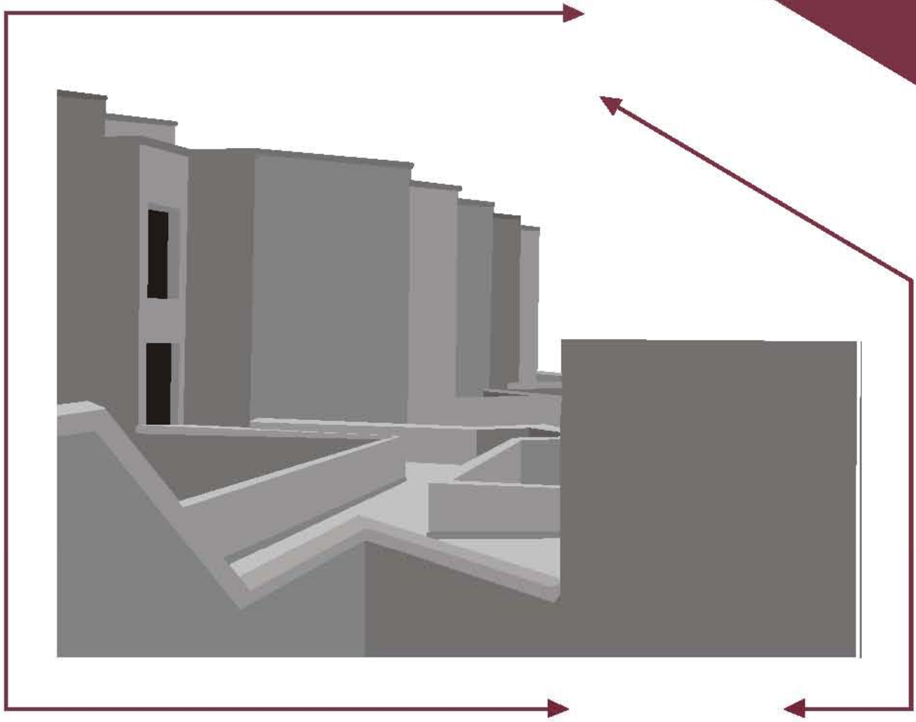


Comunicaciones del CIMAT

A BOLTZMANN BASED ESTIMATION OF DISTRIBUTION ALGORITHM

S. Ivvan Valdez, Arturo Hernández and Salvador Botello

Comunicación del CIMAT No I-08-10/25-04-2008
(CC/CIMAT)

CIMAT

A Boltzmann based Estimation of Distribution Algorithm

S. Ivvan Valdez, Arturo Hernández, and Salvador Botello
Centre for Research in Mathematics (CIMAT)
A.P. 402, C. Jalisco S/N, Guanajuato, Guanajuato,México
{ivvan,artha,botello}@cimat.mx

April 24, 2008

Abstract

The Elitist Convergent Estimation of Distribution Algorithm (ECEDA), is a definition of a class of EDA which guarantees convergence to the optimum. This paper introduces the conceptual ECEDA and a practical approach derived from it, called the Boltzmann Univariate Marginal Distribution Algorithm (BUMDA). The BUMDA uses a Gaussian model to approximate the Boltzmann distribution, requiring only one user given parameter: the population size. Several experiments and statistical analysis are used to contrast the BUMDA with state of the art EDAs.

Keywords: estimation of distribution algorithms, Boltzmann distribution, Kullback-Leibler divergence, statistical performance analysis.

1 Introduction

The Estimation of Distribution Algorithms (EDAs) were first introduced for global optimization in discrete spaces [12] [1], then several approaches were extended to continuous domains [9], [6]. Researchers have proposed general conceptual frameworks as basis for designing EDAs [10], [2]. Every framework alludes particular operating conditions, however, the quest for solutions is for the frequent question: When will an EDA perform successfully? The Elitist Convergent Estimation of Distribution Algorithm (ECEDA) is a framework defining a class of EDA which converges to the optimum assuming infinity population and generations.

The main goal of an optimizer is to find the maximum or minimum of a objective function, say $g(x)$. Population based algorithms intend to approximate the optimum by proposing a set of candidate solutions, and then spawn new individuals from a selected subset to improve the current best approximation. EDAs are population based algorithms equipped with a technique to learn a probability distribution whose main objective is to improve the optimum approximation by simulating better samples each generation.

Hereafter, without loss of generality consider a maximization problem. In order to find better solutions at each generation, a logical requirement is to increase the expectation of the objective function. This is established in Definition 1.1. Note that $\int_X$ becomes a $\sum$ in discrete cases. The ECEDA does not specify a density or probability distribution function, neither a selection method, it states that the convergence to the optimum can be achieved by increasing or maintaining the expectation of the objective function every generation.The ECEDA is called *elitist* because the probability of sampling the region containing the highest objective value is increased or at least maintained at every generation. The *convergent* characteristic is given by the Theorem 1.3.

Definition 1.1 *Consider an objective function $g(x)$, a density function $f(x)$, and sequences of consecutive generations $t = 1, 2, 3..N$, and non-consecutive generations $\tau = \tau_1, \tau_2 ... \tau_M$. An Estimation of Distribution Algorithm which fulfills:*

$$\int_X g(x)f(x,t)dx \leq \int_X g(x)f(x,t+1)dx \tag{1}$$

and

$$\int_X g(x)f(x,\tau_i)dx < \int_X g(x)f(x,\tau_{i+1})dx \tag{2}$$

For all $t \in \mathbf{N}$ and $\tau_i < \tau_{i+1} \in \mathbf{N}$, is called an **Elitist Convergent EDA (ECEDA).**

