Approximating the Search Distribution to the Selection Distribution in EDAs

S. Ivvan Valdez Peña
Center for Research in Mathematics A.C.
C. Jalisco S/N, Mineral de Valenciana, Guanajuato, Gto. México
ivvan@cimat.mx

Arturo Hernández Aguirre
Center for Research in Mathematics A.C.
C. Jalisco S/N, Mineral de Valenciana, Guanajuato, Gto. México
artha@cimat.mx

Salvador Botello Rionda
Center for Research in Mathematics A.C.
C. Jalisco S/N, Mineral de Valenciana, Guanajuato, Gto. México
botello@cimat.mx

ABSTRACT

In an Estimation of Distribution Algorithm (EDA) with an infinite sized population the selection distribution equals the search distribution. For a finite sized population these distributions are different. In practical EDAs the goal of the search distribution learning algorithm is to approximate the selection distribution. The source data is the selected set, which is derived from the population by applying a selection operator. The new approach described here eliminates the explicit use of the selection operator and the selected set. We rewrite for a finite population the selection distribution equations of four selection operators. The new equation is called the empirical selection distribution. Then we show how to build the search distribution that gives the best approximation to the empirical selection distribution. Our approach gives place to practical EDAs which can be easily and directly implemented from well established theoretical results. This paper also shows how common EDAs with discrete and real variables are adapted to take advantage of the empirical selection distribution. A comparison and discussion of performance is presented.

Track: Estimation of Distribution Algorithms.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Algorithms

Keywords

Estimation of Distribution Algorithms, Selection Methods

GECCO'09, July 8–12, 2009, Montréal, Québec, Canada.

1. INTRODUCTION

Estimation of Distribution Algorithms (EDAs) [11, 13] are an increasing and promising area in the Evolutionary Computation (EC) research field. The standard EDA is described as follows:

1. Initialize a probability model $p(x,t)$.
2. Generate a sample X_t from $p(x,t)$.
3. Evaluate X_t in the objective(s) function(s) and constraint(s).
4. Select a subset S_t from X_t (selection step).
5. Recompute the model $p(x,t+1)$ by approximating it to the underlying distribution $\hat{p}^S(x,t)$ of S_t (parameter computation step).
6. $t = t+1$
7. If the stop criterion is not reached, go to step 2.

This algorithm has been applied for discrete [13, 14] as well as for continuous [3, 8, 9, 11] variables. Notice that the selection step (step 4) and the parameter computation step (step 5) introduce a search bias that increases the probability of sampling the most promising regions. If either of these steps were omitted, no improvement of the objective function is possible. The **search distribution**, $p(x,t)$, is the model used for learning the underlying distribution of the selected set. The selection distribution or selection model depends on the selection operator (tournament, proportional, etc). The **selection distribution** is the probability of selecting any point in the search space. The selection distribution plays a central role in this paper and will be explained in detail.

1.1 Ideal EDA with an infinite sized population

The main goal of a selection operator is to bias the population towards promising regions of the search space. The four selection operators widely discussed in this paper are introduced through an example. Assume the next objective function and infinite sized population $f(x,y) = \exp(|x|-2)^2 + 4\cos(20x) + \exp(|y|-2)^2 + 4\sin(20y)$ in a continuous search space, as shown in Figure 1(a). Most of the EDAs initially draw a population from a uniform distribution such as

shown in Figure 1(b). Hence any point has the same chance to be sampled. When the population is selected by the truncation method with a function threshold of $\theta = 30$, the resulting underlying density function of the selected set looks like Figure 1(c). Observe the flat area; it means the selection probability for the population above the threshold value is the same. The truncation selection hides the roughness of the objective function by making some areas equally good, and the rest equally bad. The Boltzmann selection exponentially favors the most promising regions, as shown in Figure 1(d). Most of the probability mass is condensed on a single peak which corresponds to the function optimum. Since the remaining region is quite flat, the Boltzmann selection will deliver a selected set clustered on the peak region. The proportional selection, (adequately described by its name), selects points with a probability directly proportional to its objective value. The resulting probabilistic model of this method, shown in Figure 1(e), is very similar (although in a different scale) to the objective function. The tournament selection picks the best point found in a subset of the population. Figure 1(f) shows that the resulting probabilistic model acquires a similar roughness to that of the objective function.

