Towards a Theory of Population–Based Incremental Learning

Markus Höhfeld
Siemens AG, ZT AN 2
Otto–Hahn–Ring 6
D–81730 München / Germany
Markus.Hoehfeld@mchp.siemens.de

Günter Rudolph
Informatik Centrum Dortmund (ICD)
Joseph–von–Fraunhofer–Str. 20
D–44227 Dortmund / Germany
rudolph@icd.de

***Abstract*— Probabilistic Incremental Learning (PBIL) has been proposed as a model for binary coded evolutionary algorithms where the population is represented by its mean vector which is updated in an autoregressive manner.**

In this paper we prove that under the PBIL update rule each component of the population vector converges with probability one to 0 (1), whenever a 0 (1) in the associated bit consistently yields better fitness values than the opposite setting. As a corollary, we obtain global convergence of PBIL for linear pseudoboolean functions including the commonly investigated Counting-Ones problem.

I. Introduction

Although the idea of *population–based incremental learning* (PBIL) was presented to a broader public only as recently as 1995 at a Machine Learning Conference [1] there already exist applications to real–world optimization problems [4]. At Siemens AG, PBIL has been hybridized with gradient techniques to concurrently evolve optimal topologies and weights for neural networks in forecasting applications [3]. Meanwhile, some theoretical claims regarding a time–continuous model of PBIL are available [5]. Unfortunately, proofs are missing.

Here, we formulate and investigate PBIL as a stochastic time–discrete dynamical system. In the course of the analysis of the expected behaviour of PBIL, we prove convergence for linear pseudoboolean functions and we illustrate the convergence behaviour on a particular nonlinear problem. While our results also provide insight into PBIL's behaviour on more complex applications, a formal investigation of such settings is beyond the scope of this paper.

II. Population–Based Incremental Learning

The concept of PBIL rests on the idea that a population of admissible solutions of some minimization problem with search space $\mathbb{B}^\ell = \{\mathtt{0}, \mathtt{1}\}^\ell$ can be represented by the statistics of its gene pool. Assume that we have a finitely large population of μ individuals in $\mathbb{B}^\ell$ that may be gathered in the matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1\ell} \\ x_{21} & x_{22} & \cdots & x_{2\ell} \\ \vdots & \vdots & & \vdots \\ x_{\mu 1} & x_{\mu 2} & \cdots & x_{\mu\ell} \end{pmatrix}$$

where the row vector $x_{i\bullet} = (x_{i1}\, x_{i2}\, \dots\, x_{i\ell})$ represents the ith individual ($i = 1, \dots, \mu$) whereas the column vector $x_{\bullet j} = (x_{1j}\, x_{2j}\, \dots\, x_{\mu j})'$ represents the *gene pool* of component $j = 1, \dots, \ell$. In a usual evolutionary algorithm with *gene pool recombination* [7, p. 170] an offspring y would be generated by choosing a gene at random with uniform probability $1/\mu$ from each gene pool. Evidently, this is equivalent to the method to calculate the relative frequencies p_j of the 1s in each gene pool and to generate an offspring by drawing component y_j from a binomial distribution with $\mathsf{P}\{\, y_j = \mathtt{1}\,\} = p_j$ and $\mathsf{P}\{\, y_j = \mathtt{0}\,\} = 1 - p_j$ for $j = 1, \dots, \ell$. Thus, one only needs the vector $p \in (0,1)^\ell \subset \mathbb{R}^\ell$ of relative frequencies to represent the population and to generate λ offspring with $\mu < \lambda$. The μ best offspring are selected and their gene pool serves to determine the new vector of relative frequencies. This method is a variant of an evolutionary algorithm experimentally investigated in [2].

To elucidate the differences between this method and PBIL let us consider their update rules: Let $y_{1:\lambda}, y_{2:\lambda}, \dots, y_{\lambda:\lambda}$ denote the λ sorted trial vectors with $f(y_{1:\lambda}) \le f(y_{2:\lambda}) \le \cdots \le f(y_{\lambda:\lambda})$ where $f : \mathbb{B}^\ell \to \mathbb{R}$ is the objective function to be minimized. With $\alpha \in (0,1)$ and $t \in \mathbb{N}_0$ the update rules are given by

$$p^{(t+1)} = \frac{1}{\mu} \sum_{k=1}^{\mu} y_{k:\lambda}^{(t)} \tag{1}$$

$$p^{(t+1)} = (1-\alpha)\, p^{(t)} + \alpha\, \frac{1}{\mu} \sum_{k=1}^{\mu} y_{k:\lambda}^{(t)} \tag{2}$$

where the initial setting is $p_i^{(0)} = 1/2$ for $i = 1, \dots, \ell$. Rule (1) is associated with the evolutionary algorithm employing gene pool recombination and truncation selection as described above whereas rule (2) is associated with PBIL.