Definition 1.2 *The Gibbs or Boltzmann distribution of an objective function $g(x)$ is defined by:*

$$p(x) := \int_X \frac{\exp\left(\beta \cdot g(x)\right)}{Z} \tag{3}$$

Theorem 1.3 *Consider a sequence $\tau = \tau_1, \tau_2, ..., \tau_M$. An Elitist Convergent EDA fulfills that:*

$$\lim_{M \to \infty} E(g(x), \tau) = max\ g(x) \tag{4}$$

Proof.
By definition

$$E(g(x), \tau) = \int_X g(x)f(x,\tau)dx, \tag{5}$$

for any density function $f(x)$. By definition 1.1 the sequence of $\{E_\tau = E(g(x), \tau)\}$ is non decreasing. Also it is bounded above, with a supremum:

$$\sup E_\tau = \max g(x) = g(x^*).$$

We claim that $lim_{M\to\infty} E_\tau = g(x^*)$. If $\varepsilon > 0$, there is some E_T satisfying $g(x^*) - E_T < \varepsilon$, since $g(x^*)$ is the least upper bound of E_τ. Then if $\tau > T$ we have

$$E_\tau \geq E_T, \text{ so } g(x^*) - E_\tau \leq g(x^*) - E_T < \varepsilon.$$

This proves that $lim_{M\to\infty} E_\tau = g(x^*)$. Observe that any random global optimizer which fulfills Equation (1) also fulfills the Equation (4). And as a consequence of Equation (4), if the maximum is unique:

$$\underset{M\to\infty}{lim} P(x^* - \epsilon < X < x^* + \epsilon, \tau_M) = 1, \tag{6}$$

and

$$\underset{M\to\infty}{lim} var(x, \tau_M) = 0, \tag{7}$$

According to Definition 1.1 the ECEDA follows the steps in Figure 1.

```
Conceptual **ECEDA**

1. Assign $t = 0$, and initialize a model of probability density function
   (PDF) $f(x, t)$.
   **Repeat**

     (a) Sample $N$ candidates according $f(x, t)$.
     (b) Evaluate the candidates in the objective function $g(x)$, and
         choose the candidate $y$ with the maximum objective value as
         an approximation to the optimum, $g(y) \approx g(x^*)$.
     (c) Update $f(x, t + 1)$ such that $E(g(x), t + 1) \geq E(g(x), t + 1)$.

   **Until** termination condition is true.
2. Return $y$ as the best approximation to the optimum.
```

Figure 1: The Elitist Convergent Estimation of Distribution Algorithm (ECEDA).

By using the ECEDA framework we can derive a Boltzmann based EDA by substituting the Boltzmann probability density function, Equation (3), into Equation (1). Several well known characteristics of the Boltzmann distribution which are important when it is used in an EDA are the following:

1. When the objective function $g(x)$ increases/decreases, the **probability density function** (PDF) exponentially increases/decreases.

2. If $f(x,t)$ is the PDF, and x^* is the unique maximum, then $f(x^*,t) > f(x,t)$, for any generation t, and $x \neq x^*$.

3. The probability mass could be increased/decreased around the maxima points by simply modifying the β value.

Nevertheless, it is not possible to built efficient EDAs based on the Boltzmann distribution because in general $g(x)$ is unknown, or the cost of determining the exact Boltzmann PDF is equivalent to that of finding the optimum by exhaustive search. A common strategy to circumvent this issue is to approximate the Boltzmann PDF through a parametric PDF. For instance, Yunpeng et al. [14], approximate the integral operations by discrete summations of the variable. Thus, to compute the Kullback-Leibler divergence (KLD), the sample points are weighted by $e^{(\beta \cdot g(x))}$, with the corresponding function value $g(x)$ of each point. In Gallagher and Frean proposal [3], the goal is to find the mean of a Normal distribution (μ parameter) with fixed variance that approximates a Boltzmann with fixed temperature. They developed the analytical minimization of the KLD, and found the direction of a gradient step to be applied at every iteration. A weakness of this approach is the lack of a method to compute the variance. They used a fixed user-defined variance instead. In contrast to the mentioned approaches, particularly Yunpeng et al. work [14], we use a different KLD form. They use $KLD(Px||Qx)$ and we use $KLD(Qx||Px)$. Thus, in our approach is possible to analytically minimize the KLD, resulting in non exponential weights for the sample points which means less drastic changes during the computation of the mean and the variance. In contrast with the work of Gallagher and Frean [3], is worth notice the following differences: 1) In our approach we analytically found the expression to compute the mean which minimizes the KLD at every iteration. As mentioned, they use gradient steps to approximate the mean. 2) For our proposal the variance is not fixed, whereas theirs is fixed and user defined. We took a different approach to solve the problem and found the expression for the variances which minimizes the KLD at every iteration. The resulting proposal is called the Boltzmann Univariate Marginal Distribution Algorithm (BUMDA).

The organization of this paper is the following. Section 2 develops the formulae to approximate the Boltzmann PDF with a Normal. Section 3 explains the algorithm BUMDA. A short account of related work is given in Section 4. The Section 5 provides test problems and performance analysis for comparison with state of the art EDAs. Section 6 presents the main conclusions and discussion about the proposal presented.