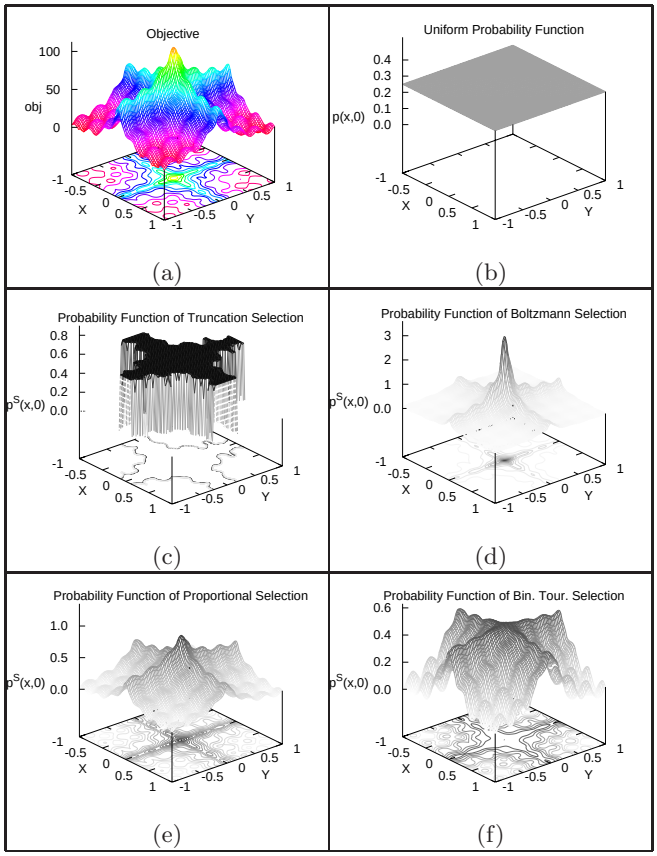


Figure 1: Initial model in EDAs (uniform) and the selection models of different selection methods.

Zhang and Mühlenbein have shown the selection models just illustrated can drive the population to the function optimum [21]. The main factor that makes convergence possible is the bias the selection operator introduces into the population. Recall we have defined the *selection model* as the

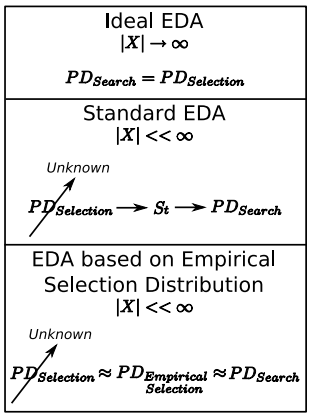


Figure 2: Strategy used in EDAs: Ideal EDA, Standard EDA, and EDA based on the empirical selection distribution.

probability of the selection operator for sampling a point of an infinite sized population. The ideal EDA with an infinite sized population analyzed by Zhang and Mühlenbein use the selection distribution as search distribution. This is not the case for a finite sized population.

1.2 Practical EDAs with a finite sized population

The advantage of a finite population is that computation is possible but nonetheless an approximation. For a finite population the selection distribution can only be approximately inferred by means of the particular points of the population. The common approach found in practical EDAs is the use of a **search distribution** to approximate the selection distribution. The search distribution is the model with a predefined structure used for learning the underlying joint probability density of the selected set, and later sampled to regenerate the next population. Examples of search distributions are: Bayesian networks [14], Polytrees [18] in discrete spaces, as well as Gaussian models [8] [3] in the continuous case, among others [9, 16, 10].

Summarizing: the main differences between infinite and finite sized populations can be briefly stated as follows. For an infinite population the remarkable characteristic is that the selection distribution and the search distribution are the same. This is shown in Figure 2, top row. However, for a finite sized population they are different. The true analytic expression of the selection distribution is unknown. The selection operator delivers a selected set (indicated as S_t) from which the search distribution builds a model (see second row of Figure 2). The goal of the search distribution is to approximate the selection distribution. The last row of Figure 2 shows the new proposal introduced by this paper. The selection distribution is unknown (for a finite population), however, the selection distribution equations (for an infinite population) are rewritten for the finite sized population case, and named the **empirical selection distribution**. The parameters of the search distribution are computed via the empirical selection distribution equations. There is no explicit application of any selection operator, however, selection happens just as in the infinite sized population case because it is inherent to the empirical selection distribution equation being used.

The paper is presented as follows: Sections 2 and 3 introduce the general method to approximate the search distribution to the selection model. Section 4 is a comparison with related work. Section 5 presents well known EDAs, which have been modified to apply the proposed method. A set of experiments is presented in Section 6 and discussion about the performance of different selection methods. Finally Section 7 shows the perspectives of future work and concludes.

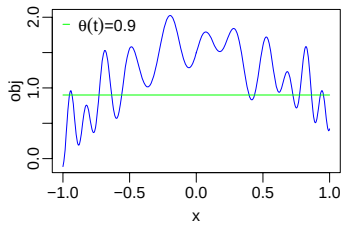


Figure 3: Unidimensional function example.

2. COMPUTING THE EMPIRICAL SELECTION DISTRIBUTION

At any generation, the population X_t contains all the information we can know about the whole search space. Hence we approximate the selection distribution $p^S(x,t)$, with the current population X_t. The **empirical selection distribution** $\hat{p}^S(x,t)$ is the exact selection distribution when the population is thought as a model of the whole search space. Four selection distribution models (of four selection operators) are shown in each row of the left column of Table 1. Each one is split into two parts. The upper half shows the selection model for an infinite sized population, $p^S(x,t)$, whereas the lower half shows the analog equation for a finite sized population. As mentioned, that is the empirical selection distribution $\hat{p}^S(x,t)$. For the sake of a complete comparison of the distributions, their plots are presented in the middle and right columns assuming the maximization of the objective function shown in Figure 3. The empirical selection distribution defines a probability $\hat{p}^S(x_i,t)$ for each individual, requiring for its computation only the objective function value at the point. The empirical distribution can be computed for univariate and multivariate problems, in continuous or discrete search spaces, as it shall be shown through several examples in Section 5. Notice that the empirical selection distribution approximates the selection model when the population size goes to infinite.

