Analyzing the Population Based Incremental Learning Algorithm by Means of Discrete Dynamical Systems

Cristina González[*]
Jose A. Lozano
Pedro Larrañaga
Intelligent Systems Group,
Department of Computer Science and Artificial Intelligence,
University of the Basque Country, Spain

In this paper the convergence behavior of the population based incremental learning algorithm (PBIL) is analyzed using discrete dynamical systems. A discrete dynamical system is associated with the PBIL algorithm. We demonstrate that the behavior of the PBIL algorithm follows the iterates of the discrete dynamical system for a long time when the parameter α is near zero. We show that all the points of the search space are fixed points of the dynamical system, and that the local optimum points for the function to optimize coincide with the stable fixed points. Hence it can be deduced that the PBIL algorithm converges to the global optimum in unimodal functions.

1. Introduction

During the 1990s many real combinatorial optimization problems were solved successfully by means of genetic algorithms (GAs). But the existence of deceptive problems, where the performance of GAs is very poor, has motivated the search for new optimization algorithms. To overcome these difficulties a number of researches have recently suggested a family of new algorithms called estimation of distribution algorithms (EDAs) [1, 2].

Introduced by Mühlenbein and Paaβ [2], EDAs constitute an example of stochastic heuristics based on populations of individuals, each of which encodes a possible solution of the optimization problem. These populations evolve in successive generations as the search progresses, organized in the same way as most evolutionary computation heuristics. In contrast to GAs, which consider the crossover and mutation operators as essential tools to generate new populations, EDAs replace

[*] Electronic mail address: ccpgomoc@si.ehu.es.

those operators by the estimation and sampling of the joint probability distribution of the selected individuals.

However, the bottleneck of this new heuristic lies in estimating the joint probability distribution associated with the database containing the selected individuals. To avoid this problem, several authors have proposed different algorithms where simplified assumptions concerning the conditional (in)dependencies between the variables of the joint probability distribution are made. A review of the different approaches in the combinatorial and numerical fields can be found in [3–5].

The population based incremental learning algorithm (PBIL) can be considered as an EDA, as proposed in [6]. PBIL supposes that all the variables are independent. At each step of the algorithm a probability vector is maintained. This vector is sampled λ times to obtain λ new solutions. The $\mu \leq \lambda$ best solutions are selected and these are used to modify the probability vector with a neural networks-inspired rule.

Recently, there has been increasing interest in PBIL and many papers have appeared in the literature. Some of these papers are concerned with applications of the algorithm [7–11]. Others are concerned with extensions of the method to general cardinality spaces [12], with continuous spaces [13–16], with parallel versions [17], and with comparisons [18, 19]. Finally there are papers concerned with modifications of the method: adding genetics information [20] and extending the algorithm to nonstatic multiobjective problems [21]. However, despite the effort devoted to applications or to creating new variants, little attention has been given to the theoretical aspects of PBIL.

In this paper we introduce a new framework for studying the PBIL algorithm theoretically. We assign a discrete dynamical system to PBIL. It will be shown that the behavior of PBIL follows the iterations of the discrete dynamical system when an algorithm's parameter α is near zero. This fact enables us to study the discrete dynamical system instead of the iterations of PBIL. We discover that all the points of the search space are fixed points for the discrete dynamical system. Moreover, the local optima are stable fixed points and the other points of the search space are unstable fixed points. This result has various outcomes, the most important of which is that PBIL converges to the global optimum in unimodal functions.

The remainder of this work is organized as follows. Section 2 describes the PBIL algorithm. Section 3 reviews the work carried out in the theoretical analysis of PBIL. Section 4 introduces the discrete dynamical system associated with PBIL. The relation between PBIL and the dynamical system is analyzed in section 5, while the dynamical system is studied in section 6. Finally we draw conclusions in section 7.

2. An introduction to the population based incremental learning algorithm

Before presenting the algorithm, let us introduce some notation. We denote a vector or a matrix by a bold-face letter and a component of a vector by a normal letter. The random variables will be written in capital letters. We use the letters m and r as component indexes, and the letters i and k as vector indexes, hence $y_{i,m}$ will represent the mth component of the $\mathbf{y}_i$ individual.

PBIL was introduced by Baluja [6] in 1994 and further improved by Baluja and Caruana [22] in 1995. This algorithm is based on the idea of substituting the individuals of a population by a set of their statistics. In our case we suppose that the function to optimize is defined in the binary space $\Omega = \{0, 1\}^l$ with $|\Omega| = 2^l = n$. Given a population, the set of statistics is expressed by a vector of probabilities $\mathbf{p} = (p_1, \ldots, p_m, \ldots, p_l)$, where p_m represents the probability of obtaining a value of 1 in the mth component.

The algorithm works as follows. At each step, drawing the probability vector $\mathbf{p}$, λ individuals are obtained and the μ best of them ($\mu \leq \lambda$), $\mathbf{y}_{1:\lambda}, \mathbf{y}_{2:\lambda}, \ldots, \mathbf{y}_{\mu:\lambda}$, are selected. These selected individuals will be used to modify the probability vector. A Hebbian-inspired rule is used to update the probability vector:

$$\mathbf{p}^{(t+1)} = (1 - \alpha)\mathbf{p}^{(t)} + \alpha \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{y}_{k:\lambda}^{(t)} \tag{1}$$

where $\mathbf{p}^{(t)}$ is the probability vector at step t, and $\alpha \in (0, 1]$ is an algorithm's parameter. Figure 1 shows a pseudocode for the PBIL algorithm.