Note that both update rules lead to stochastic algorithms that can be modeled via Markov chains. The state space of the Markov chain associated with update rule (1) is finite with cardinality $(\mu+1)^\ell$ and it is easy to see that $p^{(t)}$ can be absorbed with nonzero probability by each vector $p^{(\infty)} \in \mathbb{B}^\ell$ regardless of the objective function. As a consequence, this algorithm may converge to each admissible solution—regardless of optimality.

In contrast to the algorithm above, PBIL does not seem to have this property. Note that the state space of the Markov chain associated with update rule (2) is infinite but denumerable. Since $p^{(t)} \in (0,1)^\ell$ for all $t \ge 0$, provided that $p^{(0)} \in (0,1)^\ell$, the process cannot be trapped in points represented by the set $\mathbb{B}^\ell$. Nevertheless, it is still possible that the process stochastically converges to a corner of the hypercube. In the best case the global solution of objective function $f(\cdot)$ is the only point to which the process will stochastically converge. Even if only local solutions (with respect to Hamming distance) were candidates of such an event, PBIL would be preferable to the EA associated with update rule (1).

It is our goal to investigate the convergence properties of PBIL. Following common practice [1; 3] will shall consider PBIL's update rule for the special case $\mu = 1$ leading to

$$p^{(t+1)} = (1-\alpha)\, p^{(t)} + \alpha\, b^{(t)} \tag{3}$$

with $b^{(t)} := y^{(t)}_{1:\lambda}$. Thus, only the best out of λ trial vectors is involved in updating the vector of probabilities.

III. Theoretical Analysis

Since the stochastic sequence $(p^{(t)} : t \geq 0)$ is bounded in $(0,1)^\ell$ we may interchange limit and expectation, i.e., the relation

$$\lim_{t\to\infty} \mathsf{E}[\, p^{(t)}\,] = \mathsf{E}[\lim_{t\to\infty} p^{(t)}\,]$$

is valid. If we obtain the result $\lim_{t\to\infty} \mathsf{E}[\, p^{(t)}\,] = x \in \mathbb{B}^\ell$ we may conclude that the stochastic sequence $(p^{(t)} : t \geq 0)$ converges in mean (and therefore in probability) to the limit $x \in \mathbb{B}^\ell$. Thus, we shall investigate the (deterministic) sequence of expectations $(\mathsf{E}[\, p^{(t)}\,] : t \geq 0)$ to identify points in $[\,0,1]^\ell$ to which PBIL's stochastic process $(p^{(t)} : t \geq 0)$ will eventually converge.

The first step in the determination of the mean value recursion of PBIL consists of taking conditional expectations with respect to $p^{(t)}$ on both sides of equation (3) which leads to

$$\mathsf{E}[\, p^{(t+1)} \,|\, p^{(t)}\,] = (1-\alpha)\, p^{(t)} + \alpha\, \mathsf{E}[\, b^{(t)} \,|\, p^{(t)}\,] \tag{4}$$

where $\mathsf{E}[\, b^{(t)} \,|\, p^{(t)}\,] = F_\lambda(p^{(t)})$ is expressible as a function $F_\lambda(\cdot)$ with argument $p^{(t)}$. Notice that the conditional expectation $\mathsf{E}[\, p^{(t+1)} \,|\, p^{(t)}\,]$ above is a *random vector* representing the least–squares–best predictor of random vector $p^{(t+1)}$. Thus, we have to take expectations on both sides of equation (4). Since $\mathsf{E}[\,\mathsf{E}[\, p^{(t+1)} \,|\, p^{(t)}\,]\,] = \mathsf{E}[\, p^{(t+1)}\,]$ we obtain

$$\mathsf{E}[\, p^{(t+1)}\,] = (1-\alpha)\, \mathsf{E}[\, p^{(t)}\,] + \alpha\, \mathsf{E}[\, F_\lambda(p^{(t)})\,]\,. \tag{5}$$