3. COMPUTING THE SEARCH DISTRIBUTION

Following the usual steps of an EDA algorithm, an arbitrary large sample can be obtained from the empirical selection distribution, and then used to learn the search distribution parameters. Fortunately, sampling the empirical selection will be avoided without diminishing the advantages of using all the information at hand. It is known that the relative frequency of a point x_i when and infinite large sample is drawn is equal to the probability $\hat{p}^S(x_i,t)$. Thus, the sampling process can be avoided and $\hat{p}^S(x_i,t)$ can be used as the frequency of the point x_i.

For example, suppose that the search distribution is a Gaussian with mean μ and a variance v. These parameters must be computed to get the best approximation of the Gaussian to the selection model. Assume (for a moment) a sample $\hat{S}$ of size $|\hat{S}|$ obtained from the empirical selection distribution. As the only possible values sampled from the empirical selection distribution are those in the population, most of the points x_i would get more than one instance in the sample $\hat{S}$. Denote such number of instances of a point x_i as $freq_i$, then the estimator of the mean is:

$$\mu = \frac{\sum_{i=1}^{|X|} freq_i \cdot x_i}{|\hat{S}|} \tag{1}$$

Let us denote for short $\hat{p}_i^S = \hat{p}^S(x_i,t)$, then we know that when $|\hat{S}| \to \infty$ then $\frac{freq_i}{|\hat{S}|} = \hat{p}_i^S$. Substituting this term into Equation 1, the mean is simply computed as follows:

$$\mu = \sum_{i=1}^{|X|} \hat{p}_i^S \cdot x_i \tag{2}$$

Where $\hat{p}_i^S$ is the probability of a point computed with the empirical selection distribution. Therefore, sampling the empirical selection distribution is not necessary. Also important, note that the computation of the probabilities $\hat{p}_i^S$ are independent of the search distribution parameters. This allows to easily adapt an EDA to any of the four selection models. This is not a limitation since these models cover most EDA implementations. In addition, the computational cost of $\hat{p}_i^S$ in most of the cases is lower or equal to that of applying a selection operator. For example, the Proportional and Boltzmann selection must calculate the probabilities in Table 1 in order to obtain a sample which becomes the selected set; on the other hand, the sample is not necessary for the empirical selection distribution based EDA (ES-EDA). For the truncation method in the standard EDA as well as the ES-EDA, it is only necessary to known the individuals with objective function greater than θ_t, thus the computational cost is the same. The binary tournament selection requires comparisons of the order of the selected set size as selection method, while the empirical selection distribution requires comparisons of order $|X|(|X|-1)/2$, thus it is the unique case in which an ES-EDA have a greater cost than the standard EDA. The parameter computation in the ES-EDA in general increases its computational cost just by multiplying each individual value x_i by its corresponding frequency $\hat{p}_i^S$ which is a minimal cost.

4. RELATED WORK

In order to show that this work is an extension of previous but less general approaches, we will show the equivalence of the resulting formulae when applying frequencies from the empirical selection distribution. Yunpeng et al.[20], approximate the Boltzmann distribution with a Gaussian model by minimizing the Kullback-Leibeler divergence. The computation of the Gaussian distribution parameters is presented in Equations 3 and 4.

$$\mu = \sum_{i=1}^{|X|} e^{\Delta\beta f(x_i)} x_i / \sum_{i=1}^{|X|} e^{\Delta\beta f(x_i)} \tag{3}$$