Once we have seen Figure 1, it is useful to note that the behavior of the algorithm is the same in two functions f_1 and f_2 if:

$$\forall\, \mathbf{y}, \quad \mathbf{y}' \in \Omega, \quad f_1(\mathbf{y}) > f_1(\mathbf{y}') \Leftrightarrow f_2(\mathbf{y}) > f_2(\mathbf{y}'). \tag{2}$$

```
Obtain an initial probability vector $\mathbf{p}^{(0)}$
while no convergence do
  begin
  Using $\mathbf{p}^{(t)}$ obtain $\lambda$ individuals $\mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}, \ldots, \mathbf{y}_\lambda^{(t)}$
  Evaluate and rank $\mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}, \ldots, \mathbf{y}_\lambda^{(t)}$
  Select the $\mu \leq \lambda$ best individuals $\mathbf{y}_{1:\lambda}^{(t)}, \mathbf{y}_{2:\lambda}^{(t)}, \ldots, \mathbf{y}_{\mu:\lambda}^{(t)}$
  $\mathbf{p}^{(t+1)} = (1 - \alpha)\mathbf{p}^{(t)} + \alpha 1/\mu \sum_{k=1}^{\mu} \mathbf{y}_{k:\lambda}^{(t)}$
  end
```

Figure 1. Pseudocode for the PBIL algorithm.

Moreover the important thing is not the particular value that a function f has in an individual $\mathbf{y}$ but the ranking implied by this function on Ω. Below we assume that we have a minimization problem. Keeping in mind the previous argument we can consider a function

$$f : \Omega \to \mathbb{R}$$

as a permutation of the elements of Ω. In this case the last individual of Ω, $\mathbf{y}_n$ has the smallest function value, the penultimate $\mathbf{y}_{n-1}$ the second smallest, and so on, with the first individual $\mathbf{y}_1$ being the one with the biggest function value:

$$f(\mathbf{y}_1) \geq \cdots \geq f(\mathbf{y}_{n-1}) \geq f(\mathbf{y}_n). \tag{3}$$

Therefore the number of different injective functions on Ω is given by $n!$. In order to simplify notation, we assume in the rest of the paper that we have to optimize an injective function. However, the results can easily be extended to noninjective functions.

3. Previous approaches to modeling population based incremental learning mathematically

In the literature we have found different ways to mathematically model PBIL.

It can be modeled by means of Markov chains given that the probability vector $\mathbf{p}^{(t+1)}$ at step $t+1$ only depends on the probability vector $\mathbf{p}^{(t)}$ at step t. However, this chain has infinite numerable states once an initial probability vector $\mathbf{p}^{(0)}$, a value for the parameter α, and values for λ and μ have been established. In that case the chain is neither irreducible nor aperiodic, hence not much information can be extracted. However, using this model it is shown in [23] that for PBIL with $\lambda = 2$ and μ=1 applied to the OneMax function in two dimensions:

$$P\left(\lim_{t\to\infty} \mathbf{p}^{(t)} = (a,b)\right) \to 1 \tag{4}$$

when $\alpha \to 1$, $\mathbf{p}^{(0)} \to (a,b)$, and $(a,b) \in \{0,1\}^2$. This shows a strong dependence of PBIL on the initial parameters.

In [24] it is shown that if we denote by $\mathsf{E}[\mathbf{p}^{(t)}]$ the expectation of the probability vector in time t, then for a linear function:

$$\lim_{t\to\infty} \mathsf{E}[\mathbf{p}^{(t)}] = \mathbf{y}_* \tag{5}$$

where $\mathbf{y}_*$ is the optimum point of Ω.

In [25] a statistical framework for combinatorial optimization over fixed-length binary strings is presented and it is shown that the PBIL algorithm can be derived from a gradient dynamical system acting on Bernoulli probability vectors.

When $\alpha = 1$, a particular case of PBIL, the univariate marginal distribution algorithm (UMDA) [26] is found. In [27] it is shown that UMDA with infinite population and proportional selection stops at local optima.

4. Assigning a discrete dynamical system to population based incremental learning

A new approach can be opened if we model PBIL by means of discrete dynamical systems. This idea has been used previously for the simple GA [28, 29] obtaining important results. Our approach and notation are strongly inspired by these previous works. The key question is to associate PBIL with a discrete dynamical system, such that, the trajectories followed by the probability vectors $\{\mathbf{p}^{(t)}\}_{t=0,1,2,\ldots}$ in PBIL are related to the iterations of the discrete dynamical system.

PBIL can be considered as a sequence of probability vectors, each one given by a transition stochastic rule τ:

$$\mathbf{p}^{(0)} \xrightarrow{\tau} \mathbf{p}^{(1)} \xrightarrow{\tau} \mathbf{p}^{(2)} \xrightarrow{\tau} \cdots \xrightarrow{\tau} \mathbf{p}^{(t)} \xrightarrow{\tau} \mathbf{p}^{(t+1)} \xrightarrow{\tau} \cdots$$

that is, $\mathbf{p}^{(t+1)} = \tau(\mathbf{p}^{(t)}) = \tau^{t+1}(\mathbf{p}^{(0)})$. We are interested in the trajectories followed by the iterations of τ, and in particular, in its limit behavior:

$$\lim_{t\to\infty} \tau^t(\mathbf{p}^{(0)}). \tag{6}$$

We define a new operator $\mathcal{G}$:

$$\mathcal{G} : [0,1]^l \to [0,1]^l$$

such that $\mathcal{G}(\mathbf{p}) = (\mathcal{G}_1(\mathbf{p}), \ldots, \mathcal{G}_l(\mathbf{p})) = \mathsf{E}[\tau(\mathbf{p})]$. The operator $\mathcal{G}$ is a deterministic function that gives the expected value of random operator τ. The iterations of $\mathcal{G}$ are defined as $\mathcal{G}^t(\mathbf{p}) = \mathcal{G}(\mathcal{G}^{t-1}(\mathbf{p}))$. We are interested in the relation between the iterations of τ and the iterations of $\mathcal{G}$, and in particular we want to answer the question: Is there any relation between $\tau^t(\mathbf{p})$ and $\mathcal{G}^t(\mathbf{p})$? However, before looking for this relation we calculate the expression of $\mathcal{G}(\mathbf{p})$.

The operator $\mathcal{G}$ can be expressed as follows:

$$\begin{aligned}\mathcal{G}(\mathbf{p}) &= \mathsf{E}\left[(1-\alpha)\mathbf{p} + \alpha\frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{Y}_{k:\lambda}\right]\\ &= (1-\alpha)\mathbf{p} + \alpha\frac{1}{\mu}\mathsf{E}\left[\sum_{k=1}^{\mu}\mathbf{Y}_{k:\lambda}|\mathbf{p}\right].\end{aligned} \tag{7}$$

Hence, we have to calculate the expected sum of the μ best individuals given that λ have been sampled from the probability vector $\mathbf{p}$. Our analysis is restricted to the case $\mu = 1$ (the most frequent in the literature).

In this case the $\mathcal{G}$ operator can be calculated explicitly, which we do in several steps.

First, the expected value $\mathsf{E}[\sum_{k=1}^{\mu} \mathbf{Y}_{k:\lambda}|\mathbf{p}]$ is reduced for the case $\mu = 1$ to:

$$\mathsf{E}[\mathbf{Y}_{1:\lambda}|\mathbf{p}] = \sum_{i=1}^{n} \mathbf{y}_i P(\mathbf{Y}_{1:\lambda} = \mathbf{y}_i \mid \mathbf{p}) \tag{8}$$

where $P(\mathbf{Y}_{1:\lambda} = \mathbf{y}_i \mid \mathbf{p})$ denotes the probability of obtaining $\mathbf{y}_i$ as the best individual after an iteration of the algorithm. Let $P(\mathbf{S} = \mathbf{y}_i \mid \mathbf{p})$ denote the probability of sampling vector $\mathbf{y}_i$. The probability that $\mathbf{y}_i$ is the best individual, given that we have sampled λ individuals and the probability vector is $\mathbf{p}$, can be expressed as follows [24]:

$$P(\mathbf{Y}_{1:\lambda} = \mathbf{y}_i \mid \mathbf{p}) = P(\mathbf{S} = \mathbf{y}_i \mid \mathbf{p}) \sum_{k=1}^{\lambda} P(\Omega_i^{>} \mid \mathbf{p})^{k-1} P(\Omega_i^{\geq} \mid \mathbf{p})^{\lambda-k}, \tag{9}$$

where:

$$\Omega_i^{>} = \{\mathbf{y}_j \in \Omega \mid f(\mathbf{y}_j) > f(\mathbf{y}_i)\} \tag{10}$$

$$\Omega_i^{\geq} = \{\mathbf{y}_j \in \Omega \mid f(\mathbf{y}_j) \geq f(\mathbf{y}_i)\}. \tag{11}$$

Finally, it is important to note that, given a probability vector $\mathbf{p}$, the probability of sampling a particular individual $\mathbf{y} = (y_1, \ldots, y_m, \ldots, y_l)$ is:

$$q_{\mathbf{y}}(\mathbf{p}) = P(\mathbf{S} = \mathbf{y} \mid \mathbf{p}) = \prod_{m=1}^{l} p_m^{y_m} (1 - p_m)^{1-y_m}. \tag{12}$$

According to the previous results it can be said that the operator $\mathcal{G}$ can be expressed as a polynomial function of the probability vector $\mathbf{p}$. In fact, if we take into account that each function can be considered as a permutation of Ω (only the individuals preceding $\mathbf{y}_i$ have a bigger function value than $\mathbf{y}_i$), then:

$$P(\Omega_i^{>} \mid \mathbf{p}) = \sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{p}) \tag{13}$$

$$P(\Omega_i^{\geq} \mid \mathbf{p}) = \sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{p}). \tag{14}$$

Finally $\mathcal{G}(\mathbf{p})$ can be expressed in our study case ($\mu = 1$) as follows:

$$\begin{aligned}\mathcal{G}(\mathbf{p}) &= (1-\alpha)\mathbf{p} \\ &+ \alpha \sum_{i=1}^{n} \mathbf{y}_i q_{\mathbf{y}_i}(\mathbf{p}) \left(\sum_{k=1}^{\lambda} \left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{k-1} \left(\sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{\lambda-k} \right).\end{aligned} \tag{15}$$

5. Relationship between $\tau^t(\mathbf{p})$ and $\mathcal{G}^t(\mathbf{p})$

In this section we demonstrate that when the parameter α is near 0, then the stochastic operator τ follows the deterministic operator $\mathcal{G}$ for a long time. First we set up the relation between $\mathcal{G}$ and τ and then we study the relation between their iterates.