Unless $F_\lambda(\cdot)$ is a linear function we have $\mathsf{E}[\, F_\lambda(p^{(t)})\,] \neq F_\lambda(\mathsf{E}[\, p^{(t)}\,])$ in general. Consequently, if the map $G(\cdot)$ describes the true mean value recursion of PBIL via $\mathsf{E}[\, p^{(t+1)}\,] = G(\mathsf{E}[\, p^{(t)}\,])$, we also obtain

$$G(q) \neq (1-\alpha)\, q + \alpha F_\lambda(q)$$

in general. In any case, a closer look at function $F_\lambda(\cdot)$ is necessary. In the sequel we shall suppress the iteration counter. Since

$$\mathsf{E}[\, b \,|\, p\,] = \sum_{x\in\mathbb{B}^\ell} x \cdot \mathsf{P}\{\, b = x\,\} \tag{6}$$

we need the probability distribution of random vector b.

Theorem 1 *Let $f : \mathbb{B}^\ell \to \mathbb{R}$ be a pseudoboolean function and let $p \in (0,1)^\ell$ be the vector of probabilities $p_i = \mathsf{P}\{\, s_i = \mathtt{1}\,\}$ of the independent events to draw a* $\mathtt{1}$ *at position $i = 1,\ldots,\ell$ of random vector s and a* $\mathtt{0}$ *otherwise. If $\lambda \geq 2$ vectors s are drawn independently according to the probabilities p then the probability $\mathsf{P}\{\, b = x\,\}$ that $x \in \mathbb{B}^\ell$ is the vector with the least function value is given by*

$$\mathsf{P}\{\, b = x\,\} =$$

$$\mathsf{P}\{\, s = x\,\} \sum_{k=0}^{\lambda-1} \mathsf{P}\{\, f(s) > f(x)\,\}^k \cdot \mathsf{P}\{\, f(s) \geq f(x)\,\}^{\lambda-1-k}.$$

Proof: Assume that the λ independent and identically distributed trial vectors are drawn sequentially so that they are indexed by numbers from 1 to λ. If there is more than one vector with the least objective function value the tie is broken by choosing that vector with the smallest index. Let $x \in \mathbb{B}^\ell$ be arbitrary but fixed and let A_k with $k = 0,\ldots,\lambda-1$ denote the following event:

1. The first k trial vectors are worse than vector x.
2. The vector x is drawn in trial $k+1$.
3. The last $\lambda-1-k$ trial vectors are not better than vector x.

Each of these events A_k leads to the assignment $b := x$. As a consequence, the probability that some vector $x \in \mathbb{B}^\ell$ is the selected best one is given by

$$\mathsf{P}\{\, b_\lambda = x\,\} = \sum_{k=0}^{\lambda-1} \mathsf{P}\{\, A_k\,\} =$$

$$\sum_{k=0}^{\lambda-1} \mathsf{P}\{\, f(s) > f(x)\,\}^k \cdot \mathsf{P}\{\, s = x\,\} \cdot \mathsf{P}\{\, f(s) \geq f(x)\,\}^{\lambda-1-k}$$

and the proof is completed. □

Remark: *Elementary transformations of the sum above lead to the equivalent expression*

$$\mathsf{P}\{\, b_\lambda = x\,\} =$$

$$\mathsf{P}\{\, s = x\,\} \frac{\mathsf{P}\{\, f(s) \geq f(x)\,\}^\lambda - \mathsf{P}\{\, f(s) > f(x)\,\}^\lambda}{\mathsf{P}\{\, f(s) \geq f(x)\,\} - \mathsf{P}\{\, f(s) > f(x)\,\}}\,. \quad \square$$

Since the probabilities of the events appearing in Theorem 1 can be expressed in terms of p via

$$\begin{aligned}
\mathsf{P}\{\, s = x\,\} &= \prod_{i=1}^{\ell} p_i^{x_i}\,(1-p_i)^{1-x_i} \\
\mathsf{P}\{\, f(s) = f(x)\,\} &= \sum_{\substack{y\in\mathbb{B}^\ell \\ f(y)=f(x)}} \mathsf{P}\{\, s = y\,\} \\
\mathsf{P}\{\, f(s) > f(x)\,\} &= \sum_{\substack{y\in\mathbb{B}^\ell \\ f(y)>f(x)}} \mathsf{P}\{\, s = y\,\}
\end{aligned}$$

the expectation of the best vector b in equation (6) is given by a nonlinear function of the probabilities gathered in vector p and the sample size λ.