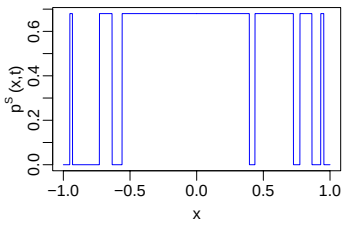
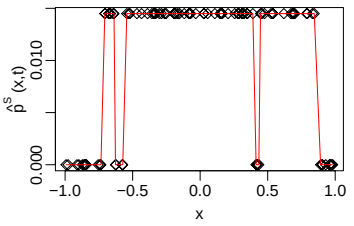
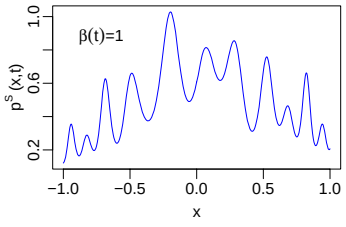
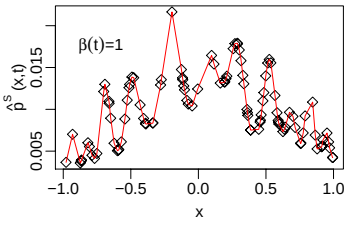
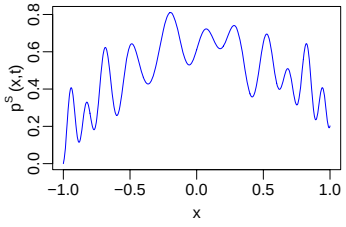
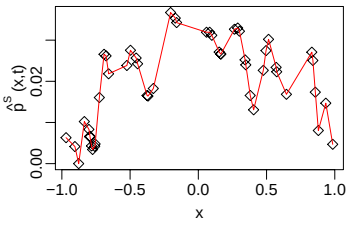
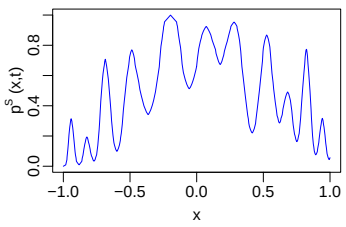
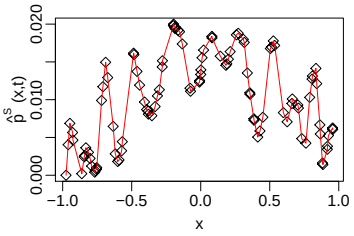
Selection model Empirical selection distribution	Selection model plot	Empirical Selection Dist. plot
Truncation: $p^S(x,t) = \begin{cases} \frac{p(x,t,\theta_t)}{\alpha(t)} & \text{if } f(x) \geq \theta_t \\ 0 & \text{otherwise} \end{cases}$ $\hat{p}^S(x_i,t) = \begin{cases} \frac{1}{\lvert S_t \rvert} & \text{if } f(x_i) \geq \theta_t \\ 0 & \text{otherwise} \end{cases}$ $\lvert S_t \rvert = \sharp$ of individuals which $f(x_i) > \theta_t$		
Boltzmann: $p^S(x,t) = \frac{e^{\beta(t)f(x)}p(x,t)}{Z(t)}$ $\hat{p}^S(x_i,t,\beta_t) = \frac{e^{\beta(t)f(x_i)}}{\sum_{j=1}^{\lvert X \rvert} e^{\beta(t)f(x_j)}}$		
Proportional: $p^S(x,t) = \frac{f(x)p(x,t)}{E(t)}$ $\hat{p}^S(x_i,t) = \frac{f(x_i)}{\sum_{j=1}^{\lvert X \rvert} f(x_j)}$		
Tournament: $p^S(x,t) = 2p(x,t)\int_{f(y)\leq f(x)} p(y,t)dy$ $\hat{p}^S(x_i,t) = \frac{\sum_{j=1}^{\lvert X \rvert} I(i,j)}{\sum \lvert X \rvert_{i=1} \sum_{j=1}^{\lvert X \rvert} I(i,j)}$ Where $I(i,j) = 1$ if $f(x_j) < f(x_i)$ and 0 otherwise		

Table 1: Left: mathematical expression of selection models and empirical selection distribution. Middle: selection model plot. Right: empirical selection distribution plot.

$$\Gamma = \frac{\sum_{i=1}^{|X|}[e^{\Delta\beta f(x_i)}(x_i-\mu)(x_i-\mu)^T]}{\sum_{i=1}^{|X|} e^{\Delta\beta f(x_i)}} \quad (4)$$

Where μ is the mean, Γ is the covariance matrix, X the population, and $|X|$ the population size.

Note that this approach which requires costly analytical work, gives exactly the same formulae that the empirical distribution presented in Table 1 when using the Boltzmann empirical selection distribution in UMDA$_c$ and EMNA$_{global}$ (with β= $\Delta\beta$, in Yunpeng et al. approach).

5. MODIFYING SUCCESSFUL EDAS WITH THE EMPIRICAL SELECTION DISTRIBUTION

This section presents several successful EDAs which have been modified to accept the relative frequencies given by the empirical selection distribution. Continuous and discrete EDAs with univariate and multivariate models are presented. Table 2 presents the notation used. A standard EDA algorithm is presented in Table 3. Since the parameter computation in line 8 is the only section that should be changed according to the particular search distribution, a pseudo-code is separately presented for each algorithm. The selection procedure in line 6 must be understood as the computation of the empirical selection distribution (according to Equations in Table 1).

5.1 UMDA

The Univariate Marginal Distribution Algorithm (UMDA) was introduced by Mühlenbein and PaaB [13]. It is based on the estimation of univariate marginal probabilities. This model considers all variables are statistically independent. It uses the simplest model for discrete distributions. Each variable x_i has attached a probability vector $b_{i,k}$, that is, the probability of x_i taking the value k, is: $b_{i,k} = p(x_i = k)$. Note that in the original UMDA the computation of the

$p(x,t) \rightarrow$ Search distribution in generation t.
$n \rightarrow$ Number of variables.
$X_t \rightarrow$ Population.
$\|X\| \rightarrow$ Population size.
$\hat{p}_j^s \rightarrow$ Frequency for x_j, according to $\hat{p}_j^s = \hat{p}^s(x_j,t)$ in Table 1.
$F_t \rightarrow$ Population objective values.
$X_{best} \rightarrow$ An optimum approximation.
$I(x_i,k) \rightarrow$ Indicator, 1 if $x_i = k$, and 0 otherwise.