Lemma 1. Given $\epsilon > 0$ and $\gamma < 1$, there exists $\alpha_0 > 0$ independent of the probability vector $\mathbf{p}$ such that with probability at least γ:

$$\alpha < \alpha_0 \Rightarrow \|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| < \epsilon. \tag{16}$$

Proof. Suppose that $\tau(\mathbf{p}) = (1-\alpha)\mathbf{p} + \alpha\mathbf{y}$. The discrepancy between $\mathcal{G}$ and τ can be bounded by:

$$\begin{aligned}\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| &= \|(1-\alpha)\mathbf{p} + \alpha\mathbf{y} - (1-\alpha)\mathbf{p} - \alpha\mathsf{E}[\mathbf{Y}_{1:\lambda}|\mathbf{p}]\| \\ &= \alpha\|\mathbf{y} - \mathsf{E}[\mathbf{Y}_{1:\lambda}|\mathbf{p}]\| \leq \alpha l.\end{aligned} \tag{17}$$

Since the right-hand side goes to zero as $\alpha \to 0$ the proof is completed. ■

Theorem 1. Given $k > 0$, $\epsilon > 0$, and $\gamma < 1$, there exists α_0 such that with probability at least γ and for all $0 \leq t \leq k$:

$$\alpha < \alpha_0 \Rightarrow \|\tau^t(\mathbf{p}) - \mathcal{G}^t(\mathbf{p})\| < \epsilon. \tag{18}$$

Proof. We make the proof by induction on k. The base case $k = 1$ coincides with Lemma 1. Given that $\mathcal{G}$ is uniformly continuous (it is continuous in the compact $[0,1]^l$), choose δ such that

$$\|\tau^{k-1}(\mathbf{p}) - \mathcal{G}^{k-1}(\mathbf{p})\| < \delta \Rightarrow \|\mathcal{G}(\tau^{k-1}(\mathbf{p})) - \mathcal{G}(\mathcal{G}^{k-1}(\mathbf{p}))\| < \frac{\epsilon}{2}. \tag{19}$$

By the inductive hypothesis, if $\alpha < \alpha_1$ then with probability at least $1 - (1-\gamma)/2$ we have

$$\|\tau^{k-1}(\mathbf{p}) - \mathcal{G}^{k-1}(\mathbf{p})\| < \delta. \tag{20}$$

By Lemma 1 (applied to $\tau^{k-1}(\mathbf{p})$ instead of $\mathbf{p}$) let α_2 be such that with probability at least $1 - (1-\gamma)/2$

$$\alpha < \alpha_2 \Rightarrow \|\tau^k(\mathbf{p}) - \mathcal{G}(\tau^{k-1}(\mathbf{p}))\| < \frac{\epsilon}{2}. \tag{21}$$

It follows that if $\alpha < \alpha_0 = \min\{\alpha_1, \alpha_2\}$, then with probability at least γ

$$\begin{aligned}\|\tau^k(\mathbf{p}) - \mathcal{G}^k(\mathbf{p})\| \leq\ & \|\mathcal{G}(\tau^{k-1}(\mathbf{p})) - \mathcal{G}^k(\mathbf{p})\| \\ & + \|\tau^k(\mathbf{p}) - \mathcal{G}(\tau^{k-1}(\mathbf{p}))\| < \frac{\epsilon}{2} + \frac{\epsilon}{2}. \blacksquare\end{aligned} \tag{22}$$

Theorem 1 means that when α is near 0, the stochastic operator τ follows, with high probability and for a long time, the iterations of the deterministic operator $\mathcal{G}$.

The operator $\mathcal{G}$ can be thought of as a discrete dynamical system:

$$\mathbf{p}, \mathcal{G}(\mathbf{p}), \ldots, \mathcal{G}^t(\mathbf{p}), \ldots.$$

6. The discrete dynamical system $\mathcal{G}$

In this section we try to find properties of the discrete dynamical system $\mathcal{G}$ that will give some information concerning the behavior of the PBIL algorithm. Before studying the discrete dynamical system $\mathcal{G}$, it is important to note that the following discrete dynamical system $\overline{\mathcal{G}}$:

$$\overline{\mathcal{G}}(\mathbf{p}) = \sum_{i=1}^{n} \mathbf{y}_i q_{\mathbf{y}_i}(\mathbf{p}) \left(\sum_{k=1}^{\lambda} \left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{k-1} \left(\sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{\lambda-k} \right) \tag{23}$$

has the same behavior as $\mathcal{G}$ in the points of the search space. All the results obtained here for $\overline{\mathcal{G}}$ are valid for $\mathcal{G}$. In particular they have the same singular points, so we study the dynamical system of equation (23) instead of the operator $\mathcal{G}$.

Theorem 2. All the points of Ω are fixed points of $\overline{\mathcal{G}}$.

Proof. Given $\mathbf{y} \in \Omega$, clearly $\mathsf{E}[\mathbf{Y}_{1:\lambda} | \mathbf{p} = \mathbf{y}] = \mathbf{y}$, because the probability of sampling an individual different from $\mathbf{y}$ given $\mathbf{p} = \mathbf{y}$ is zero. Hence:

$$\overline{\mathcal{G}}(\mathbf{y}) = \mathsf{E}[\mathbf{Y}_{1:\lambda} | \mathbf{p} = \mathbf{y}] = \mathbf{y}. \blacksquare \tag{24}$$

Before introducing Theorem 3, we define what we mean by a "local optimum" for the Hamming distance. We also give a result borrowed from page 126 in [30], that will be useful in the proof of the theorem.