To illustrate the line of thoughts appearing in the subsequent proofs we shall explicitly elaborate the simplest nontrivial case with $\ell = 2$ and $\lambda = 2$ for the counting–ones problem with objective function

$$f(x) = \sum_{i=1}^{\ell} x_i \, .$$

Following the general approach presented in Theorem 1 we straightforwardly obtain the probability distribution of b and hence its expectation. Since

$$\begin{aligned}
\mathsf{P}\left\{ b = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\} &= 1 - (p_1 + p_2 - p_1\, p_2)^2 \\
\mathsf{P}\left\{ b = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} &= (1 - p_1)\, p_2\, (p_1 + p_2) \\
\mathsf{P}\left\{ b = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} &= p_1\, (1 - p_2)\, (p_1 + p_2) \\
\mathsf{P}\left\{ b = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} &= (p_1\, p_2)^2
\end{aligned}$$

the expectation of b conditioned by p is

$$\mathsf{E}[\, b \,|\, p\,] = F_2(p) = \begin{pmatrix} p_1\,(1-p_2)\,(p_1+p_2) + (p_1\,p_2)^2 \\ p_2\,(1-p_1)\,(p_1+p_2) + (p_1\,p_2)^2 \end{pmatrix} . \tag{7}$$

Figure 1 shows the vector field of the map $F_2(p)$. The interpretation is as follows: For given probabilities p the least–squares–best prediction of the probabilities in the next iteration is $F_2(p)$ and the arrows represent the directions from p to $F_2(p)$.

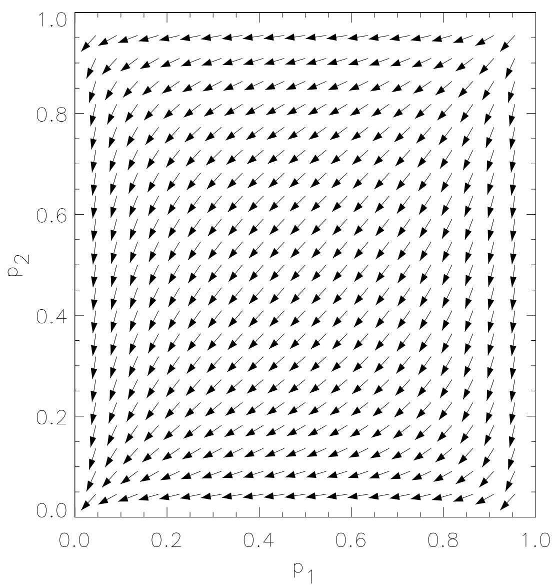

Fig. 1: Vector field of map $F_2(p)$.

A closer look at Figure 1 provides evidence that the relation $F_2(p) < p$ could be valid. In fact, this is true. As a consequence, equation (4) can be bounded via

$$\begin{aligned}
\mathsf{E}[\, p^{(t+1)} \,|\, p^{(t)}\,] &= (1-\alpha)\, p^{(t)} + \alpha\, F_2(p^{(t)}) \\
&< (1-\alpha)\, p^{(t)} + \alpha\, p^{(t)} \;=\; p^{(t)}
\end{aligned}$$

and we obtain

$$\mathsf{E}[\, \mathsf{E}[\, p^{(t+1)} \,|\, p^{(t)}\,]\,] = \mathsf{E}[\, p^{(t+1)}\,] < \mathsf{E}[\, p^{(t)}\,] \, .$$

We may conclude that $\mathsf{E}[\, p_i^{(t)}\,] \to 0$ as $t \to \infty$ for $i = 1, 2$.

This example shows that PBIL will converge in mean to a point in $\mathbb{B}^\ell$ if the flux of the conditional expectation moves in only one direction for each dimension. A more formal and general statement of this fact is given below.