Table 2: Notation used in EDAs of this section.

1	Initialize the probability model $p(x,0)$
2	`Sampling`$(p(x,0),\ X_t)$
3	`Evaluation`(X_t,F_t)
4	$t \leftarrow 1$
5	`While`(*the stopping criterion is not met*) {
6	`Selection`$(\hat{p}^s(x,t),X_t,F_t\ )$
7	`Parameter_computing`$(\hat{p}^s(x,t),X_t,F_t\ )$
8	`Sampling`$(p(x,t),\ X_t)$
9	`Evaluation`(X_t,F_t)
10	`Elitism`(X_t,F_t,X_{best})
11	}
12	`Return` X_{best}

Table 3: General Empirical Selection Distribution based EDA

parameter $b_{i,k}$ is basically made by a counting bits of a variable of the selected set. It is quite simple to adapt to the relative frequencies given by the empirical selection distribution. The pseudo-code of the UMDA-ES (ES=Empirical Selection) parameter computation is shown in Table 4.

1	`For`$(i = 1$ to $n)$ {
2	`For`$(k = 1$ to $m_i)$ {
3	$b_{i,k} = \sum_j^{\|X\|} I(x_j,k) * \hat{p}_j^s$
4	}
5	}

Table 4: Parameter computation for UMDA using the empirical selection.

5.2 UMDAc

The Univariate Marginal Distribution Algorithm for continuous domains (UMDA$_c$) was introduced by Larrañaga et al. [8]. It uses an univariate model, in the specific case of UMDA$_c^G$, there are n univariate Gaussian distributions (for an n-dimensional problem). Two parameters are needed for Gaussian at each dimension i, the mean μ_i and standard deviation σ_i. The computation of both parameters is simply done by weighting each point by its corresponding relative frequency (probability).

1	`For`$(i = 1$ to $n)$ {
2	$\mu_i = \sum_j^{\|X\|} \hat{p}_j^s \cdot x_{j,i}$
3	$\sigma_i^2 = \sum_j^{\|X\|} \hat{p}_j^s \cdot (x_{j,i} - \mu_i)^2$
4	}

Table 5: UMDA$_c$ using the empirical selection to compute the search distribution parameters.

Where $\hat{p}_i^S$ is the probability of a point computed with the empirical selection distribution.

5.3 K2-Bayesian-network based EDA

Bayesian networks have been successfully used in EDAs, by instance the Bayesian Optimization Algorithm introduced by Pelikan et al. [14]. A BOA-like algorithm based on the K2 algorithm [4] is presented. The parameter computation has been modified to use the empirical selection distribution. The K2 is a greedy heuristic search method, for maximizing the probability $P(B_S,D)$ of the structure B_S and the data D. For maximizing $P(B_S,D)$ the K2 maximizes $g(i,\pi_i)$, which is a measure related with the probability of x_i given a set of parents π_i.

$$g(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)} \prod_{k=1}^{r_i} N_{ijk}! \qquad (5)$$

where x_i has r_i possible discrete values. Each variable x_i in B_S has a set of parents, which are represented by a list of variables π_i. N_{ijk} is the number of cases in D in which the variable x_i has the value v_{ik}, and π_i is instantiated as w_{ij}. w_{ij} denote the jth unique instantiation of π_i relative to D, and q_i are the number of such unique instantiations of π_i. $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. For deeper explanation of the K2 algorithm the reader is directed to [4].

Learning the Bayesian network with the K2 is a counting process of point instances. Then, in order to compute Equation 5 an integer frequency for each point is needed. For obtaining such integer frequency let us define a sample size for the selection method, say $|\hat{S}_t|$. When using the empirical selection distribution we a have a relative frequency $\hat{p}_l^s$ associated with a point l, if that point is such a case for N_{ijk}, then instead of summing a single case we will sum $integer(|\hat{S}_t| \cdot \hat{p}_l^s)$. As $|\hat{S}_t|$ grows, the approximation of the K2 Bayesian network to the empirical selection distribution will be more accurate, but the computational cost of Equation 5 increases as well. Once the K2 Bayesian Network Structure has been learned, the frequencies $|\hat{S}_t| \cdot \hat{p}_l^s$ must be used to compute the conditional probabilities.

5.4 EMNAglobal

The EMNA$_{global}$ was introduced by Larrañaga et al. [9]. It is based on the estimation of a multivariate normal density function. This model can represent linear dependencies between Normal variables in the covariance matrix. The pseudo-code of the EMNA$_{global}$-ES parameter computation is shown in Table 6. Note that it is quite easy to insert the empirical selection distribution in univariate as well as multivariate algorithms, also it was inserted in discrete and continuous domains with a minimum analytical and computational effort.