Definition 1. Given a real function f defined in Ω, a point $\mathbf{y}$ is a *local minimum* for the Hamming distance d_H if:

$$\text{for all } \mathbf{y}' \text{ such that } d_H(\mathbf{y}', \mathbf{y}) = \sum_{m=1}^{l} |y'_m - y_m| = 1 \Rightarrow f(\mathbf{y}') \geq f(\mathbf{y}). \tag{25}$$

Lemma 2. Let $\mathbf{x}$ be a fixed point of a discrete dynamical system $\mathcal{G}$.

- If the eigenvalues of $D\mathcal{G}(\mathbf{x})$ all have absolute value less than 1 then $\mathbf{x}$ is a stable fixed point of $\mathcal{G}$.
- If some eigenvalue of $D\mathcal{G}(\mathbf{x})$ has absolute value greater than 1 then $\mathbf{x}$ is an unstable fixed point of $\mathcal{G}$.

Theorem 3. Given a real function f defined on Ω, we have the following.

- All the local optima of f with respect to the Hamming distance are stable fixed points of $\overline{\mathcal{G}}$.

- All the nonlocal optima of f with respect to the Hamming distance are unstable fixed points of $\overline{\mathcal{G}}$.

Proof. We use Lemma 2 in order to prove these affirmations.

We first show that all the local optima are stable points; and, in a second step, the last result.

Let $\mathbf{y} \in \Omega$ be a local optimum of a function f (i.e., all the individuals in Ω whose Hamming distance from $\mathbf{y}$ is one, precede $\mathbf{y}$ in the order imposed by f in Ω). We will see that in this case $D\overline{\mathcal{G}}|_{\mathbf{y}} = \mathbf{0}$. In fact, we will see that

$$\left.\frac{\partial \overline{\mathcal{G}}_r}{\partial p_m}\right|_{\mathbf{y}} = 0 \text{ for all } r, m = 1, 2, \ldots, l, \tag{26}$$

where $\overline{\mathcal{G}}_r$ is the rth component of equation (23):

$$\overline{\mathcal{G}}_r(\mathbf{p}) = \sum_{i=1}^{n} y_{i,r} q_{\mathbf{y}_i}(\mathbf{p}) \left(\sum_{k=1}^{\lambda} \left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{k-1} \left(\sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{p}) \right)^{\lambda-k} \right). \tag{27}$$

To show equation (26), we must first take into account the following results which can be checked easily using the definition of $q_{\mathbf{y}}(\mathbf{p})$ (equation (12)):

$$q_{\mathbf{y}_k}(\mathbf{y}) = 0 \text{ for all } \mathbf{y} \neq \mathbf{y}_k \tag{28}$$

$$q_{\mathbf{y}_k}(\mathbf{y}_k) = 1 \text{ for all } \mathbf{y}_k \in \Omega \tag{29}$$

$$\left.\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\right|_{\mathbf{y}_k} = \begin{cases} 1 & \text{if } y_{k,m} = 1 \\ -1 & \text{if } y_{k,m} = 0 \end{cases} \tag{30}$$

$$\left.\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\right|_{\mathbf{y}} = 0 \text{ for all } \mathbf{y} \text{ such that } d_H(\mathbf{y}, \mathbf{y}_k) \geq 2 \tag{31}$$

$$\left.\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\right|_{\mathbf{y}} = \begin{cases} 1 & \text{if } d_H(\mathbf{y}, \mathbf{y}_k) = 1 \text{ and } y_{k,m} = 1, y_m = 0 \\ -1 & \text{if } d_H(\mathbf{y}, \mathbf{y}_k) = 1 \text{ and } y_{k,m} = 0, y_m = 1 \end{cases} \tag{32}$$

$$\left.\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\right|_{\mathbf{y}} = 0 \text{ if } d_H(\mathbf{y}, \mathbf{y}_k) = 1 \text{ and } y_{k,m} = y_m. \tag{33}$$

The partial derivative of $\overline{\mathcal{G}}_r$ with respect to p_m can be expressed as:

$$\left.\frac{\partial \overline{\mathcal{G}}_r}{\partial p_m}\right|_{\mathbf{y}} = \sum_{i=1}^{n} \frac{\partial}{\partial p_m} \left[y_{i,r} q_{\mathbf{y}_i} \left(\sum_{k=1}^{\lambda} \left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j} \right)^{k-1} \left(\sum_{j=1}^{i} q_{\mathbf{y}_j} \right)^{\lambda-k} \right) \right]\Bigg|_{\mathbf{y}}. \tag{34}$$

We analyze each adding term of $\partial\overline{\mathcal{G}}_r/\partial p_m|_{\mathbf{y}}$ separately and split it into three different cases. The partial derivative of a term of equation (34) must be written first ($y_{i,r}$ has been eliminated, because it is a constant term):

$$\left.\frac{\partial q_{\mathbf{y}_i}\left(\sum_{k=1}^{\lambda}(\sum_{j=1}^{i-1} q_{\mathbf{y}_j})^{k-1}(\sum_{j=1}^{i} q_{\mathbf{y}_j})^{\lambda-k}\right)}{\partial p_m}\right|_{\mathbf{y}} =$$

$$\left.\frac{\partial q_{\mathbf{y}_i}}{\partial p_m}\right|_{\mathbf{y}} \left(\sum_{k=1}^{\lambda}\left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{y})\right)^{k-1}\left(\sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{y})\right)^{\lambda-k}\right)$$
$$+q_{\mathbf{y}_i}(\mathbf{y})\left.\frac{\partial\left(\sum_{k=1}^{\lambda}(\sum_{j=1}^{i-1} q_{\mathbf{y}_j})^{k-1}(\sum_{j=1}^{i} q_{\mathbf{y}_j})^{\lambda-k}\right)}{\partial p_m}\right|_{\mathbf{y}}. \qquad (35)$$

We split the problem into the following three different cases.