Theorem 2 *Let $f : \mathbb{B}^\ell \to \mathbb{R}$ be some pseudoboolean function and $(p^{(t)} : t \geq 0)$ be the random sequence generated by PBIL. Then*

$$\mathsf{E}[\, p_i^{(t)}\,] \to x_i^* = \begin{cases} 0 & \textit{if} \quad \forall p \in (0,1)^\ell : \mathsf{E}[\, b_i \,|\, p\,] < p_i \\ 1 & \textit{if} \quad \forall p \in (0,1)^\ell : \mathsf{E}[\, b_i \,|\, p\,] > p_i \end{cases}$$

as $t \to \infty$. □

It is clear that the condition of the theorem above will not be fulfilled for arbitrary pseudoboolean functions. Moreover, the verification of the conditions requires tedious calculations. Therefore it is useful to develop sufficient criteria that are much easier to check.

Theorem 3 *Let e_i be the i–th unit vector with dimension ℓ and $\bar{e}_i$ its binary complement. If*

$$\frac{\mathsf{P}\{\, b = x \vee e_i \,\}}{\mathsf{P}\{\, s = x \vee e_i \,\}} < \frac{\mathsf{P}\{\, b = x \wedge \bar{e}_i \,\}}{\mathsf{P}\{\, s = x \wedge \bar{e}_i \,\}} \tag{8}$$

for arbitrary $x \in \mathbb{B}^\ell$, $p \in (0,1)^\ell$, and $i \in \{1, \ldots, \ell\}$ then $\mathsf{E}[\, b_i \,|\, p\,] < p_i$. If the inequality sign in (8) is reversed one obtains $\mathsf{E}[\, b_i \,|\, p\,] > p_i$.

Proof: With $r_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{\ell} p_j^{x_j}\,(1-p_j)^{1-x_j} > 0$ one obtains

$$\left.\begin{aligned}
\mathsf{P}\{\, s = x \wedge \bar{e}_i \,\} &= (1-p_i)\, r_i(x) \\
\mathsf{P}\{\, s = x \vee e_i \,\} &= p_i\, r_i(x)
\end{aligned}\right\} \tag{9}$$

Insertion of the identities (9) into inequality (8) and rearrangement leads to

$$\mathsf{P}\{\, b = x \vee e_i \,\} < p_i\,(\, \mathsf{P}\{\, b = x \wedge \bar{e}_i \,\} + \mathsf{P}\{\, b = x \vee e_i \,\}\,) \, . \tag{10}$$

The conditional expectation of b_i can be bounded via

$$\begin{aligned}
\mathsf{E}[\, b_i \,|\, p\,] &= \sum_{x \in \mathbb{B}^\ell} x_i\, \mathsf{P}\{\, b = x \,\} \\
&= \frac{1}{2} \sum_{x \in \mathbb{B}^\ell} \mathsf{P}\{\, b = x \vee e_i \,\} \qquad (11) \\
&< \frac{1}{2}\, p_i \sum_{x \in \mathbb{B}^\ell} (\mathsf{P}\{\, b = x \wedge \bar{e}_i \,\} + \mathsf{P}\{\, b = x \vee e_i \,\}) \\
&= \frac{1}{2}\, p_i \sum_{x \in \mathbb{B}^\ell} 2\, \mathsf{P}\{\, b = x \,\} \;=\; p_i
\end{aligned}$$

where inequality (10) was inserted into (11). The proof for reversed inequality sign in (8) is analogous and therefore omitted. □

Thanks to Theorem 3 we are exempted from the task to determine explicit bounds on the conditional expectation. As it is shown below, the conditions of Theorem 3 are fulfilled for linear pseudoboolean functions.

Theorem 4 *Let $f(x) = c_0 + c'x$ be a linear function with $x \in \mathbb{B}^\ell$ and $c_i \in \mathbb{R} \setminus \{0\}$ for all $i = 0, 1, \ldots, \ell$. Then*

$$\frac{\mathsf{P}\{\, b = x \vee e_i \,\}}{\mathsf{P}\{\, s = x \vee e_i \,\}} < \frac{\mathsf{P}\{\, b = x \wedge \bar{e}_i \,\}}{\mathsf{P}\{\, s = x \wedge \bar{e}_i \,\}} \quad \text{if } c_i > 0 \text{ and}$$

$$\frac{\mathsf{P}\{\, b = x \vee e_i \,\}}{\mathsf{P}\{\, s = x \vee e_i \,\}} > \frac{\mathsf{P}\{\, b = x \wedge \bar{e}_i \,\}}{\mathsf{P}\{\, s = x \wedge \bar{e}_i \,\}} \quad \text{if } c_i < 0$$

for all $i = 1, \ldots, \ell$ simultaneously if $p \in (0,1)^\ell$.