2 Approximating the Boltzmann PDF with a Gaussian Model

The PDF model for independent variables is given by Equation (8). The advantages of this model are simplicity and low computational cost, not to mention the promising results reported, such as the UMDA$_c^G$ [6], PBIL [1], BG-UMDA [14], etc. A payoff for the low computational cost is the bias of the univariate model to solve problems which present weak variable correlation.

$$Q(x) = \prod_{i=1}^{n} Q_i(x_i) \tag{8}$$

In order to avoid the complexity of computing the exact Boltzmann distribution, we aim to approximate it by using the univariate normal distribution, presented in Equation (9).

$$Q_x = N(\mu, v) = \frac{1}{(2\pi v)^{1/2}} e^{\left[-\frac{(x-\mu)^2}{2v}\right]} \tag{9}$$

A widely used measure of the difference between two distributions $P(x) = P_x$ and $Q(x, \mu, v) = Q_x$ is the Kullback-Leibler divergence given in Equation (10). To approximate the normal distribution Q_x to the Boltzmann distribution P_x, we minimize the Kulback-Leibler divergence with respect to the normal parameters (μ, v).

$$K_{Q,P} = \int_x Q_x \log \frac{Q_x}{P_x} dx \tag{10}$$

Deriving $K_{Q,P}$ with respect to a model parameter θ:

$$\frac{\partial K_{Q,P}}{\partial \theta} = \int_x \left[1 + \log \frac{Q_x}{P_x}\right] \frac{\partial Qx}{\partial \theta} dx \tag{11}$$

And,

$$\log\ Q_x = -\tfrac{(x-\mu)^2}{2v} - \log(2\pi v)^{1/2}, \quad \log\ P_x = -\log\ Z + \beta g(x)$$

Substituting the logarithms into (11), we get:

$$\int_x \left[1 - \frac{(x-\mu)^{1/2}}{2v} - \log 2\pi v^{1/2} + \log Z - \beta g(x)\right] \frac{\partial Q_x}{\partial \theta} dx \tag{12}$$

The derivative of Q_x with respect μ, is:

$$\frac{\partial Q_x}{\partial \mu} = Q_x \frac{(x-\mu)}{v} \tag{13}$$

By substituting (13) into (12) we get Equation (14).

$$\begin{aligned} \frac{\partial K_{Q,P}}{\partial \mu} &= \int_x \left[1 - \frac{((x-\mu)^2)}{2v}\right] Q_x \frac{(x-\mu)}{v} dx \\ &- \int_x \left[\log 2\pi v^{1/2} - \log Z + \beta g(x)\right] Q_x \frac{(x-\mu)}{v} dx. \end{aligned} \tag{14}$$

The fact that $(x-\mu)$ is odd about μ becomes useful to evaluate the integrals, which become equal to 0. We get:

$$\frac{\partial K_{Q,P}}{\partial \mu} = -\frac{\beta}{v}\int_x Q_x(x-\mu)g(x)dx \approx -\frac{\beta}{v}\sum_{x_i \in X}(x_i-\mu)g(x_i). \tag{15}$$

Gallagher and Frean [3] used that gradient approximation and μ^t to compute μ^{t+1}. In this work we propose to directly compute the μ value which best fits the data distribution, as shown in Equation (16):

$$\tfrac{\beta}{v}\textstyle\sum_{x_i \in X}(x_i-\mu)g(x_i) = 0,$$

$$\mu \approx \frac{\sum_i g(x_i)x_i}{\sum_i g(x_i)}. \tag{16}$$

In the same way we propose to estimate the variance, as follows:

$$\frac{\partial Q_x}{\partial v} = Q_x\left(\frac{(x-v)^2}{2v^2} - \frac{1}{2v}\right). \tag{17}$$

Substituting (17) into (11):

$$\begin{aligned}\frac{\partial K_{Q,P}}{\partial v} = \int_x \left[1 + \log\frac{Q_x}{P_x}\right] Q_x \left(\frac{(x-v)^2}{2v^2} - \frac{1}{2v}\right) dx = \\ \int_x \left[1 + \log(2\pi v)^{1/2}\right] Q_x \left[\frac{(x-\mu)^2}{2v^2} - \frac{1}{2v^2}\right] dx + \\ \int_x \left[-\frac{(x-\mu)^2}{2v} + \log Z - \beta g(x)\right] Q_x \left[\frac{(x-\mu)^2}{2v^2} - \frac{1}{2v^2}\right] dx\end{aligned} \tag{18}$$

By using the equalities:

$$\textstyle\int_x Q_x(x-\mu)^2 dx = v, \;\; \int_x Q_x dx = 1, \text{ and } \int_x (x-\mu)^4 Q_x dx = \frac{3\sqrt{2}}{8}v^2,$$

the Equation (18) can be reduced to Equation (19). It is equal to 0 for minimization.

$$-\frac{3\sqrt{2}}{32v} - \frac{\beta}{2v^2}\int_x g(x)Q_x(x-\mu)^2 dx + \frac{1}{4v} + \frac{\beta}{2v}\int_x g(x)Q_x dx = 0 \tag{19}$$

Finally the expression to compute the variance is given by the next equation:

$$v = \frac{\int_x g(x)(x-\mu)^2 Q_x dx}{\frac{3+4\sqrt{2}}{8\sqrt{2}\beta} + \int_x g(x)Q_x dx} \tag{20}$$

The numerical approximation used by BUMDA for Equation (20) is the following:

$$v \approx \frac{\sum_i g(x_i)(x_i - \mu)^2}{T' + \sum_i g(x_i)} \tag{21}$$

Where:

$$T' = \frac{(3 + 4\sqrt{2})}{8\sqrt{2}\beta} n \tag{22}$$

In Equation (22) n is the sample size (selected set), it becomes involved when the integrals are approximated by summations.

3 The Boltzmann Estimation of Distribution Algorithm (BUMDA)

As mentioned in Section 1, the characteristics of an ECEDA are: the increasing expectation, and the convergence to the optimum. A simple way to ensure convergence is to apply a truncation method which increases the mean of the population, such as explained in algorithm in Figure 2.

Truncation Method

- For the initial generation $t = 0$, let be $g(x_i, 0)$ for $i = 1..N$, the objective values of the initial population. Define: $\theta_0 = \min g(x_i, 0)$.
- For $t > 0$, set:
  $\theta_t = \max\left(\theta_{t-1}, \min(g(x_i, t) | g(x_i, t) \geq \theta_{t-1})\right)$.
- If for the decreasingly sorted individuals $g(x_{N/2}) \geq \theta_t$, set $\theta_t = g(x_{N/2})$. Where N is the population size.
- Truncate the population such that $g(x_s, t) \geq \theta_t$. Where x_s are all the individuals which objective values is equal or greater than θ_t.