1. Let $\mathbf{y}_i \in \Omega$ be an individual such that $d_H(\mathbf{y},\mathbf{y}_i) \geq 2$. In this case, using $\partial q_{\mathbf{y}_i}/\partial p_m|_{\mathbf{y}} = 0$ (equation (31)) and $q_{\mathbf{y}_i}(\mathbf{y}) = 0$ (equation (28)), equation (35) has a value of 0.

2. Let $\mathbf{y}_i \in \Omega$ be an individual such that $d_H(\mathbf{y},\mathbf{y}_i) = 1$. In this case, in the second term we again have $q_{\mathbf{y}_i}(\mathbf{y}) = 0$ (equation (28)) but, in the first, $\partial q_{\mathbf{y}_i}/\partial p_m|_{\mathbf{y}}$ could be different from zero. However in this case (because $\mathbf{y}_i$ is before $\mathbf{y}$ in the order in Ω and then $q_{\mathbf{y}_j}(\mathbf{y}) = 0$ for all $j = 1,\ldots,i$) the second multiplicative term $\left(\sum_{k=1}^{\lambda}(\sum_{j=1}^{i-1} q_{\mathbf{y}_j})^{k-1}(\sum_{j=1}^{i} q_{\mathbf{y}_j})^{\lambda-k}\right)$ has a value of zero in $\mathbf{y}$. Hence equation (35) has a value of zero for this kind of individual.

3. Finally we take into account the term corresponding to individual $\mathbf{y}$. If $y_m = 0$ this term does not appear in the sum. In the other case ($y_m = 1$), if i represents the place that individual $\mathbf{y}$ takes in the ordering of Ω ($\mathbf{y} = \mathbf{y}_i$), equation (35) can be expressed by:

$$\left.\frac{\partial q_{\mathbf{y}}}{\partial p_m}\right|_{\mathbf{y}} \left(\sum_{k=1}^{\lambda}\left(\sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{y})\right)^{k-1}\left(\sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{y})\right)^{\lambda-k}\right)$$
$$+q_{\mathbf{y}}(\mathbf{y})\left.\frac{\partial\left(\sum_{k=1}^{\lambda}(\sum_{j=1}^{i-1} q_{\mathbf{y}_j})^{k-1}(\sum_{j=1}^{i} q_{\mathbf{y}_j})^{\lambda-k}\right)}{\partial p_m}\right|_{\mathbf{y}}$$
$$= \left.\frac{\partial q_{\mathbf{y}}}{\partial p_m}\right|_{\mathbf{y}}\left(\sum_{k=1}^{\lambda} A_i(\mathbf{y})^{k-1}B_i(\mathbf{y})^{\lambda-k}\right) + q_{\mathbf{y}}(\mathbf{y})\left.\frac{\partial\left(\sum_{k=1}^{\lambda} A_i^{k-1}B_i^{\lambda-k}\right)}{\partial p_m}\right|_{\mathbf{y}}, \qquad (36)$$

where $A_i(\mathbf{y}) = \sum_{j=1}^{i-1} q_{\mathbf{y}_j}(\mathbf{y})$ and $B_i(\mathbf{y}) = \sum_{j=1}^{i} q_{\mathbf{y}_j}(\mathbf{y})$.

In the next reasoning, note that $A_i(\mathbf{y}) = 0$ and $B_i(\mathbf{y}) = 1$.
The first term of equation (36) is:

$$\left.\frac{\partial q_{\mathbf{y}}}{\partial p_m}\right|_{\mathbf{y}}\left(B_i(\mathbf{y})^{\lambda-1} + A_i(\mathbf{y})B_i(\mathbf{y})^{\lambda-2} + \cdots + A_i(\mathbf{y})^{\lambda-1}\right) = 1. \qquad (37)$$

The second term of equation (36) can be expressed as:

$$\begin{aligned}
&\underbrace{q_{\mathbf{y}}(\mathbf{y})}_{=1}\,\frac{\partial\left(B_i^{\lambda-1}+A_iB_i^{\lambda-2}+\cdots+A_i^{\lambda-1}\right)}{\partial p_m}\Bigg|_{\mathbf{y}} = (\lambda-1)B_i(\mathbf{y})^{\lambda-2}\frac{\partial B_i}{\partial p_m}\bigg|_{\mathbf{y}}\\
&\qquad +\frac{\partial A_i}{\partial p_m}\bigg|_{\mathbf{y}} B_i(\mathbf{y})^{\lambda-2} + A_i(\mathbf{y})(\lambda-2)B_i(\mathbf{y})^{\lambda-3}\frac{\partial B_i}{\partial p_m}\bigg|_{\mathbf{y}}\\
&\qquad +\cdots+(\lambda-1)A_i(\mathbf{y})^{\lambda-2}\frac{\partial A_i}{\partial p_m}\bigg|_{\mathbf{y}}\\
&\quad = (\lambda-1)B_i(\mathbf{y})^{\lambda-2}\frac{\partial B_i}{\partial p_m}\bigg|_{\mathbf{y}} + \frac{\partial A_i}{\partial p_m}\bigg|_{\mathbf{y}} B_i(\mathbf{y})^{\lambda-2}.
\end{aligned}\tag{38}$$