Proof: Owing to Theorem 1 we have

$$\frac{\mathsf{P}\{\, b = x \,\}}{\mathsf{P}\{\, s = x \,\}} = \sum_{k=0}^{\lambda-1} \mathsf{P}\{\, f(s) > f(x) \,\}^k \cdot \mathsf{P}\{\, f(s) \geq f(x) \,\}^{\lambda-1-k}. \quad (12)$$

With $f_i(x) = c_0 + \sum_{\substack{j=1 \\ j \neq i}}^{\ell} c_j \, x_j$ we obtain:

1. If $c_i > 0$ then
$$f(x \vee e_i) = c_i + f_i(x) > f_i(x) = f(x \wedge \bar{e}_i)$$
and hence
$$\mathsf{P}\{\, f(s) \overset{(>)}{\geq} f(x \vee e_i) \,\} < \mathsf{P}\{\, f(s) \overset{(>)}{\geq} f(x \wedge \bar{e}_i) \,\}. \quad (13)$$
2. If $c_i < 0$ then
$$f(x \vee e_i) = c_i + f_i(x) < f_i(x) = f(x \wedge \bar{e}_i)$$
and hence
$$\mathsf{P}\{\, f(s) \overset{(>)}{\geq} f(x \vee e_i) \,\} > \mathsf{P}\{\, f(s) \overset{(>)}{\geq} f(x \wedge \bar{e}_i) \,\}. \quad (14)$$

Replacement of x by $x \vee e_i$ and $x \wedge \bar{e}_i$, respectively, in (12) yields two finite sums. Taking into account the inequalities (13) and (14) a pairwise comparison of the summands in both sums leads to the desired result. □

The corollary below offers a summary and a strengthening of the results:

Corollary 1 *Let $f(x) = c_0 + c'x$ be a linear pseudoboolean function with $x \in \mathbb{B}^\ell$ and $c_i \in \mathbb{R} \setminus \{0\}$ for $i = 0, 1, \ldots, \ell$. If $(p^{(t)} : t \geq 0)$ denotes the random sequence generated by PBIL, then $p_i^{(t)} \to x_i^*$ for all $i = 1, \ldots, \ell$ in mean and with probability 1 as $t \to \infty$, where x^* is the global solution of the objective function $f(\cdot)$.*

Proof: Theorem 4 implies that the conditions of Theorem 3 are fulfilled if PBIL is applied to linear pseudoboolean functions. Theorem 3 in turn satisfies the conditions of Theorem 2 and we may conclude that $p^{(t)} \to x^*$ in mean as $t \to \infty$.

As for convergence with probability 1, note that Theorem 3 also implies that

$$\mathsf{E}[\, \|p^{(t+1)} - x^*\| \,|\, p^{(t)} \,] < \|p^{(t)} - x^*\| \quad (15)$$

for all $t \geq 0$, where $\|\cdot\|$ denotes the 1–norm (i.e., sum of absolute values). Inequality (15) reveals that the random sequence $(\|p^{(t)} - x^*\| : t \geq 0)$ is a nonnegative supermartingale [6] that converges with probability 1 to its finite limit v, say.

Since both convergence in mean and convergence with probability 1 implies convergence in probability, we obtain the result that $\|p^{(t)} - x^*\|$ converges in probability to zero as well as to v. But since the limits must be unique, we may conclude that $\mathsf{P}\{\, v = 0 \,\} = 1$. □

Note that linear pseudoboolean functions have exactly one local solution with respect to Hamming distance. The next example provides evidence that the behavior of PBIL in case of nonlinear problems with more than one local solution is different from the linear case.