Figure 2: Truncation method to ensure convergence in a population based algorithm.

We must ensure that there is always at least one element in the selected set by preserving the elite individual. Now, we have all the elements needed to introduce the BUMDA, in Figure 3. A simplification for the variance computing was done by setting $T' = 1$. This means that the Boltzmann distribution is used with a fixed temperature. The reader must observe that the fixed temperature does not imply a fixed distribution, because of the following reasons:

1. The data (selected set) comes from the probability distribution we are approximating;
2. If the data change, then the underlying distribution we are approximating changes;

3. The population size and truncation method affect the magnitude of the changes of the data.

```
BUMDA

1. Give the parameter and stop criterion:
   **nsample** ← Number of individuals to be sample.
   **minvar** ← minimum variance allowed.
2. Uniformly generate the initial population $P_0$, set $t = 0$.
3. While $v > minvar$ for all dimensions
   (a) $t \leftarrow t + 1$
   (b) Evaluate and truncate the population according algorithm in Figure 2.
   (c) Compute the approximation to $\mu$ and $v$ (for all dimensions) by
       using the selected set (of size $nselec$), and Equations (16) and
       (21), as follows:
       $\mu \approx \frac{\sum_1^{nselec} x_i \bar{g}(x_i)}{\sum_1^{nselec} x_i \bar{g}(x_i)}$,
       $v \approx \frac{\sum_1^{nselec} \bar{g}(x_i)(x_i - \mu)^2}{1 + \sum_1^{nselec} \bar{g}(x_i)}$,
       where $\bar{g}(x_i) = g(x) - g(x_{nselec}) + 1$.
       Note: the individuals can be sorted to simplify the computation,
       and $g(x_{nselec})$ is the minimum objective value of the
       selected individuals.
   (d) Generate $nsample - 1$ individuals from the new model $Q(x,t)$.
       And insert the elite individual.
4. Return the elite individual as the best approximation to the optimum.
```

Figure 3: Pseudo-code for BUMDA

Several non-reported experiments, suggest that the performance is more impacted by the change of the population size than the change in the fixed value of the T' parameter.

The BUMDA presents the follow interesting advantages:

1. It converges to the best approximation to the optimum.

2. The variance tends to 0 for a large number of generations.

3. It only needs **one** parameter (population size).

4. The estimation of the parameters results in a fast automatic adaptation. The variance could be increased or decreased, according the solutions in the selected set and their objective values, and the mean moves faster to the region with best solutions found.

The first advantage listed is related with the mean of the selected set, it is lower bounded by θ_t, and upper bounded by the best approximation to the maximum $\hat{x}^*$, as θ_t tends to $\hat{x}^*$, then the mean converges to $\hat{x}^*$. The second is a consequence of the first. The last advantage is shown in Figure 4. This figure compares the Normal PDF parameters returned by our approach BUMDA (black line), and the Normal PDF parameters computed via the standard formulas (dashed line). The population is split into two clusters shown on the objective function. The first cluster contains most of the population (around $x = -1$), while the elements of the second cluster are several new promising solutions (around $x = 6$). When computing the normal parameters by using the maximum likelihood estimators of the mean and variance, we get the Gaussian density shown in Figure 4 (dashed line), which as expected has condensed the probability mass on the main cluster. This is the estimation proposed by UMDA$_c^G$ [6]. On the other hand, when computing the parameters with the Equations (16) and (21) of the BUMDA, we get the density function plotted with the black line. Observe how the mean computed by BUMDA (vertical line) is closer to the best solutions, and the variance is larger (which improves exploration).

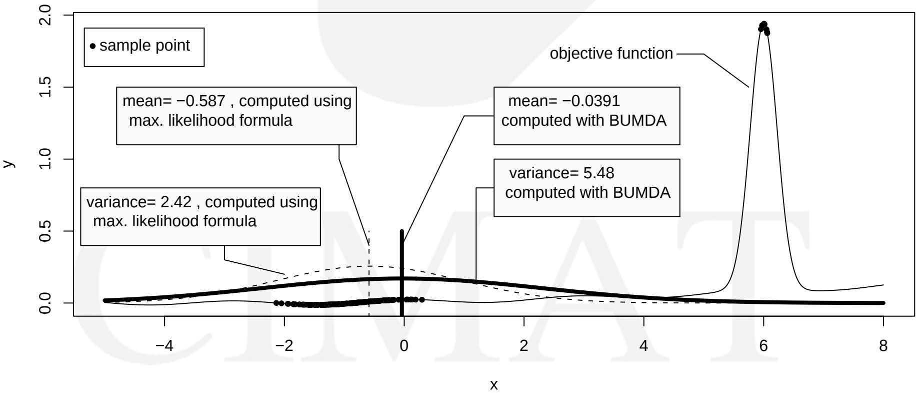


Figure 4: Comparison of BUMDA vs UMDA: the BUMDA quickly adapts its parameters to improve the exploration, when new promising solutions are found. The mean moves to promising regions and the variance is increased.

4 Related Work

Two important issues related with this work have been assessed by researchers:

1. Convergence of EDAs.
2. EDAs based on the Boltzmann distribution, and
3. parameter and structure learning.

Several lanes of work have been applied in order to obtain a convergent EDA [11, 7, 15, 5]. For example, Mühlenbein et al. [11] proposed to increase the sampling probability of the optimum. They proved the convergence of the algorithm for a specific distribution and selection in discrete spaces. Zhang and Mühlenbein [15] carried out the analysis of convergence for different selection methods. Grahl et al. [5], identified three phases during the convergence of the algorithm. For each phase they proposed an approach for the optimal sampling variance which maximizes the proportion of solutions in the optimal region.

