree-EDA-B for all the instances except $i1$. For this instance, statistical differences appear for a test with significance level 0.1.

The experiments show that disregarding benign interactions does not influence the performance of the Tree-EDA. Nevertheless, capturing malign interactions is essential. For these problems, approximations based solely on malign interactions are enough to solve the problem efficiently.

5 Conclusions

In this paper we have analyzed two issues relevant to the understanding of EDAs. We have approached the problem of the relationship between the structure of the fitness function and the dependencies that arise in the probability distributions determined by the EDA. In this way, the role played by selection in the arousal of dependencies for the class of random functions has been emphasized. An important conclusion drawn from the experiments conducted is that even if the function structure plays an important role in the creation of dependencies, this role is mediated by selection. The complexity of EDAs will be sensitive to the way selection is defined. Our results are related with previous work [21] that identified selection as a determining factor in the performance of EDAs.

The other related issue considered in this paper is whether there are alternatives for problems in which the independence graph is too complex. We have introduced a new way to approximate probability distributions in EDAs. Our method is based on removing benign interactions hoping that these interactions will naturally arise during the sampling process.

The Tree-EDA-M has been introduced as an algorithm that only employs malign interactions to construct the model. Our preliminary results show that this algorithm can perform similar to the Tree-EDA, which does not distinguish between malign and benign interactions. However, the algorithm that only considers benign interactions suffers from a reduced performance.

The search for efficient and accurate approximations remains a challenging field in EDAs. We have shown that the type of interactions to be included in the probabilistic model can be used as one of the criteria to approximate the distributions. For our analysis we have assumed that whenever the combined effect of two variables on a third variable depends not only on their individual contributions but also on their combined effect, these two variables interact. Independence tests are able to detect these types of interactions.

The results of our work are preliminary. However, for the cases considered we have shown that it is possible to remove benign interactions from the probabilistic model without worsening the algorithm performance. A number of questions remain unanswered and deserve further research:

1. Can EDAs based on models that use only malign interactions outperform EDAs that employ all type of interactions?

[3] www.informatik.uni-koeln.de/ls_juenger/projects/sgs.html

2. How can one detect and take advantage of higher order malign interactions?

3. Can probability approximations based on malign interactions be useful in domains different from EDAs?

4. Which is the number of benign interactions actually included in the probabilistic model learned by EDAs?

If we can simplify graphical models by discarding benign interactions without a loss in efficiency that will be a gain for EDAs. Whether this is possible is one of the goals of our further research.

6 Acknowledgments

The authors thank three anonymous referees for their valuable comments on the reading. This work was supported by the Saiotek S-PEO4UN25 and Etortek-Biolan projects from the Basque Government.

Bibliography

[1] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[2] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2002.

[3] H. Mühlenbein and G. Paaß, "From recombination of genes to the estimation of distributions I. Binary parameters," in *Parallel Problem Solving from Nature - PPSN IV*. Berlin: Springer Verlag, 1996, pp. 178–187, lNCS 1141.

[4] Y. Gao and J. C. Culberson, "Space complexity of estimation of distribution algorithms," *Evolutionary Computation*, vol. 13, no. 1, pp. 125–143, 2005.

[5] H. Mühlenbein, T. Mahnig, and A. Ochoa, "Schemata, distributions and graphical models in evolutionary optimization," *Journal of Heuristics*, vol. 5, no. 2, pp. 213–247, 1999.

[6] Q. Zhang, "On stability of fixed points of limit models of univariate marginal distribution algorithm and factorized distribution algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 1, pp. 80–93, 2004.

[7] Q. Zhang and H. Mühlenbein, "On the convergence of a class of estimation of distribution algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 127–136, 2004.

[8] P. Larrañaga, *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2002, ch. An introduction to probabilistic graphical models, pp. 25–54.

[9] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction and Search*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, 1993, vol. 81.

[10] R. Santana, "Estimation of distribution algorithms with Kikuchi approximations," *Evolutionary Computation*, vol. 13, no. 1, pp. 67–97, 2005.

[11] K. Sastry and D. E. Goldberg, "Modeling tournament selection with replacement using apparent added noise," in *Intelligent Engineering Systems Through Artificial Neural Networks. Proceedings of the Conference ANNIE 2001*, vol. 2, 2001, pp. 129–134.

[12] J. Neil, "Mathematical Models of Estimation of Distribution Algorithms," Ph.D. dissertation, The University of Birmingham, School of Computer Science, 2005.

[13] G. J. E. Rawlins, Ed., *Foundations of Genetic Algorithms*. San Mateo: Morgan Kaufmann Publishers, 1991.

[14] T. Blickle and L. Thiele, "A comparison of selection schemes used in evolutionary algorithms," *Evolutionary Computation*, vol. 4, no. 4, pp. 361–394, 1997.

[15] T. Miquélez, E. Bengoetxea, and P. Larrañaga, "Evolutionary computation based on Bayesian classifiers," *International Journal of Applied Mathematics and Computer Science*, vol. 14, no. 3, pp. 101–115, 2004.

[16] D. Thierens, "Scalability problems of simple genetic algorithms. evolutionary computation," *Evolutionary Computation*, vol. 7, no. 4, pp. 331–352, 1999.

[17] R. Santana, A. Ochoa, and M. R. Soto, "The mixture of trees factorized distribution algorithm," in *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2001*. San Francisco, CA: Morgan Kaufmann Publishers, 2001, pp. 543–550.

[18] M. Pelikan, "Bayesian Optimization Algorithm: From Single Level to Hierarchy." Ph.D. dissertation, University of Illinois, 2002.

[19] L. Kallel, B. Naudts, and R. Reeves, "Properties of fitness functions and search landscapes," in *Theoretical Aspects of Evolutionary Computing*, L. Kallel, B. Naudts, and A. Rogers, Eds. Springer Verlag, 2000, pp. 177–208.

[20] S. Baluja and S. Davies, "Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space," in *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 30–38.

[21] A. Johnson and J. Shapiro, "The importance of selection mechanisms in distribution estimation algorithms," in *Proceedings of EA 2001*, ser. Lecture Notes in Computer Science, P. Collet, Ed., vol. 2310. Springer Verlag, 2002, pp. 91–103.