Let $f : \mathbb{B}^2 \to \mathbb{R}$ represent a class of objective functions obeying the relation $f(\mathtt{00}) < f(\mathtt{11}) < f(\mathtt{01}) < f(\mathtt{10})$. For example, an instance of this class is the objective function $f(x_1, x_2) = 3\,x_1 + 2\,x_2 - 4\,x_1\,x_2$. Every instance of this class is a nonlinear function with two local minima attained at the positions $x^* = (\mathtt{0}, \mathtt{0})'$ and $x^* = (\mathtt{1}, \mathtt{1})'$.

If the number of trial vectors is set to $\lambda = 2$ the conditional expectation of b is

$$\mathsf{E}[\, b \,|\, p \,] = G_2(p) = \begin{pmatrix} p_1 \,[\, p_1 + 2\, p_2^2\,(1 - p_1)] \\ p_2 \,[\, p_2 + 2\, p_1\,(1 - p_2)] \end{pmatrix} \quad (16)$$

where we have used the method presented previously to obtain the probability distribution of b. Figure 2 shows the motion of the conditional expectation for given p.

Evidently, PBIL will converge to both optima with a certain probability depending on the initial setting of $p^{(0)}$. Therefore PBIL will not stochastically converge to the global solution in the general case.

IV. Conclusions

We presented an analysis of the convergence behavior of the PBIL algorithm. It was proven that the simple dynamics of PBIL's update rule ensure convergence with probability 1 to the global optimum in the case of linear pseudoboolean functions.

As for nonlinear problems, the behavior of PBIL becomes more complex. Based on a simple example we have shown graphically that PBIL may be attracted by local (non–global) solutions as well.

Acknowledgments

Both authors gratefully acknowledge financial support by the project EVOALG (grants 01 IB 403 C and 01 IB

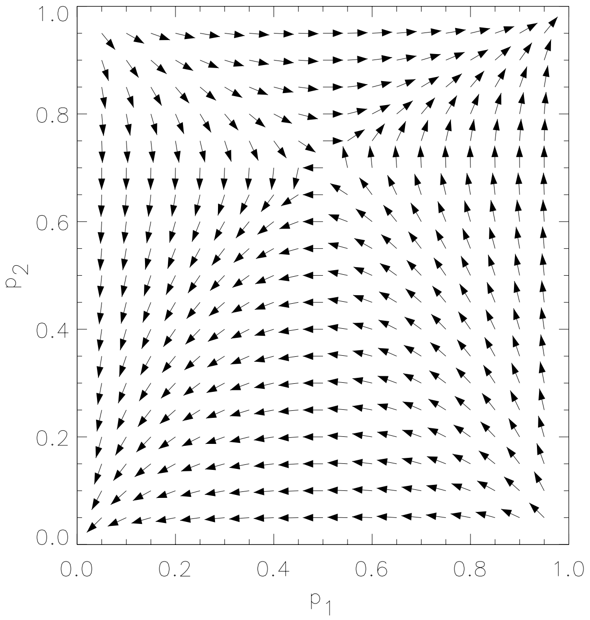

Fig. 2: Vector field of map $G_2(p)$.

403 A) from the German Federal Ministry of Education, Science, Research and Technology (BMBF).

References

[1] S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithm. In A. Prieditis and S. Russel, editors, *Proceedings of the International Conference on Machine Learning (ML–95)*, pages 38–46. Morgan Kaufmann, San Mateo (CA), 1995.

[2] L. J. Eshelman and J. D. Schaffer. Crossover's niche. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 9–14. Morgan Kaufmann, San Mateo (CA), 1993.

[3] E. Galić and M. Höhfeld. Improving the generalization performance of multi–layer–perceptrons with population–based incremental learning. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Parallel Problem Solving From Nature — PPSN IV*, pages 740–750. Springer, Berlin, 1996.

[4] J. R. Greene. Population–based incremental learning as a simple versatile tool for engineering optimisation. In *Proceedings of the First International Conference on Evolutionary Computation and Its Applications (EvCA'96)*, pages 258–269. Russian Academy of Science, Moscow, 1996.

[5] V. Kvasnicka, M. Pelikan, and J. Pospichal. Hill climbing with learning (an abstraction of genetic algorithm). In *Proceedings of the First International Conference on Genetic Algorithms (MENDEL'95)*, pages 65–73. Brno (Czech Repbublic) 1995.

[6] J. Neveu. *Discrete–Parameter Martingales.* North Holland, Amsterdam and Oxford, 1975.

[7] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.* Birkhäuser, Basel, 1977.