An earlier approach of discrete EDAs based on the Boltzmann distribution, was the Boltzmann Estimation of Distribution Algorithm (BEDA) [11]. Important remarks about this work are the convergence properties, and the derivation of practical approaches such as the Factorized Distribution Algorithm (FDA) [11, 8]. The BEDA based approaches as well as the BGUMDA [14] are based on the so called Boltzmann Selection; they need an annealing schedule for the adjustment of the β parameter in Equation 1.2. Then, we can discern two main parameter learning strategies:

1. Truncation methods. They can be seen as an indicator assignment procedure: the solutions have a weight of 1 in the parameter (and structure) computing if they were selected, and 0 if they were truncated.
2. Weighting methods. Each solution have a non-binary weight for the parameter computing (this weights can be seen as an a priori probability). Normally the weights are related with the objective function.

The proposals based on the Boltzmann selection are in the second group, and the weights of the solutions are dependent on the objective function as well as the β value. The $\beta_{t+1} = \beta_t + \Delta\beta_t$ will determine the weight of the solutions in the new distribution. To compute $\Delta\beta$ researchers [8] have proposed schemes such as Equation 23.

$$\Delta\beta(t) = \beta(t+1) - \beta(t) = \frac{W_f^{new}(t) - W_f(t)}{\sigma_f^2(\beta(t))} \tag{23}$$

Where W_f is the fitness average, and σ_f^2 the fitness variance. A similar schedule for continuous variables was proposed in [14]. These models manage the selection pressure by proposing a desired fitness mean for the next generation. The weighting methods can regulate slow or premature convergence, while the truncation methods usually ensures convergence and avoid extra parameter setting. Our proposal uses both methods, the solutions are weighted according their objective function, but also we truncate the population to compute the parameters. This strategy has reduced the number of parameters without badly impacting the performance of BUMDA. A more complex strategy could be used by searching for a β value in Equation 20.

5 Test Problems and Performance Analysis

This section presents experiments and comparison among state of the art EDAs proposed by different researchers and the BUMDA. These problems test different characteristics of a continuous EDA, such as:

- The capability to find a minimum/maximum which requires a high precision, as the Sum Cancellation-like problems.
- The capability to escape from local maxima/minima, such as the presented in the Griewangk and Ackley functions.
- The capability to converge to the optimum in a low number of evaluations in convex problems such as the sphere, ellipsoid, etc. functions.
- The scalability of the algorithm, that is, how much varies the number of evaluations, or the population size when the dimensionality of the problems is increased or decreased.

The Test 1 compares the precision achieved by algorithms in the state of the art and the BUMDA. The Test 2 compares multi-modal functions and the Sum Cancellation function with well performed continuous EDAs. The Test 3 is a general comparison, they are multi-modal functions, and convex functions (sphere) to test convergence speed, and they are solved in different dimensions (10 and 50). The comparison for the third set is performed among BUMDA, the best performed EDA in other comparison (EMNA-B, [14]), and the BG-UMDA, a similar approach which uses an univariate normal distribution and the Boltzmann function. Finally the Test 4 compares principally the scalability, by plotting how the number of evaluations changes versus the dimensionality of the problem. It is important to remark that the plots and data of the algorithms we are comparing with, as well as statistical comparisons, are done considering the information in the papers where the other algorithms were introduced. The reader must take into account, that even if there are several cases when we can not (statistically) say that the BUMDA performs better, it does not mean that it performs worse.

5.1 Problem Test 1.

The first test problem set is taken from [13], in order to compare the results of BUMDA with a novel strategy which uses the Boltzmann distribution called BG-UMDA, and other state of the art continuous EDAs [14]. The mean and standard deviation are presented in Table 2, taken from reference [14]. The BUMDA results are added at the end of the table for comparison. The three functions of this test are similar to the Sum Cancellation problem, with the optimum on a high peak surrounded by a low gradient plane. These problems need high precision to get the optimum value of 1×10^7. The functions F1 and F2 have a strong variable correlation, and F3 have weak variable correlation, it means that if all variable values are maintained but one, the function could be minimized with respect the variable which is changing. Thus, as BUMDA uses a univariate model, we expect a competitive performance in F3. The results are summarized in Table 2. Note that sc-PBIL and BG-UMDA also use univariate models.

Stopping Criterion. All the algorithms were tested for 2×10^5 function evaluations.

BUMDA Parameter Setting. The unique BUMDA parameter is the population size, it is 2500 for F1 and F2, and 250 for F3.

Statistical test. Considering the number of runs for each algorithm (20), the lack of data (we only know the sample mean and standard deviation). We used the ***z test*** to find out if the BUMDA reports a better mean value of the objective function than the other approaches presented in this test. Thus, with a significance $\alpha = 0.05$ we test the null hypothesis $H_0 : \bar{F}_{BUMDA} = \bar{F}_{other\ algorithm}$, and alternative hypothesis $H_1 : \bar{F}_{BUMDA} > \bar{F}_{other\ algorithm}$. The *t-test* was not used because we can not assume variance homogeneity, according with the F_{max} test. The results of the test are presented in Table 2. If the alternative hypothesis H_1 is accepted, the BUMDA is better. Otherwise the null hypothesis is not rejected, therefore, there is not enough statistical evidence to say which algorithm is better.

Name	**Definition**	**Value to reach**
F1	$1/\left(10^{-5} + \sum_{i=1}^{n} \lvert y_i \rvert\right)$ where: $y_1 = x_1$ and $y = x_i + y_{i-1}$, for $i \geq 2$	10^5
F2	$1/\left(10^{-5} + \sum_{i=1}^{n} \lvert y_i \rvert\right)$ where: $y_1 = x_1$ and, $y = x_i + \sin y_{i-1}$, for $i \geq 2$	10^5
F3	$1/\left(10^{-5} + \sum_{i=1}^{n} \lvert y_i \rvert\right)$ where: $y = 0.024(i+1) - x_i$, for $i \geq 1$	10^5

Table 1: Test 1: Functions and values to reach.