Substituting again the values of $A_i(\mathbf{y})$ and $B_i(\mathbf{y})$ in equation (38):

$$\begin{aligned}
&(\lambda-1)\sum_{j=1}^{i}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}} + \sum_{j=1}^{i-1}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}} =\\
&\qquad (\lambda-1)\left(\underbrace{\sum_{\substack{d_H(\mathbf{y}_j,\mathbf{y})\geq 2\\ j<i}}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=0} + \sum_{d_H(\mathbf{y}_j,\mathbf{y})=1}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}} + \underbrace{\frac{\partial q_{\mathbf{y}}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=1}\right)\\
&\qquad +\left(\underbrace{\sum_{\substack{d_H(\mathbf{y}_j,\mathbf{y})\geq 2\\ j<i}}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=0} + \sum_{d_H(\mathbf{y}_j,\mathbf{y})=1}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}}\right).
\end{aligned}\tag{39}$$

Taking into account that $\mathbf{y}_k$ is such that $d_H(\mathbf{y},\mathbf{y}_k)=1$, $y_{k,m}=0$, and $y_m=1$ we find that the second term of equation (36) is:

$$(\lambda-1)\left(\underbrace{\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=-1} + \underbrace{\sum_{\substack{d_H(\mathbf{y}_j,\mathbf{y})=1\\ y_{j,m}=y_m}}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=0} +1\right)+\left(\underbrace{\frac{\partial q_{\mathbf{y}_k}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=-1} + \underbrace{\sum_{\substack{d_H(\mathbf{y}_j,\mathbf{y})=1\\ y_{j,m}=y_m}}\frac{\partial q_{\mathbf{y}_j}}{\partial p_m}\bigg|_{\mathbf{y}}}_{=0}\right) = -1.\tag{40}$$

Hence $D\overline{\mathcal{G}}(\mathbf{y}) = \mathbf{0}$ for all local optimum points $\mathbf{y}$ and all the eigenvalues have a value of zero. Moreover, we have shown that the local optimum points for a function f are stable fixed points of the discrete dynamical system $\overline{\mathcal{G}}$.

In the case of a point $\mathbf{y}$ of Ω that is not a local optimum, following arguments similar to the previous case, it can be shown that:

$$\left.\frac{\partial \overline{\mathcal{G}}_r}{\partial p_m}\right|_{\mathbf{y}} = 0 \text{ for all } r \neq m. \tag{41}$$

In all the previous cases the adding terms have a value of zero, except for the case that there exists $\mathbf{y}'$ such that $d(\mathbf{y}, \mathbf{y}') = 1$, $y'_m \neq y_m$ and in addition $f(\mathbf{y}') > f(\mathbf{y})$ ($\mathbf{y}$ is before $\mathbf{y}'$ in the order imposed by f in Ω), the adding terms corresponding to $\mathbf{y}$ and $\mathbf{y}'$ are different from zero but their sum is zero.

In the case $r = m$, there exists $m \in \{1, \ldots, l\}$ such that

$$\left.\frac{\partial \overline{\mathcal{G}}_m}{\partial p_m}\right|_{\mathbf{y}} > 1. \blacksquare$$

From Theorem 3 it can be deduced that when the value of α is near 0, then the PBIL algorithm can only converge to the local optima of f. So, the PBIL algorithm converges to the global optimum in unimodal functions.

7. Conclusions and future work

We have opened a new approach to the theoretical study of PBIL: using discrete dynamical systems. Associating an appropiate dynamical system we have shown that the PBIL algorithm follows the iterates of the dynamical system. In addition, we have seen that PBIL can only converge to local optima; meaning, in the case of unimodal functions, that PBIL converges to the global optimum.

This is a preliminary analysis and much work remains to be done. Our first objective is to generalize the results to the case in which $\mu > 1$ individuals are selected. Furthermore, we plan to study the size of the basin of attraction of each local optimum to calculate the probability of convergence to each local optimum.

Acknowledgments

This work was supported by the University of the Basque Country under grant 9/UPV/EHU 00140.226-12084/2000. Also C. González is supported by UPV-EHU.

References

[1] P. Larrañaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation* (Kluwer Academic Publishers, 2001).

[2] H. Mühlenbein and G. Paaβ, "From Recombination of Genes to the Estimation of Distributions: I. Binary Parameters," in *Parallel Problem Solving from Nature, PPSN-IV,* edited by H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel (Springer-Verlag, Berlin Heidelberg, 1996).

[3] P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña, "Combinatorial Optimization by Learning and Simulation of Bayesian Networks," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, UAI-2000*, edited by C. Boutilier and M. Goldszmidt (Morgan Kaufmann, San Francisco, 2000).

[4] P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña, "Optimization in Continuous Domains by Learning and Simulation of Gaussian Networks," in *Proceedings of the Genetic and Evolutionary Computation Conference, Workshop Program*, edited by A. S. Wu (Las Vegas, Nevada, 2000).