6. EXPERIMENTS AND PERFORMANCE ANALYSIS

This section provides a set of experiments to address the performance and advantages of the empirical selection distribution. Two kind of experiments are presented:

1. Graphically we show that the search distribution based on the empirical selection distribution constitutes a robust and better approximation to the selection model.

```
1   For(i = 1 to n)  {
2       μ_i = Σ_j^|X| p̂_j^s · x_{j,i}
3       For (k = 1 to i)  {
4           σ_{i,k} = Σ_j^|X| p̂_j^s · (x_{j,i} − μ_i)(x_{j,k} − μ_k)
5           σ_{k,i} = σi, k
6       }
7   }
```

Table 6: EMNA$_{global}$ using the empirical selection to compute the search distribution parameters.

2. Using the EMNA$_{global}$ presented in Section 5, we shown the impact of the empirical selection distribution on the EDA performance.

6.1 Graphical Comparison

For better plots, the unidimensional objective function shown in Figure 3 is used. This analysis contrasts the three discussed concepts summarized in Figure 2: the ideal EDA, the standard EDA, and the EDA with the empirical selection distribution, combined with the four discussed selection methods.

Even though the exact selection distribution can be computed for the ideal EDA, to perform the correct comparison we used a predefined Gaussian model for the three approaches, say we show the Gaussian approximation with a very large population (considered as infinite), contrasted with the approximation computed by using the selected set from a standard sized population, and the approximation computed by using the empirical selection distribution. The experiments are designed as follows:

- The ideal EDA. A very large population (10^4 individuals) equally spaced is used, then a larger selected set (10^5) is extracted using the selection method. Using the huge selected set (10^5 individuals) the parameters of the search distribution are computed. This approximation will be called the exact approximation. A special case is the truncation method which uses the same population size, but a smaller selected set is obtained by truncating at $\theta = 0.9$.

- The standard EDA. A relatively small population (30 individuals) is randomly drawn, the selection method is applied delivering the selected set used to compute the search distribution. Most of the selection methods add randomness to the procedure (except the truncation method which delivers a deterministic selected set), by consequence the search distributions could differ after applying several times the selection method to the same population. Thus, we present the different search models computed when selecting 30 selected sets of 15 individuals from the same population.

- The EDA with the empirical selection distribution. Using the same population of 30 individuals used by the standard EDA, the empirical selection distribution is used to compute the search distribution parameters. Notice that the empirical selection distribution is unique for a given population as well as the search model delivered by this method.

Truncation selection. The truncation method shown in Figure 4(a) is basically the same for the standard EDA and the empirical selection distribution, because of the points used are the same in both approximations. Thus, in this case the empirical selection performs at least as the common selection method.

Boltzmann selection. As shown in Figure 4(b), the empirical selection method delivers a very good approximation to the exact model. The search models computed by the standard EDA approximation are a belt of Gaussians, the empirical selection approximation is at the middle of this belt, it is less affected by a single point or small set of points (robustness). It is possible that the randomness of the common selection method guides the search to a better optimum approximation, but in general it is not the expected behavior given its difference with the exact model, the small size of the selected set (usually 50% of the population) and the consequent lost of information which favor the tendency of being biased to sub-optimal regions. Also, it is expected that the behavior of the Boltzmann selection varies according the β value. The empirical selection computes the same search model from the same population, thus a more stable behavior is expected in contrast with the common EDA, this could be specially useful when designing annealing schedules, because with the empirical selection distribution a similar performance under similar conditions is expected.

Proportional selection. The proportional selection does not use extra parameters, therefore, no expensive parameter tuning is required. At the same time, however, there is no way to control the high selection pressure exerted over the fitted individuals that usually lead the population to premature convergence. Thus a search distribution which better represents the selection method could be useful, also a more predictable search model is useful when tuning other controls such as the population size, because the expected behavior could be better inferred. Figure 4(c), shows 30 search models (points) computed when applying the selection method 30 times on the same population, notice that this models could be easily biased to different regions, most of them suboptimal.

Binary tournament selection. The binary tournament selection seems to be the most robust selection method according to Figure 4(d). The models delivered by the standard EDA are more similar among them than those delivered by other selection methods. This selection could be a starting point when looking for the best parameters for other methods, because it is parameter free, and it is less sensible to large objective values in the population. As shown the empirical selection distribution computes an accurate approximation when compared with the exact method.

6.2 Performance Comparison

It has been previously shown that the estimation of the search distribution according with the empirical selection distribution, accurately approximate the exact search distribution with an infinite sized population. To support this conclusion, we present a comparison using the EMNA$_{global}$ and two problems widely used to test EDAs performance [9]: Sphere and Griewangk. In order to maximize the function and convert it as to have positive objective values, the objective function $g(x)$ was adjusted as: $f(x) = -g(x) + 1.8e7$, and $f(x)$ is used as fitness function.