5.2 Problem Test 2.

The second test is taken from [14], it was also used in [6] for testing continuous EDAs. This test presents a variety of characteristics, the Sum Cancellation function needs a large precision, the Schwefel function is a multimodal function with many local minima and maxima, with valleys of different sizes. The Griewangk function have many local minima and maxima, produced by a cosine mounted on a parabolic surface. The SumCan can not be precisely compared because the standard deviation is not reported in the original work.

Algorithm	F1	F2	F3	$H_1 : \bar{F}_{BUMDA} > \bar{F}_{other}$ F1	F2‘	F3
sc-PBIL	4.43 ± 0.4	7.54 ± 0.36	18.7 ± 0.63	no	no	yes, p=0
(10-50)-ES	2.91 ± 0.45	7.56 ± 1.52	399.07 ± 6.97	no	no	yes, p=0
PBILc	4.76 ± 0.78	10.99 ± 1	4083 ± 4986	no	no	yes, p=0
BG-UMDA	4.83 ± 0.57	11.32 ± 0.72	$10^7 \pm 0.0002$	no	no	no
EMNA-B	337 ± 110	326 ± 79	5.80 ± 0.99	no	no	yes, p=0
BUMDA	1.7 ± 0.15	4.94 ± 0.226	$10^7 \pm 0.003$			

Table 2: Comparison for 100-dimensional problem test (Test 1), for 20 independent runs.

Name	**Definition**	**Value to reach**
SumCan	$1/\left(10^{-5} + \sum_{i=1}^{n} \lvert y_i \rvert\right)$ where: $y_1 = x_1$ and $y = x_i + y_{i-1}$, for $i \geq 2$	10^5
Schewel	$\sum_{i=1}^{n} \{(y_1 - x_i^2)^2 + (x_i - 1)^2\}$	0
Griewangk	$1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} + \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$	0

Table 3: Test 2: Functions and values to reach.

Algorithm	SumCan	Schwefel	Griewangk	$H_1 : \bar{F}_{BUMDA}$ *better than* $\bar{F}_{other}$ SumCan	Schwefel	Griewangk
PBIL	91002 ± 28611	*unstable*	0.11 ± 0.57	no	NP	yes, p=0.027
ES	5910	0	0.034477	NP	NP	NP
UMDAc	53460	0.13754	0.011076	NP	NP	NP
EGNA	$1E5$	0.0250	0.008175	NP	NP	NP
BG-UMDA	$8E4 \pm 1.8E4$	0.009 ± 0.003	0.001 ± 0.0056	no	no	yes, p=0.037
EMNA-B	$1E5 \pm 0$	$2.7E\text{-}31 \pm 1E\text{-}31$	$5.8E\text{-}5 \pm 5.8E\text{-}4$	no	no	no
BUMDA	$7.6E3 \pm 7.9E3$	$0.233 \pm 1.9E\text{-}2$	$\mathbf{0 \pm 0}$			

Table 4: Comparison for 10-dimensional problem test (Test 2), for 100 independent runs.

The weakest variable correlation is given by the Griewangk function, observe that the weight of the cosines become important when the solutions are close to 0. Thus, we expect a good performance of the univariate EDAs in this problem. As shown in Table 4 the BUMDA reports the best performance for the Griewangk problem.

Stopping Criterion. All the algorithms were tested for 3×10^5 function evaluations.

BUMDA Parameter Setting. The population sizes for this test are 3000 for the Sum Cancellation and Schwefel functions,and 300 for Griewangk.

Statistical test. The $z - test$ with $\alpha = 0.05$ is used to compare the objective values means. The rightmost column of Table 4 shows the $z - test$ results.

5.3 Problem Test 3.

The third set of problems compares the BUMDA with the algorithms EMNA-B and BG-UMDA reported in [14] (BG-UMDA also uses a univariate Gaussian functions to approximate a Boltzmann distribution). This set of functions have been widely used to compared EDAs [7]. Some of these functions have many local maxima/minima. Also, as the functions are defined for any dimension, this set could be used to analyze the scalability of the algorithms. For this set of problems we imitate the conditions of the experiments reported in the reference [14]. The BUMDA is the most competitive approach for three problems of this set, as shown in Table 5. The BUMDA finds the best average value of the objective functions in most of the cases, but as we are using the solution error as stopping criterion, the real comparison is given by the number of function evaluations. The BUMDA uses the less average number of evaluations for all cases when it finds the solution. Observe that there is not a great difference between the number of evaluations for 10 dimensions and 50 dimensions. Even though the dimensionality was increased 5 times, the number of evaluations increased less than 3 times (when the optimum was found by BUMDA).
Stopping Criteria. All the algorithms were tested for 3×10^5 function evaluations or when they found a solution with an error less or equal to 10^{-6}.
BUMDA Parameter Setting. The population sizes for this test are 3000 for the Sum Cancellation, and 300 for all the other functions.
Statistical test. The $z - test$ with $\alpha = 0.05$ is used to compare both, objective values and function evaluations. This is the recommendable test because the only data available are the means and standard deviations. The $t - test$ should not be used, because in general the variances could not be considered homogeneous, according with the F_{max} test. The rightmost column of Tables 5 and 6 show the $z - test$ results. If the alternative hypothesis H_1 is accepted, the BUMDA is better. Otherwise the null hypothesis is not rejected, therefore, there is not enough statistical evidence to say which algorithm is better.