[5] M. Pelikan, D. E. Goldberg, and F. Lobo "A Survey of Optimization by Building and Using Probabilistic Models," Technical Report IlliGAL, 99018 (University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, 1999).

[6] S. Baluja, "Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning," Technical Report CMU-CS-94-163 (Computer Science Department, Carnegie Mellon University, Pittsburgh, 1994).

[7] E. Galić and M. Höhfeld, "Improving the Generalization Performance of Multi-Layer Perceptrons with Population-Based Incremental Learning," in *Parallel Problem Solving from Nature, PPSN-IV,* edited by H.-M. Voight, W. Ebeling, I. Rechenberg, and H.-P. Schwefel (Springer-Verlag, Berlin Heidelberg, 1996).

[8] B. Maxwell and S. Anderson, "Training Hidden Markov Models Using Population-Based Learning," *Genetic and Evolutionary Computation Conference, GECCO-99* (1999).

[9] I. Inza, M. Merino, P. Larrañaga, J. Quiroga, B. Sierra, and M. Girala, "Feature Subset Selection by Genetic Algorithms and Estimation of Distribution Algorithms. A Case Study in the Survival of Cirrhotic Patients Treated with TIPS," *Artificial Intelligence in Medicine* (in press).

[10] R. Sukthankar, S. Baluja, and J. Hancock, "Evolving an Intelligent Vehicle for Tactical Reasoning in Traffic," *International Conference on Robotics and Automation, ICRA'97* (1997).

[11] M. Gallagher, "Multi-layer Perceptron Error Surfaces: Visualization, Structure and Modelling," PhD thesis, (Department of Computer Science and Electrical Engineering, University of Queensland, 2000).

[12] M. P. Servais, G. de Jaer, and J. R. Greene, "Function Optimization Using Multiple-Base Population Based Incremental Learning," in *Proceedings of the Eighth South African Workshop on Pattern Recognition* (1997).

[13] V. Kvasnicka, M. Pelikan, and J. Pospichal, "Hill Climbing with Learning (an Abstraction of Genetic Algorithms)," *Neural Networks World*, **6** (1996) 773–796.

[14] S. Rudlof and M. Köppen, "Stochastic Hill Climbing by Vectors of Normal Distributions," in *Proceedings of the First Online Workshop on Soft Computing, WSC1* (University of Nagoya, Nagoya, Japan, 1996).

[15] M. Sebag and A. Ducoulombier, "Extending Population-Based Incremental Learning to Continuous Search Spaces," in *Parallel Problem Solving from Nature, PPSN-V* (Springer, 1998).

[16] A. Berny, "An Adaptative Scheme for Real Function Optimization Acting as a Selection Operator," in *First IEEE Symposium of Combinations of Evolutionary Computation and Neural Networks* (2000).

[17] S. Baluja, "Genetic Algorithms and Explicit Search Statistics," *Advances in Neural Information Processing Systems,* edited by M. Mozer, M. Jordan, and T. Petsche (The MIT Press, Cambridge, MA, 1997).

[18] N. Monmarché, E. Ramat, L. Desbarats, and G. Venturini, "Probabilistic Search with Genetic Algorithms and Ant Colonies," in *Proceedings of the Genetic and Evolutionary Computation Conference, Workshop Program,* edited by A. S. Wu (Las Vegas, Nevada, 2000).

[19] N. Monmarché, E. Ramat, G. Dromel, M. Slimane, and G. Venturini, "On the Similarities between AS, BSC, and PBIL: Toward the Birth of a New Meta-Heuristic," Technical Report, 215, E3i (Laboratoire d'Informatique pour l'Industrie (E3i), Université de Tours, 1999).

[20] M. Schmidt, K. Kristensen, and T. R. Jensen, "Adding Genetics to the Standard PBIL Algorithm," in *Congress on Evolutionary Computation, CEC'99* (1999).

[21] C. Fyfe, "Structured Population-Based Incremental Learning," *Soft Computing*, **2** (1999) 191–198.

[22] S. Baluja and R. Caruana, "Removing the Genetics from the Standard Genetic Algorithm," in *Proceedings of the International Conference on Machine Learning,* edited by A. Prieditis and S. Russell (Morgan Kaufmann, 1995).

[23] C. González, J. A. Lozano, and P. Larrañaga, "The Convergence Behavior of the PBIL Algorithm: A Preliminary Approach," in *Fifth International Conference on Artificial Neural Networks and Genetic Algorithms, ICANNGA'2001* (in press).

[24] M. Höhfeld and G. Rudolph, "Towards a Theory of Population-Based Incremental Learning," in *Proceedings of the Fourth IEEE Conference on Evolutionary Computation* (IEEE Press, 1997).

[25] A. Berny, "Selection and Reinforcement Learning for Combinatorial Optimization," in *Parallel Problem Solving from Nature, PPSN-VI,* edited by M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel (Springer-Verlag, Berlin Heidelberg, 2000).

[26] H. Mühlenbein, "The Equation for Response to Selection and its Use for Prediction," *Evolutionary Computation*, **5** (1998) 303–346.

[27] H. Mühlenbein and T. Mahnig, "Evolutionary Computation and Wright's Equation," *Theoretical Computer Science* (in press).

[28] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory* (MIT Press, 1999).

[29] M. D. Vose, "Random Heuristic Search," *Theoretical Computer Science*, **229**(1-2) (1999) 103–142.

[30] E. R. Sheinerman, *Invitation to Dynamical Systems* (Prentice-Hall, 1996).