Experiments description: 30 independent runs were performed with 50 variables, the domain for all variables is $[-600, 600]$, population size of 2000 for the truncation, 1000

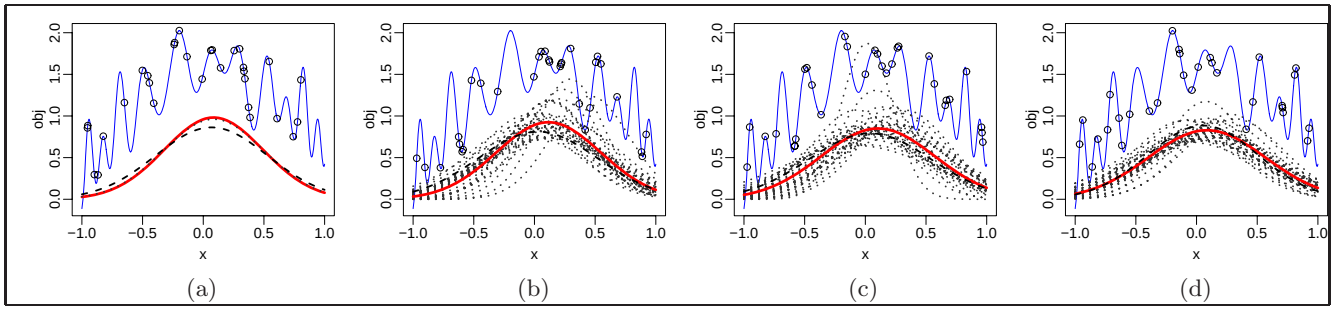


Figure 4: Selection methods:(a)truncation, (b) Boltzmann, (c)proportional and (d) tournament. Search distributions for: 1)The exact approximation with dashed line, 2)the standard EDA approximation with doted lines, and 3) the empirical selection approximation with solid line. The objective function (solid line), and the population used for practical EDAs (circles).

	Trunc.	**Boltzmann**	**Propor.**	**Bin. Tour.**
		Sphere		
St.	6.1222e-4 (2.2574e-3)	8.3358e+3 (2.6645e+3)	1.5882e+5 (4.2970e+4)	3.2699e+2 (2.2010e+2)
ES	1.9540e-3 (4.9814e-3)	1.6882e+3 (2.5745e+2)	8.5791e+4 (1.4717e+4)	1.9615e-2 (2.1285e-2)
Best	N	**ES**	**ES**	**ES**
		Griewangk		
St.	0.003645 (0.0025)	81.8394 (17.09844)	98.4952 (23.9788)	0.9760 (0.1506)
ES	4.358e-3 (0.002949)	128.4277 (20.213)	83.21131 (17.1586)	8.3975e-03 (0.001228)
Best	**N**	**St**	**ES**	**ES**

Table 7: Mean and (Standard deviation), of 30 independent runs of the standard EDA (St.) and the Empirical Selection distribution-based EDA (ES), and the result of the hypothesis test ran to determine which one of them is better, N =neither one, ES =Empirical Selection Distribution, and St =Standard.

for the proportional and 1500 for the other selections, selected set size for the standard EDA of 500. The annealing schedule for the Boltzmann selection was fixed by initializing $\beta_0 = 0$, $\delta\beta = (2E-5)/200$, and $\beta_t = \beta_{t-1} + \delta\beta$. Truncation threshold at 50% of the population, and elitism according to the general EDA in Table 3.

Stopping criterion: 300 000 function evaluations.

Results: Table 7 compares the results obtained by the standard EDA (denoted by **St.**), and the Empirical Selection based EDA (denoted by **ES**). We report the mean and standard deviation as well as the result of a hypothesis test based on the Bootstrap methodology with a significance of 5% [5]. We tested that the mean of the ES-based EMNA is less than the mean of the standard EMNA, and vice versa. The row "Best", says which algorithm performs the best according to the hypothesis test. As shown, with the truncation method neither algorithm could be considered superior. However, with the remaining methods there is sufficient statistical evidence to say that the Empirical Selection-based EMNA is the best. The consistent improvement of the performance can be noted in the smaller standard deviation of the ES-based algorithm.

7. PERSPECTIVES, FUTURE WORK AND CONCLUSIONS

EDAs researches have approximated the selection distribution since the first approaches [12, 3]. This paper proposes a general method for this purpose. The only information used in standard EDAs to learn the search distribution is a set of points and its frequency (selected set), this is the same information delivered by the empirical selection distribution, thus any standard EDA can be improved with the empirical selection distribution.

By instance bivariate models [15] and histograms [19] are completely based on frequencies, another important remark are clustering based algorithms [9], for example the k-means algorithm is based on distances, when using the empirical selection distribution in clustering, instead of using a single point in the position x_i, we use its relative frequency $\hat{p}^S(x_i, t)$. This measurement will move the mean of the cluster to the regions with highest fitness values, helping to perform a better search.

Important issues on EDAs such as diversity and premature convergence can be tackled using the empirical selection distribution. Studies on convergence phases [7] have shown that the maximum likelihood variance might not be the best way of proceeding to perform an optimum search. Thus, a possible line of work is: how to favor diversity using the empirical selection distribution?. A possibility is by simply modifying the fitness function into the empirical selection distribution formulae. This line of work could be an alternative to recent proposals on variance scaling [6, 2].