				$H_1 : \bar{F}_{BUMDA}$ *better than* $\bar{F}_{other}$	
Function	**BUMDA**	EMNA-B	BG-UMDA	EMNA-B	BG-UMDA
SumC 10d	$7.5E3 \pm 8.4E3$	$1E5 \pm 1.1E\text{-}7$	$5.8E4 \pm 2.3E4$	no	no
SumC 50d	2.07 ± 0.12	99910 ± 160	1.39 ± 0.1	no	yes
Griew. 10d	**7.3E-7 ± 1.7E-7**	$7.4E\text{-}7 \pm 1.1E\text{-}7$	$1.27E\text{-}4 \pm 4E\text{-}4$	no	no
Griew. 50d	**9E-7 ± 8.4E-8**	$9.2E\text{-}7 \pm 5E\text{-}8$	$8.8E\text{-}7 \pm 7E\text{-}8$	no	no
Sphere 10d	**7E-7 ± 1.6E-7**	$7.5E\text{-}7 \pm 2.1E\text{-}7$	$5.9E\text{-}7 \pm 1.8E\text{-}7$	no	no
Sphere 50d	**8.7E-7 ± 8.1E-8**	$8.8E\text{-}7 \pm 1.1E\text{-}7$	$8.4E\text{-}7 \pm 8E\text{-}8$	no	no
Rosen. 10d	8.1 ± 0.08	6.33 ± 0.37	7.74 ± 0.08	no	no
Rosen. 50d	47.7 ± 0.18	47.08 ± 0.44	47.54 ± 0.07	no	no
Ackley 10d	**8.3E-7 ± 1.2E-7**	$8.4E\text{-}7 \pm 1E\text{-}7$	$8.3E\text{-}7 \pm 1.6E\text{-}7$	no	no
Ackley 50d	**9.3E-7 ± 4.3E-8**	$9.42E\text{-}7 \pm 4E\text{-}8$	$9.6E\text{-}7 \pm 4E\text{-}8$	no	yes

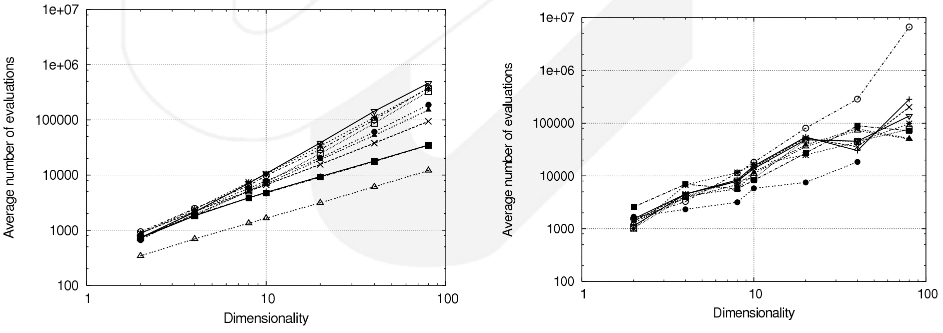
Table 5: Mean and standard deviation of best function value found in 20 runs for the Test 3.

Function	**BUMDA**	EMNA-B	BG-UMDA	$H_1 : \bar{N}^{eval}_{BUMDA} < \bar{N}^{eval}_{other}$ EMNA-B	 BG-UMDA
SumCan 10d	$3E5 \pm 0$	92520 ± 840	300400 ± 0	NP	NP
SumCan 50d	$3E5 \pm 0$	301000 ± 0	300400 ± 0	NP	NP
Griewangk 10d	$\mathbf{17262 \pm 384}$	134000 ± 47000	$229E3 \pm 64E3$	yes, p=5.8E-29	yes, p=7.8E-50
Griewangk 50d	$\mathbf{39675 \pm 342}$	170100 ± 1700	71880 ± 420	yes, p=0	yes, p=0
Sphere 10d	$\mathbf{14541 \pm 261}$	35200 ± 420	35720 ± 840	yes, p=0	yes, p=0
Sphere 50d	$\mathbf{40695 \pm 325}$	192900 ± 1600	82400 ± 460	yes, p=0	yes, p=0
Rosenbrock 10d	$3E5 \pm 0$	300400 ± 0	300400 ± 0	NP	NP
Rosenbrock 50d	$3E5 \pm 0$	301000 ± 0	300400 ± 0	NP	NP
Ackley 10d	$\mathbf{23257 \pm 287}$	43560 ± 610	44000 ± 530	yes, p=0	yes, p=0
Ackley 50d	$\mathbf{58850 \pm 348}$	231800 ± 4300	98920 ± 530	yes, p=0	yes, p=0

Table 6: Average and standard deviation of evaluations for Test 3.

5.4 Problem test 4

This set of problems is taken from [4]. All the functions are convex and have been generalized for any number of dimensions. Most of these problems can be solved by well performed EDAs as the presented in [4]. Then, an objective comparison of scalability must relate the number of evaluations with problem dimensionality.



(a) (b)

Figure 5: (a)CMA-ES plot for Average Number of Evaluations vs Dimensionality. (b)CT-AVS-IDEA plot for Average Number of Evaluations vs Dimensionality.