Multi-objective applications and insertion of the Pareto dominance criterion in the selection distribution is another

possible research line, the Pareto ranking seems the most natural way of tackling this important issue.

Ever since the very first proposed EDAs [1] to the most recent works [17], incremental learning has been applied to the learning distribution phase. The future work must contemplate how to insert the empirical selection distribution into incremental approaches, or how to use historical or previous empirical selection distributions.

The selection methods presented are just a sample of the possibilities, other methods such as combined truncation-proportional, truncation-Boltzmann, 3-tournament, etcetera must be explored.

Finally, the presented method is easy to implement, has a wide range of application, low computational as well as analytical cost, avoids to be wrongly biased by a single solution or a small set, and uses all the information on the population to accurately approximate the selection distribution. The perspectives and future use and applications are promising, and the possible lines of work are really extensive.

8. REFERENCES

[1] S. Baluja. Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[2] P. A. N. Bosman, J. Grahl, and F. Rothlauf. SDR: A Better Trigger for Adaptive Variance Scaling in Normal EDAs. In *GECCO '07: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 516–522. ACM, 2007.

[3] P. A. N. Bosman and D. Thierens. Expanding from Discrete to Continuous Estimation of Distribution Algorithms: The IDEA. In *PPSN VI: Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, pages 767–776, London, UK, 2000. Springer-Verlag.

[4] G. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistics Networks from Data. *Machine Learning*, 9:309–347, 1992.

[5] B. Efron and R. J.Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall,Monographs on Statistics and Applied Probability, New York, 1993.

[6] J. Grahl, P. A. Bosman, and F. Rothlauf. The Correlation-Triggered Adaptive Variance Scaling IDEA. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 397–404, New York, NY, USA, 2006. ACM.

[7] J. Grahl, P. A. N. Bosman, and S. Minner. Convergence Phases, Variance Trajectories, and Runtime Analysis of Continuos EDAs. In *GECCO '07: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 516–522. ACM, 2007.

[8] P. Larrañaga, R. Etxeberria, J. Lozano, and J. Peña. Optimization by Learning and Simulation of Bayesian and Gaussian Networks. Technical Report EHU-KZAA-IK-4/99, Department of Computer Science and Artificial Intelligence, University of the Basque Country, 1999.

[9] P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[10] J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, editors. *Towards a New Evolutionary Computation*, volume 192 of *Studies in Fuzziness and Soft Computing*. Springer, 2006.

[11] H. Müehlenbein[1], J. Bendisch[1], and H.-M. Voight[2]. From Recombination of Genes to the Estimation of Distributions II. Continuous Parameters, 1996.

[12] H. Mühlenbein. The Equation for Response to Selection and Its Use for Prediction. *Evolutionary Computation*, 5(3):303–346, 1997.

[13] H. Mühlenbein and G. Paaβ. From Recombination of Genes to the Estimation of Distributions I. Binary Parameters. In *PPSN IV: Proceedings of the 4th International Conference on Parallel Problem Solving from Nature*, pages 178–187, London, UK, 1996. Springer-Verlag.

[14] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz. BOA: The Bayesian Optimization Algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99*, volume I, pages 525–532, Orlando, FL, 1999. Morgan Kaufmann Publishers, San Fransisco, CA.

[15] M. Pelikan and H. Mühlenbein. The Bivariate Marginal Distribution algorithm. *Advances in Soft Computing Engineering Design and Manufacturing*, pages 521–535, 1999.

[16] M. Pelikan, K. Sastry, and E. Cantú-Paz, editors. *Scalable Optimization via Probabilistic Modeling*, volume 33 of *Studies in Computational Intelligence*. Springer, 2006.

[17] M. Pelikan, K. Sastry, and D. E. Goldberg. iBOA: The Incremental Bayesian Optimization Algorithm. In *GECCO '08: Proceedings of the 2008 Conference on Genetic and Evolutionary Computation*, pages 455–462. ACM, 2008.

[18] M. Soto and A. Ochoa. A Factorized Distribution Algorithm based on Polytrees. In *Proceedings of the 2000 Congress on Evolutionary Computation (CEC 2000)*, pages 232 – 237. IEEE, 2000.

[19] S. Tsutsui, M. Pelikan, , and D. E. Goldberg. Evolutionary Algorithm using Marginal Histogram Models in Continuous Domain. Technical Report 2001019, Illinois Genetic Algorithms Laboratory, 2001.

[20] C. Yunpeng, S. Xiaomin, and J. Peifa. Probabilistic Modeling for Continuous EDA with Boltzmann Selection and Kullback-Leibeler divergence. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 389–396, New York, NY, USA, 2006. ACM.

[21] Q. Zhang and H. Mühlenbein. On the Convergence of a Class of Estimation of Distribution Algorithms. *IEEE Transactions on Evolutionary Computation*, 8(2):127–136, April 2004.