For these problems we report a plot of the problem dimensionality (2, 4, 8, 10, 20, 40, 80) versus the average number of evaluations (to preserve the experimental conditions of the results presented in [4], and because of numerical results were not reported). The comparison includes well performed algorithms reviewed in [4]: the Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), and the Correlation-Triggered Adaptive Variance Scaling IDEA (CT-AVS-IDEA). The BUMDA successfully solves 30 independent consecutive runs for all the test problems except the Rosenbrock. Figure 5(a) presents the results of the CMA-ES, as shown the number of evaluations are in the range of $1 \times 10^2 < evaluations < 1 \times 10^6$. They are linearly proportional to the dimensionality in the log scale, and sub-quadratically in the linear scale. Figure 5(b) shows that the number of evaluations of CT-AVS-IDEA exponentially grows with the dimensionality. On the other hand, the BUMDA plot in Figure 6 shows linear behavior. Although BUMDA uses an univariate model the linear behavior is not necessarily expected. Because the data are projected in each dimension, and the objective function on the projected data becomes noisy, and the noise increases when the dimensionality is increased. The symbols used to represented the different test problems are: **Cigar + , Cigar tablet ×, Different powers *, Ellipsoid □, Parabolic Ridge ■, Rosenbrock ○(not presented for BUMDA) , Sharp Ridge •, Sphere △, Tablet ▲, Two axes ▽.**

Stopping Criteria. All the algorithms use the closeness to the optimum as termination criterion, it was set in $10^{-}10$ for all the functions except the different powers function which optimum closeness was set in $10^{-}15$.

BUMDA Parameter Setting. The population sizes for this test are 3000 for the Sum Cancellation,and 300 for all the other functions.

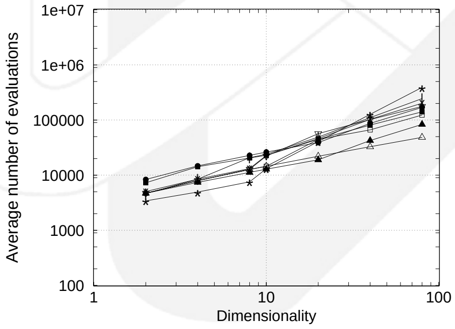


Figure 6: BUMDA plot for Average Number of Evaluations vs Dimensionality.

Parameter Setting and Comments about Tests. As expected the BUMDA performs very well in the problems which presents a weak variable correlation. When comparing the BUMDA with similar approaches (as BG-UMDA) it is very competitive (or the best in many cases as shown in Test 3). In addition, when BUMDA is compared with approaches that uses multivariate models as EMNA-B or CT-AVS-IDEA it is competitive or better as in Test 3 and 4 for most of the cases. Observe that BUMDA have a low computational cost (linear), and for the set of problems presented here, the number of evaluations needed to find the optimum grows slow with respect to the number of dimensions. The BUMDA only needs the population size as user given parameter. In the case when the optimum is known, this parameter could be set in a straightforward way, increasing the population size until the best optimum approximation is found or the performance does not change. When the optimum is not known a good indicator of the population size needed is the number of selected individuals by the truncation method: if less than the 50% of the population is selected in most of the generations, then the population size must be increased. The algorithm in Figure 3 uses a *minvar* value as stop criterion, this value stops the algorithm when a poor exploration is detected, and the optimum approximation difficultly will be improved.

6 Conclusions

This paper presents a novel proposal for designing EDAs called *the Elitist Convergent EDA (ECEDA)*, this conceptual proposal principally indicates the characteristics of a successful EDA, such as: convergence to the optimum, high probability of return solutions very closed to the optimum for a large number of generations, and the tendency of the variance to 0. These characteristics are achieved if we ensure that the expectation of the objective function is maintained/increased every generation. As it is not possible to built efficient ideal ECEDAs, we present a practical proposal: the Boltzmann Univariate Marginal Distribution Algorithm (BUMDA), which ensures convergence to the best solution found for a large number of (increasing-expectation) generations. This proposal uses a Gaussian model to approximate a Boltzmann distribution which is one of the most important functions in global random optimization, because its special characteristics mentioned in Section 1. The BUMDA algorithm can solve an extensive type of problems with a very competitive effort (number of evaluations). Also, the computational cost required to calculate the parameters of the probabilistic model is $O(nm)$ (linear) with the number of dimensions n, and the population size m. One of the most important goals of the Estimation of Distribution Algorithms is the reduction of user-given parameters, the BUMDA requires just ***one*** user-given parameter which can be easily tuned: the population size. We suggest to use a minimum variance as stopping criterion, this value detects when the algorithm have a poor exploration and the optimum approximation could be rarely improved. The Test problems presented in Section 5, are used to contrast BUMDA with many other approaches in the state of the art, the results indicate that the BUMDA is a very competitive algorithm when it is compared with different approaches which use univariate models such as BGUMDA, PBILc, and sc-PBIL presented in Test 1,2 and 3, as well as multivariate models such as: EMNA-B, CT-AVS-IDEA and CMA-ES, presented in Test 3 and 4 respectively. Future work will contemplate the approximation of the Boltzmann distribution by more complex model which capture dependencies among variables.

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 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.

[3] M. Gallagher[1] and M. Frean[2]. Population-based continuous optimization and probabilistic modelling. Technical report, [1]School of computer Science and Electrical Engineering, Univerity of Queensland, Australia. [2]School of Mathematical and Computing Sciences, Victoria Univerty, New Zealand., University of Queensland. 4072 Australia., 2001.

[4] 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.

[5] 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.

[6] 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.

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

[8] T. Mahnig and H. Mühlenbein. Comparing the adaptive boltzmann selection schedule sds to truncation selection. In *Proceedings of the Third International Symposium on Adaptive Systems ISAS 2001, Evolutionary Computation and Probabilistic Graphical Models*, pages 121–128, La Habana, Cuba, 2001.

[9] H. Müehlenbein[1], J. Bendisch[1], and H.-M. Voight[2]. From recombination of genes to the estimation of distributions ii. continuous parameters, 1996.

[10] H. Mühlenbein. The equation for response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346, 1997.

[11] H. Mühlenbein, T. Mahnig, and A. O. Rodriguez. Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics*, 5(2):215–247, 1999.

[12] 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.

[13] M. Sebag and A. Ducoulombier. Extending population-based incremental learning to continuous search spaces. In *PPSN V: Proceedings of the 5th International Conference on Parallel Problem Solving from Nature*, pages 418–427, London, UK, 1998. Springer-Verlag.

[14] 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.

[15] 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.